TUM School of Computation, Information and Technology
Technische Universität München

Promoting Human-Centered AI by Integrating Human Factors into Model Design

Yao Rong

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Georg Groh

Prüfende der Dissertation:

1. Prof. Dr. Enkelejda Kasneci
2. Prof. Dr. Xia Hu

Die Dissertation wurde am 20.12.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 19.05.2024 angenommen.

Acknowledgements

First and foremost, I would like to thank my amazing advisor Professor Enkelejda Kasneci. I truly enjoyed being her student. Her vision in human-centered AI technologies and expertise in eye tracking has laid a robust foundation for my doctoral project. I vividly recall the hands-on assistance she provided from my very first paper, a testament to her dedication and supportive nature. Without her support, my current research achievements would not have been possible. Beyond the academic guidance, she offered invaluable advice that significantly shaped my career path. Her cheerful attitude and encouragement have helped me stay positive, believe in myself, successfully complete my doctoral studies and venture confidently into my professional career. For that, I am forever grateful to her.

I would also like to thank my other thesis committee members. I extend my gratitude to my thesis committee chair, Professor Georg Groh, for his commitment, and the time and dedication he has invested in this role. I thank Professor Xia Hu, my host professor during my research stay at Rice University, for providing insightful advice for both my current research and future career. He encouraged me to learn from other brilliant peers in the field of explainable AI techniques.

I am very fortunate to receive advice and guidance from mentors and collaborators across diverse research areas. In particular, I thank Professor Gjergji Kasneci for his mentorship and support throughout my doctoral studies. I always received constructive and insightful feedback from him on our research projects and his positive energy instilled confidence in me. I extend my thanks to Professor Zeynep Akata for her visionary inspiration in integrating human attention with computer vision models. I have gained considerable knowledge from her comments throughout our manuscripts. I am grateful to Professor Vaibhav Unhelkar for his invaluable technical feedback on our collaborative projects about human-centered explainable AI.

Many thanks go to *all* my colleagues and friends both at the Human-Computer Interaction team and Data Science & Analytics groups at the University of Tübingen. Especially, I would like to thank Tobias for those thought-provoking discussions we shared, and for his mental support during those high-pressure moments as we were catching deadlines. I would also like to thank Nora, Wolfgang, and Efe who provided invaluable guidance in my research. I also thank *all* my colleagues from the group Human-Centered Technologies for Learning and the group Reliable Data Science, with whom I have shared so many delightful memories. Without all these colleagues, I would never have enjoyed my doctoral studies that much.

I would like to express my heartfelt gratitude to all my fantastic colleagues from Rice University, especially the colleagues from DATA Lab and Human-Centered AI and Robotics group at Rice University. Thanks to them, I had a wonderful time in Houston.

I extend my special thanks to Bernhard, who always offered me encouragement and assistance in refining my papers with great patience.

Finally, I dedicate this dissertation to my family, especially my mom and dad. Unfortunately, I could not spend enough time with my family during my doctoral studies, yet their unconditional love and support have been my strength, enabling me to chase my goals. I am so lucky to have them always by my side.

Abstract

Recently, we have witnessed advanced artificial intelligence (AI) models used in many domains, such as medicine, finance, and education. However, there have been concerns raised that AI systems could become so advanced that they might replace people in many jobs and make decisions over human control. To address these concerns, many prominent institutes propose the strategy of Human-centered AI (HAI) to ensure that the next generation of AI will technically reflect human behaviors, focus on the impact of AI on humans, and augment humans' capabilities rather than replace them.

This dissertation focuses on building HAI models that can capture the complexity of human intelligence, understand human needs, and provide human-understandable explanations to build trustworthy and reliable AI systems. Concretely, this dissertation studies three inseparable research challenges to promote HAI:

(1) **Human attention**: incorporating human expertise in visual perception into the decision-making to enhance the model capability. In particular, this part proposes to use human gaze-based attention data to enhance the model attention module, which improves model performance in challenging visual tasks such as exploring fine-grained distinct features.

(2) **Human intentions**: designing models that can better foresee a human's reaction, i.e., intentions, within specific contexts for achieving effective human-AI collaboration. This part uses high-level autonomous driving applications as an illustrative case, where AI models are integrated within the driving cabin to facilitate seamless human-AI interaction.

(3) **Human comprehension**: ensuring solutions provided by models are explainable and user-friendly. This research proposes to integrate the human factor of reasoning into procedures for evaluating various explanation methodologies and for novel model explanation design to enhance human comprehension of AI.

This dissertation delves into the practical implications of integrating artificial intelligence in critical domains like autonomous driving and medical diagnostic support systems. It proposes novel perspectives to improve model capabilities through the incorporation of human expertise, to infer human intentions from actions using AI algorithms, and to provide tailored model explanations to enhance human comprehension. These contributions are beneficial in fostering an effective and reliable collaboration between humans and AI models.

Zusammenfassung

Heutzutage werden immer mehr Fortschrittliche künstliche Intelligenz (KI)-Modelle in vielen Bereichen wie Medizin, Finanzen und Bildung routinemäßig eingesetzt. Damit einhergehend wächst die Sorge, dass solche KI-Systeme Menschen in vielen Jobs ersetzen und Entscheidungen fernab menschlicher Kontrolle treffen könnten. Um diese Bedenken anzugehen, schlagen viele prominente Institute eine Strategie der Human-centered AI (HAI) vor, um sicherzustellen, dass die nächste KI-Generation technisch menschliche Intelligenz widerspiegelt, sich auf die Auswirkungen der KI auf Menschen konzentriert und Menschen eher verbessert als ersetzt.

Meine Dissertation befasst sich mit dem Aufbau von humanzentrierten KI-Modellen, die die Komplexität menschlicher Intelligenz erfassen, menschliche Bedürfnisse verstehen und menschlich nachvollziebare Erklärungen liefern können, um vertrauenswürdige und zuverlässige KI-Systeme zu entwickeln.

Konkret untersuche ich drei untrennbare Forschungsherausforderungen zur Förderung von HAI:

(1) **Menschliche Aufmerksamkeit**: Integration von menschlichem Fachwissen in die Entscheidungsfindung zur Verbesserung der Modellfähigkeit. Insbesondere schlage ich vor,
zusätzlich zu gängigen Dateneingaben Aufmerksamkeitsdaten, die auf Blickerfassung von Menschen basieren, in das Modell einzupflegen. Dies verbessert nachweislich die Modellleistung bei anspruchsvollen visuellen Aufgaben wie der Erkundung von feinkörnigen Unterscheidungsmerkmalen.

(2) **Menschliche Bedürfnisse**: Entwicklung von Modellen, die menschliche Absichten in spezifischen Kontexten besser verstehen können, um eine effektive Zusammenarbeit zwischen Mensch und KI zu erreichen. Ich verwende Anwendungen im Bereich des autonomen Fahrens als Beispiele, in denen KI-Modelle in die Fahrerkabine integriert sind, um eine nahtlose Interaktion zwischen Mensch und KI herzustellen.

(3) **Menschliches Verständnis**: Gewährleistung, dass von KI-Modellen bereitgestellte Lösungsvorschläge erklärbar und benutzerfreundlich sind. Diese Forschung integriert den menschlichen Faktor *logisches Denken* in Anwendungen zur Verbesserung des menschlichen Verständnisses von KI und in die empirische Bewertung verschiedener Erklärungsmethoden.

Meine Forschung hat bedeutende Auswirkungen auf eine Vielzahl praktischer Anwendungen, beispielsweise autonomes Fahren und medizinische Diagnoseunterstützungssysteme und liefert potenzielle Lösungen, die eine *effiziente und sichere Zusammenarbeit* zwischen Menschen und KI erleichtern und so eine Synergie erreichen, bei der das Ergebnis die Summe der Einzelbeiträge übertrifft.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence.
AIMDSS	AI-based Medical Diagnosis Support System.
AUC	Area Under Curve.
CNN	Convolutional Neural Network.
ConvLSTM	Convolutional-LSTM.
CUB	Caltech-UCSD Birds.
FC	Fully-Connected.
FLOPs	Floating point operations per second.
HA	Human Attention.
HAI	Human-centered Artificial Intelligence.
HCI	Human-Computer Interaction.
HXAI	Human-centered Explainable Artificial Intelligence.
LLM	Large Language Model.
LSTM	Long Short-Term Memory.
ML	Machine Learning.
PLDA	Probabilistic Linear Discriminant Analysis.
SGD	Stochastic Gradient Descent.
SOTA	State-Of-The-Art.
XAI	Explainable Artificial Intelligence.

1 Introduction

1.1 Research Goal

Artificial intelligence (AI) models have recently made significant strides toward *human-like* behaviors. We have witnessed advanced models such as GPT offering conversational experiences closely resembling human interactions, or robots from Boston Dynamics exhibiting human-like behaviors in their ability to interact and navigate in real-world environments. Some concerns have been raised that AI systems would become so advanced that they might replace people in many jobs and make decisions over human control [19]. To alleviate these concerns and position AI as a support for enhancing human performance instead of replacing humans, ***Human-centered AI*** (HAI) has been proposed [19, 20]. More specifically, [19] summarizes strategies of HAI from multiple prestigious institutes: HAI research should focus on developing AI technologies that can technically reflect human intelligence, studying the impact of AI on humans, and designing AI applications that augment human capabilities [21, 19]. Together, these technological innovations in the mission of HAI are redefining the boundaries of human-AI interaction and fostering new collaborative patterns.

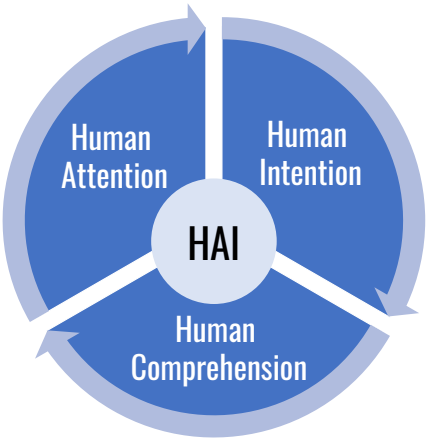


Figure 1.1: Illustration of three research challenges for addressing Human-centered AI in this dissertation.

In this dissertation, the focus is on developing AI-driven systems and exploring their interaction with human cognitive abilities. This area of study squarely aligns with the field of *Cognitive Ergonomics*. Often synonymous with "human factors engineering," cognitive ergonomics delves into understanding mental processes such as perception, memory, reasoning, and motor response, and their modulation during interactions with the various elements of the system under observation [22]. The research presented here aims to examine three fundamental **human factors** that reflect the processes of perception, reasoning, and response: human attention, related to the process of perception; human comprehension, aligning with the reasoning process; and human intentions, associated with the response mechanism.

More specifically, as shown in Figure 1.1, this dissertation incorporates human factors in the model decision-making across various *stages* organized as follows: (1) **Human**

Part I. Incorporating Human Attention
Chapter 3. Human Gaze-based Attention in Classification
(1) Human Attention in Fine-grained Classification Y Rong, W Xu, Z Akata, E Kasneci *British Machine Vision Conference, 2021*
Part II. Predicting Human Intention
Chapter 4. Driver Intention Prediction
(2) Driver Intention Anticipation based on In-cabin and Driving Scene Monitoring Y Rong, Z Akata, E Kasneci *IEEE 23rd International Conference on Intelligent Transportation Systems, 2020*
(3) Where and What: Driver Attention-based Object Detection Y Rong, NR Kassautzki, W Fuhl, E Kasneci *Proceedings of the ACM on Human-Computer Interaction, 2022*
Part III. Enhancing Human Comprehension
Chapter 5. Evaluating Model Explanations
(4) A Consistent and Efficient Evaluation Strategy for Attribution Methods Y Rong, T Leemann, V Borisov, G Kasneci, E Kasneci *The 39th International Conference on Machine Learning, 2022*
(5) Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations Y Rong, T Leemann, T Nguyen, L Fiedler, P Qian, V Unhelkar, T Seidel, G Kasneci, E Kasneci *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023*
Chapter 6. Tailoring Explanations to User Expertise
(6) I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise Y Rong, P Qian, V Unhelkar, E Kasneci *The 38th Annual AAAI Conference on Artificial Intelligence, 2024*

Table 1.1: Publications used in each chapter of this dissertation.

Attention - It can be benificial to incorporate human attention into model training and architecture design, allowing models to process *inputs* more effectively. (2) **Human Intentions** - It is essential for models to comprehend and predict human intentions in their *outputs*, facilitating a more efficient collaboration. (3) **Human Comprehension** - In the *post-decision* phase, models should offer explanations to enhance user understanding, thereby optimizing the user experience of model utilization. This dissertation explores innovative approaches to integrating human factors at each of the three stages.

More specifically, through multiple projects on human behaviors across various stages of interaction with models, it is observed that humans and AI models often utilize different attentional or reasoning mechanisms when making decisions. As a result, each possesses distinct advantages. Recognizing this, it is crucial to integrate their complementary strengths to cultivate synergistic collaboration between the two entities.

1.2 Thesis Overview

This dissertation discusses the three types of human factors in the model design based on several publications that are organized in the following three parts. Table 1.1 lists the publications relevant to each chapter. In Part I, regarding human attention, *human gaze-based attention* is studied - an important asset of human knowledge that humans acquire

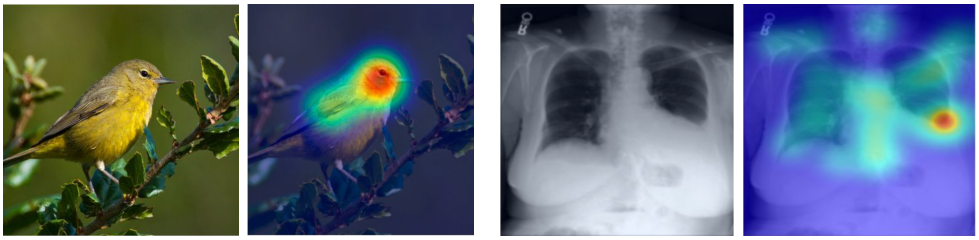


Figure 1.2: Human gaze-based attention examples on two datasets: **Left**: CUB-200-2011 [1] for fine-grained bird species classification; **Right**: Chest X-Ray [2] for diagnostic classification.

through life-long learning. To gain insight into human intentions, Part II introduces a model specifically aimed at predicting *human intended objects* within driving videos. Explainable AI (XAI) is widely used in improving human comprehension in AI models. Part III concentrates on developing efficient methods to evaluate model explanations, including automatic and human-grounded evaluations on XAI techniques. Furthermore, it contains a case study in the medical field for examining human feedback on model explanations, with the goal of refining these explanations through user feedback.

1.2.1 Human Attention

The Human factor *perception* is a crucial aspect of how we interact with our environment and the tools we use. Specifically, human visual perception plays an important role in solving different tasks. Human attention acts as a filter for the vast amount of visual information our eyes perceive, allowing us to focus on what is most relevant in our surroundings.

Part I explores human attention, aiming at improving AI models in their processing inputs. Human visual attention has been the subject of extensive study for many years, including fields such as cognitive psychology and neuroscience [23]. When encountering many objects at once, humans focus on task-relevant objects. The essential role of human attention mechanisms in efficiently selecting relevant objects for tasks in a controlled, top-down approach is well acknowledged [24, 25, 26]. The pivotal and unique contribution of human attention in resolving visual tasks has sparked interest in its study within artificial intelligence research, for example, as noted in [27]. This is evident in numerous computer vision applications that incorporate human gaze data. Such applications include classification tasks [28, 29], systems aiding in medical diagnosis [2, 30], and the selection or cropping of significant objects in images and videos [31, 32, 33, 34].

These mechanisms are visualized generally through the application of a Gaussian filter on fixation points, creating a *saliency* map [35]. Figure 1.2 illustrates human gaze-based attention in two complex classification tasks. In these saliency maps, regions highlighted in red denote areas of concentrated human attention. Such a visualization highlights how human attention is directed towards important features in decision-making processes, such as identifying bird species or diagnosing diseases. By utilizing human attention

Figure 1.3: Illustration of driver intention prediction. In this video, the car slows down at a cross and turns right. The prediction is made after every second. If the prediction is correct, there is a ✓. (Frames from Brain4cars [3].)

saliency maps, Part I demonstrates the power of human attention in processing and classifying images, which can enhance the capability of deep learning models.

Our work addresses this research gap and the hypotheses that (1) human attention focuses on essential features for solving the task (e.g. fine-grained classification); (2) using human attention also allows improving model performance in accomplishing the task. To validate the first hypothesis, we first capture and present human attention in the style of a saliency map. We compare the regions that human attention covers with the ones that are discovered by the model (model explanation), and show that human attention hints on the regions that are more discriminative in the classification. We propose two modules which make use of the essential features revealed by human gaze to validate the second hypothesis: we use Gaze Augmentation Training (GAT) to train a better classifier and a Knowledge Fusion Network (KFN) to integrate the human attention knowledge into models.

Our contributions are as follows: (1) We collect human gaze data for the fine-grained data set CUB, enhance it by incorporating human attention and coin this new dataset as CUB-GHA (Gazed-based Human Attention). For this novel dataset, we also validate the efficiency of human gaze data in discovering discriminative features. (2) We propose two novel modules to incorporate human attention knowledge in classification tasks: Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN). (3) To showcase the relevance of our work for highly relevant applications, we evaluate our methods not only on our novel CUB-GHA dataset, but also on chest radiograph images from a recently released dataset CXR-Eye (which contains also gaze data). Our work shows that human attention knowledge can be successfully integrated in classification models and help improve the model performance with regard to the state-of-the-art in different classification tasks.

1.2.2 Human Intention

The second human factor studied in this dissertation is *reaction*, which refers to the way humans respond to various stimuli in their environment. The studied human reactions in this work are physical behaviors. In the context of advanced **human-AI collaboration**, understanding these human behaviors can better assist humans in solving a task together, i.e., the AI system is able to predict human intentions.

By analyzing data from various sources such as past human actions, AI models are designed to anticipate human needs and intentions. This predictive capability is particularly important in areas like autonomous driving, where each individual driver's needs should be considered. Part II presents models designed to understand human actions and then predict human intentions. Figure 1.3 illustrates the intention prediction in the context of autonomous driving, concretely, the prediction is made based on in-cabin and outside videos. It predicts human maneuver intention at each second based on videos before that second. Intuitively, the outside video, i.e., the scene perspective, should be very informative and provide information that the inside video does not convey. Therefore, our work aims (1) to extract the vehicle motion information from the traffic videos effectively and improve the results that only used one video stream; (2) to propose an end-to-end method without using manual encoding information, and (3) to keep the model as light-weighted (less parameters) as possible to offer applicability to resource-limited mobile platforms.

To approach these aims, we propose a deep learning framework, which combines the information from the driver monitoring videos with the outside view. In our framework, a ConvLSTM [36] based encoder (shown in the upper branch) extracts the motion information, which is interpreted in optical flow images. Meanwhile, the 3D ResNet-50 (shown in the bottom branch) acquires features from the driver video. The motion decoder for outside motion features is integrated into the classifier. This novel classifier leverages features from both sides, i.e., driver and scene, jointly to produce a maneuver anticipation.

To gain a deeper insight into human intentions beyond human actions, it is important to identify not just *where* a person focuses, but also *which* object is in the area of attention, a concept known as gaze-object mapping [37]. This understanding is vital in numerous research endeavors, particularly in studying the learning processes of students [38] and examining human cognitive functions [39]. In the context of driving, human drivers utilize their gaze-based attention to identify crucial objects and take informed actions while driving. Human intention prediction by detecting the objects of the human's intention is also studied in Chapter 4. As demonstrated in Figure 1.4, a model predicts the objects the driver focuses on. This offers a clear insight into driver intentions, i.e., the driver cautiously passes the cars in the right lane. This intention cannot be easily observed if all objects are detected using a standard object detector. To bridge the research gap between driver gaze prediction and semantic object detection in the current research landscape of autonomous driving applications, we propose (1) to predict where and what the drivers look at. Furthermore, we aim (2) at a model that is efficient in computation, since resources on self-driving cars are limited. Specifically,

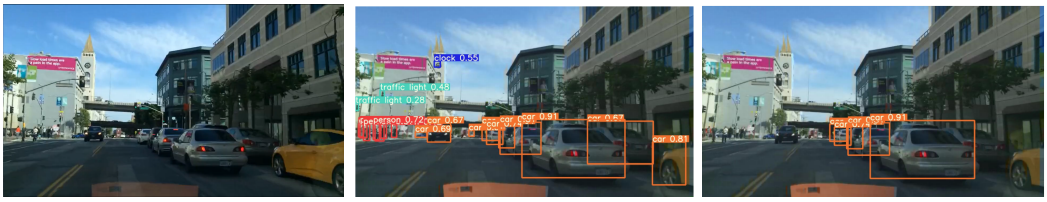


Figure 1.4: Driver attentive object detection: **Left**: This frame, sourced from the BDD-A dataset [4], depicts a vehicle approaching a crossroad and overtaking stationary vehicles in the right lane. **Middle**: Objects detected using YOLOv5 [5]. **Right**: Objects of the driver's intention, suggesting the driver intends to cautiously pass the cars on the right lane.

we designed a novel framework for efficient attention-based object detection based on human driver gaze. Our approach provides not only pixel-level attention saliency maps, but also the information of objects appearing in attention areas. A feature encoder is first used in our framework to encode the information in the input image. Then, the extracted features are used to predict gaze and detect objects in the image at the same time. Since obtaining accurate high-level (object) information is our final goal, instead of low-level (pixel) accuracy in saliency map prediction, we predict salient areas in a grid-based style to save computational costs while still maintaining high performance in the critical object detection task.

1.2.3 Human Comprehension

The third studied human factor in this dissertation is *reasoning*. The reasoning process in AI model design corresponds to generating model explanations, which can be addressed by using Explainable AI (XAI) methods. XAI is increasingly recognized as an indispensable component in the field of AI research. It plays a crucial role in enhancing human understanding of AI models, which is essential for their acceptance, especially in high-stakes applications.

Creating XAI applications is a complex task because the quality of an explanation does not only depend on the AI model itself, but largely on how the person receiving it perceives and understands it. One primary research challenge existing in XAI is the misalignment between the technical methods in XAI and the actual goals of users in practical applications [40]. Multiple studies have been unable to demonstrate that incorporating XAI elements always enhances user performance in real-world tasks that involve AI assistance (refer to Table 5 in [41]). Moreover, the effectiveness of an explanation relies on the human's background knowledge and their purpose for seeking the explanation, as well as various other human factors. Much of current XAI research does not focus adequately on the user receiving the explanation, often producing *one-size-fits-all* explanations that may not suit each individual user's needs. In summary, the development of XAI should adopt a human-centric approach. This approach emphasizes meeting the individual needs for explanation of people and measuring success through user interaction experiences.

To advocate **human-centered XAI**, this dissertation considers humans in both model explanation evaluation and design, which are addressed in the two chapters in Part III. This chapter first addresses the drawbacks of automatic evaluation. Evaluation strategies proposed to compare different attribution methods commonly follow an ablation approach by perturbing the input features, e.g., image pixels, deemed most or least important. Specifically, perturbing pixels assigned high importance should decrease predictive quality whereas perturbing unimportant pixels, should hardly affect the predictions. These measures aim to capture the *fidelity* of explanations [42], i.e., how well the explanation genuinely reflects the prediction of the underlying model. Fidelity based on a single data sample is known as local fidelity, while global fidelity is measured on the whole data set [42]. The outcome of evaluation strategies is highly sensitive to parameters such as the perturbation function and order. Such removal strategies often lead to highly contradictory results depending on the order chosen, i.e., *most relevant pixels first* or *least relevant pixels first*. For instance, local attribution methods that seem to perform well in one order may perform rather poorly in the other [42, 43, 13]. This inconsistency makes it hard for researchers to impartially compare between different attribution methods and it is not well understood where the inconsistencies stem from. Moreover, for conducting the global fidelity check, a retraining step is required by some methods [13], which is prohibitively expensive in practice [42].

Chapter 5 then addresses the human-grounded evaluation in XAI, highlighting the importance of incorporating human feedback in the XAI development cycle. Many functionally-grounded measures have been proposed to evaluate XAI algorithms (see [44] for review), however, the difficult comparability between different automatic evaluation measures is a common problem [45]. Another drawback of automated measures is that there is no guarantee that they truly reflect humans' preferences [46, 47]. Consequently, user studies in XAI, especially when moving towards real-world products, are inevitable if one wishes to test more general beliefs of the quality of explanations [48]. However, only a small portion (about 20 %) of XAI evaluation projects consider human subjects [44]. There exist efforts in developing taxonomies or introducing the definitions or implications of different human-centric evaluations [49, 50, 51], but the recent generation of user studies and their findings have not been systematically discussed yet. Moreover, Yang et al. [52] point out that XAI is growing separately and treated differently in different communities (e.g., machine learning and HCI). Hence, effective guidance in XAI user study design is crucial to better let both XAI algorithm and application designers recognize the users' real needs. This work aims to bridge this research gap in modern XAI user study design by distilling practical guidelines for user studies through a comprehensive and structured literature review.

Chapter 6 presents an algorithm that improves human understanding of models by providing explanations based on each user's expertise. Specifically, Chapter 6 introduces a novel framework named I-CEE, which provides **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise. Differing from previous approaches, I-CEE considers the user's expertise in reasoning when selecting example images, thus tailoring the examples for each user, as illustrated in Figure 1.5. This chapter represents a significant advancement towards *personalized* model explanations. We advocate that human modeling is critical

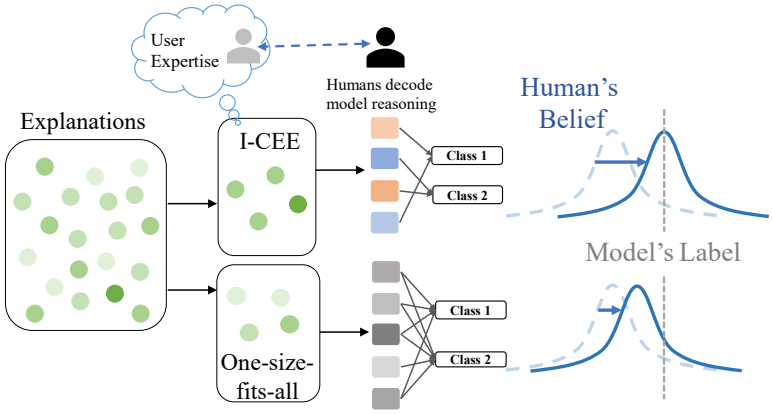


Figure 1.5: I-CEE tailors the explanation process to each user by considering their expertise. By selecting the most informative explanations based on estimated user expertise, I-CEE can improve the user's understanding of the ML model's decision.

to XAI research because explainability is inherently centered around humans [40]. A few works focusing on explaining reinforcement learning policies use cognitive science theories to model the human user and generate explanations based on the human model [53, 54, 55, 56]. Closer to our focus, the works of [57] and [58] utilize a Bayesian Teaching framework to model human perception and then generate human-centered explanations. One limitation of these works is that all human users are treated the same by the modeling method, presuming that an identical set of explanations will work for *all* users. In contrast, we attempt to generate tailored explanations for each user by modeling their *task-specific expertise*. Our approach to modeling user expertise is informed by human annotator models used in active and imitation learning [59, 60]. Similar to these works, our user model aims to capture both the decisions and reasoning process (expertise in concepts used for image classification) of the human user in the context of a given classification task.

1.3 Research Contributions

This thesis presents significant research contributions across various key areas, as elaborated in each respective section.

1.3.1 Exploration using Eye Tracking

In my doctoral studies, I used eye tracking as a tool for exploring human attention mechanisms in solving computer vision tasks. We deployed an image comparison game to collect human gaze-based attention in fine-grained classification, in which participants are prompted to focus on distinct features while comparing two similar images from different categories. This task is intentionally made difficult to yield more meaningful insights, by selecting two very alike classes for each comparison pair. We gathered hu-

Figure 1.6: CUB-GHA (Gaze-based Human Attention) dataset. **Left**: a static aggregated heatmap; **Right**: a sequence of fixation heatmaps for human attention.

man gaze data on the CUB fine-grained classification dataset, creating the **CUB-GHA (Gaze-based Human Attention) dataset** [1]. With the help of such a dataset, we can (1) analyze and contrast the critical features employed in decision-making processes by both humans and models; and (2) study the efficacy of human attention in the context of fine-grained classification. In **CUB-GHA**, the complete sequence of fixations from each participant for an image is accessible. Users have the option to utilize either the aggregated static attention maps for an image or a sequence of gazes to analyze scan paths (exploration by each participant). An illustration of the two types of gaze data offered in CUB-GHA is presented in Figure 1.6.

We have also supplied scripts for processing heatmaps, which are adaptable for use with other data sources. Our dataset presents possibilities for exploring the integration of eye tracking with existing comprehensive annotations, including textual explanations, attributes, and bounding boxes. This enables researchers to evaluate various applications where human gaze is essential in machine interaction.

1.3.2 XAI Evaluation Guidelines

In the context of the growing field of XAI, ensuring fair and effective evaluation of these techniques is crucial. In my thesis, I explored two widely recognized types of evaluation metrics: automatic and human-grounded evaluations. One of the most popular automatic evaluation metrics is *fidelity* (or *faithfulness*). We studied the existing biases in one of the most popularly used evaluation frameworks called "ROAR" (RemOve And Retrain) [13] using an information-theoretic analysis. Drawing on our theoretical understanding, we introduce a new evaluation framework named **ROAD** (Remove and Debias). Our framework provides two key advantages compared to ROAR: (1) it reduces the influence of confounders, leading to greater consistency across evaluation methods, and (2) ROAD eliminates the need for resource-intensive retraining, thereby cutting computational expenses by up to 99%. Our algorithm's source code has been made publicly available and is now integrated into "*Quantus*", a toolkit to evaluate neural network explanations [2]. This demonstrates that our evaluation algorithm has effectively responded to the challenges faced by researchers and practitioners in assessing XAI techniques.

[1] CUB-GHA can be found at `https://github.com/yaorong0921/CUB-GHA`.

[2] Quantas can be found at `https://github.com/understandable-machine-intelligence-lab/Quantus`.

In addition to conducting automated evaluations of XAI, I ventured into human-grounded evaluation metrics and took the lead on projects aimed at emphasizing the significance of incorporating human elements into the evaluation of XAI techniques. After a comprehensive study of approximately one hundred state-of-the-art works focused on user studies for XAI, we distilled the best practices and formulated guidelines for the design of user studies for XAI algorithms. Furthermore, we emphasized potential research directions for human-centered XAI and advocated for collaborative efforts across different communities. Our guidelines have garnered recognition from numerous researchers with diverse expertise backgrounds. As one researcher noted,"*your paper does a commendable job of triaging multiple threads of work.*"

1.3.3 Deployment in Practical Applications

Throughout my doctoral research, I consistently prioritized practical applications and emphasized its importance in assisting humans in completing various tasks. I mainly studied two primary application domains: **autonomous driving** and **healthcare**. Within driving applications, I collaborated with automotive companies, gaining valuable insights into the industry's specific requirements and demands. Concretely, I worked closely with *Continental*, a German multinational automotive parts manufacturing company, in AI-driven in-cabin applications. For instance, I designed an algorithm to predict driver intentions by analyzing video footage from both inside the vehicle cabin and the surrounding road traffic. Additionally, we conducted research on the trust and comfort levels of drivers and passengers within the driving cabin, particularly in high-level autonomous driving scenarios. This involved exploring methods to measure these factors and enhancing them through the design of a co-drive assistant. During the collaboration with *Horizon Robotics*, a Chinese company specializing in autonomous vehicle development, we developed an effective feature fusion approach for LiDAR 3D object detection. This project provided a valuable lesson in the importance of algorithmic *efficiency* in real-world applications.

I collaborated closely with radiologists, gaining valuable insights from diverse areas of expertise. Together, we worked on a project aimed at designing a medical diagnosis support system, which provided me with the opportunity to understand their perspectives on working with AI. For example, while model explanation algorithms produced faithful saliency maps that aligned with the model's decisions, medical professionals still found it challenging to comprehend the underlying reasoning of the model. Some radiologists expressed a preference for a different *form* of explanations rather than saliency maps. This project highlighted the significance of understanding the specific requirements of experts from various domains and assisting them in effectively integrating AI models into their workflows.

1.3.4 Publication List

My doctoral studies resulted in several publications, listed below. The papers included in this dissertation are highlighted in blue. The papers marked with underline are the core publications (led by me) for this dissertation:

- **Rong, Y.**, Qian, P., Unhelkar, V., & Kasneci, E. (2023)
  I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise
  Pre-print. (To appear at the 38th Annual AAAI Conference on Artificial Intelligence (AAAI)).
- **Rong, Y.**, Leemann, T., Nguyen, T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2023)
  Towards Human-centered Explainable AI: User Studies for Model Explanations
  IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)
- Leemann. T., **Rong, Y.**, Nguyen, T., Kasneci, E., & Kasneci, G. (2023)
  Caution to the Exemplars: On the Intriguing Effects of Example Choice on Human Trust in XAI
  XAI in Action: Past, Present, and Future Applications
- **Rong, Y.**, Wei, X., Lin, T., Wang, Y., & Kasneci, E. (2023)
  DynStatF: An Efficient Feature Fusion Strategy for LiDAR 3D Object Detection
  In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*
- Leemann, T., Kirchhof M., **Rong, Y**, Kasneci E., & Kasneci, G. (2023)
  When are Post-hoc Conceptual Explanations Identifiable?
  In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*
- **Rong, Y.**, Leemann, T., Borisov, V., Kaneci, G., & Kasneci, E. (2022)
  A Consistent and Efficient Evaluation Strategy for Attribution Methods
  In *Proceedings of the 39th International Conference on Machine Learning (ICML)*
- **Rong, Y.**, Kassautzki, N.-R., Fuhl, W., & Kasneci, E. (2022)
  Where and what: Driver attention-based object detection
  In *Proceedings of the ACM on Human-Computer Interaction (PACMHCI)*
- **Rong, Y.**, Castner, N., Bozkir, E., & Kasneci, E. (2022)
  User Trust on an Explainable AI-based Medical Diagnosis Support System
  TRAIT at Conference on Human Factors in Computing Systems (CHI-TRAIT)

- **Rong, Y.**, Xu, W., Akata, Z., & Kasneci, E. (2021)
  Human attention in fine-grained classification
  In *2021 British Machine Vision Conference (BMVC)*
- **Rong, Y.**, Han, C., Hellert, C., Loyal, A., & Kasneci, E. (2021)
  Artificial intelligence methods in in-cabin use cases: A survey
  IEEE Intelligent Transportation Systems Magazine
- **Rong, Y.**, Akata, Z., & Kasneci, E. (2020)
  Driver intention anticipation based on in-cabin and driving scene monitoring
  In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*

2 Related Work

This chapter reviews the state-of-the-art (SOTA) methods related to the proposed algorithms for solving the three research challenges discussed in Table 1.1. In Section 2.1, how human attention is used in AI models and is compared with model attention will be discussed. In Section 2.2, human intention prediction in the context of driving will be introduced. This section contains the methods used in driver maneuver prediction using videos. Moreover, gaze-based prediction methods, for instance, gaze-based attention and gaze-object mapping, are introduced. Finally, in Section 2.3, recent research on how humans understand black-box AI models via XAI is examined. More specially, this section includes recent works regarding how to evaluate model explanations using human feedback, and how human reasoning is addressed in model explanation generation.

2.1 Human Attention in AI

In this section, gaze-based human attention in AI-based applications will be first introduced, followed by attention in neural networks which can be divided into two parts: learnable and post-hoc attention [61]. Additionally, this section focuses on comparing machine (post-hoc) attention to human attention.

2.1.1 Gaze-based Attention in AI

Gaze data can provide insights into human attention, and is frequently utilized for analyzing attention patterns. It has emerged in many branches of AI because of its effectiveness and irreplaceability in many tasks [27]. Recent advancements in hardware technology have enabled precise tracking of eye movements across various activities. This includes human-computer interactions [62, 63], as well as more dynamic tasks like driving [64, 65] and robotics [66, 67, 68]. The processing of visual information can also provide insights into a person's cognitive strategies or skill levels [69]. In the medical field, gaze data has shown potential in augmenting AI models for disease classification, for example, recognizing Pneumonia and Congestive Heart Failure [2].

In computer vision, the application of gaze data is diverse and valuable [34, 31, 70, 61]. For instance, research in [70] gathers gaze data (like coordinates and duration) for 60 bird classes to assist in zero-shot learning. Another study, [61] demonstrates that human attention data outperforms an attention module in generating attention maps. Additionally, [34] introduces a system for photograph cropping using fixation data to highlight key content areas. Eye tracking is also instrumental in identifying principal objects in videos [31].

2.1.2 Attention in Neural Networks

Learnable Attention. Many studies [71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81] have successfully incorporated attention mechanisms into neural networks. These mechanisms are essential for identifying key areas in fine-grained classification tasks, thereby enhancing model performance. A group of these works [75, 77, 73, 76, 78] utilize the Recurrent Attention Model (RAM) [82]. In this RAM module, an attention agent predicts the locations of critical regions, and the classifier is trained based on these identified areas. This module employs reinforcement learning to train the attention agent, addressing the non-differentiability issue caused by the cropping operation. However, this approach is complex and computationally demanding. In contrast, other studies [71, 72, 74, 79, 80, 81] have developed attention modules that leverage outputs from intermediate network layers to enforce the model attending to significant features.

Post-hoc Attention. Post-hoc attention refers to the attention maps that are generated using various strategies from neural networks that have been fully trained [83]. These strategies are *model explanation* techniques. Recent work has looked at attribution-based machine explanation methods that perform in the same way as the human gaze. For example, in [84, 85], the impact of input image features was quantified by attaching importance to pixels. Such a quantification was also performed at the level of super-pixels in [86]. In addition to attribution-based methods, there are many other types of model explanations. For example, gradient-based methods [85, 87, 88], on the other hand, back-propagate the gradient for model prediction to the activation maps [88, 85], input image features [11, 89], or the biases [87]. Perturbation-based methods determine the importance of features by measuring the model prediction after perturbing these features as in [86, 90]. In prototype-based explanations, image samples [91, 92] or local image patches [84, 93] are treated as prototypes and attribute the model decision to these previously learned prototypes. In another approach, concept-based explanations identify higher-level, human-understandable concepts that are important to the model prediction [94, 95].

2.1.3 Comparison Between Human and Model Attention.

There is, to date, little research investigating the relationship between human attention and machine-generated attention [96, 97, 61]. More specifically, Das et al. [96] collect human attention data by asking Amazon Mechanical Turk (AMT) annotators to mouse-click the important regions in a blurry image in order to answer a visual question. Researchers come to the conclusion that the VQA model attention [98, 99] does not focus on the same regions as humans do. In a recent work by Sen et al. [97], the authors collect binary human attention by asking annotators to click on important words for a review (text) classification task and compare human attention with soft machine attention. However, this human attention data is collected in an uncontrolled environment on AMT, which may not be fully reliable [100]. Recently, [61] uses human gaze data as human attention and compares it to learnable model attention. However, the human

and model attention in [61] are also collected in different experimental settings. More specifically, human attention is collected for a 2-way classification, while model attention for a 60-way classification.

2.2 Driver Intention Prediction

This section discusses the prior works in human intention prediction. Given that the case study centers around autonomous driving, the discussion will primarily concentrate on intention prediction in driving contexts. Specifically, this section narrows down to two primary types of intentions: driver maneuver behavior and gaze attention. This section begins with exploring prior works in predicting driver intentions using machine learning models based on driver monitoring videos. Then, it discusses gaze-object mapping within various applications, establishing the groundwork for the innovative aspects of the proposed method in this thesis. In the end, current driver gaze-based attention prediction models will be introduced.

2.2.1 Maneuver Behavior Prediction.

In progressing towards fully autonomous driving, it is crucial to enhance current Advanced Driver Assistance Systems (ADAS) for effective cooperation with human drivers. Consequently, accurately predicting the driver's intentions is essential to offer them optimal assistance. Driver maneuvers can be inferred from related behaviors like glancing at mirrors or windows. As such, techniques from human action recognition have been effectively employed in this area. A key focus of recent studies has been on predicting a driver's maneuver intentions prior to their execution. Notably, the Brain4cars [101] and Honda Research Institute Driving Dataset (HDD) [102] datasets were created specifically for studying driver behavior. HDD, for instance, [102], utilizes three high-definition cameras, GPS, LiDAR sensors, and vehicle CAN-Bus data to capture traffic scenarios. Brain4cars [101] includes both internal and external car videos, GPS data, and vehicle dynamics. These videos offer insights into various driver behaviors and traffic conditions. The literature suggests that driver intentions can be predicted from video analysis, particularly noting how drivers check side mirrors. Studies using the Brain4cars dataset, such as [101, 3, 103, 104, 105], have successfully predicted maneuvers. While these results are promising, there are aspects that warrant further examination. Primarily, previous research in driver maneuver prediction has largely relied on video observations of the driver. Research indicates that driver behavior, particularly eye movement, is not only useful for activity recognition [106, 107] but also vital for ensuring safe control transitions in semi-autonomous driving [65]. The following paragraph will delve into predictions based on driver gaze and attention.

2.2.2 Gaze-Object Mapping

Earlier research [108, 109] aims to release the burden of manual labeling by employing gaze-object mapping. This technique labels objects at the fixation point, essentially

marking the object being observed. A widely used method involves checking if a fixation falls within the object's bounding box as predicted by an object detector based on deep neural networks [38, 37, 110], such as YOLOv4 [111]. Wolf and colleagues [108] recommend using Mask-RCNN [112] for object segmentation to detect object areas. These studies train their object detectors with a limited range of object data and classes for annotation. Conversely, Panetta et al. [39] opt for a bag-of-visual-words classification model [113] instead of deep neural networks, due to a lack of sufficient training data.

Barz et al. [114] introduce a "cropping-classification" approach where a small area around the fixation is cropped and classified using a network pre-trained on ImageNet [17]. This method from [114] is applicable in Augmented Reality for enhancing cognition-aware mobile user interactions. In subsequent research [37], the authors evaluate mapping algorithms based on image cropping (IC) against those using object detectors (OD), assessing metrics like precision and recall. Their findings indicate that while IC achieves higher precision, it has a lower recall rate compared to OD.

There has been little effort in the previous works about gaze-object mapping for autonomous driving applications, due to the need for a remote eye tracker to identify the objects being focused on. However, this is a useful feature for semi-autonomous driving cars, i.e., a model acts like a "second driver", by providing safety alerts on critical traffic objects that human drivers might miss. In the context of fully autonomous driving, where human driver fixation data is unavailable, a model is required to replicate human driver's fixation patterns.

2.2.3 Gaze-based Attention Prediction

The surge in interest in (semi-)autonomous driving has led to a heightened focus on understanding and predicting human drivers' attention. Recent studies have shown advancements in simulated driving scenarios by employing driver gaze in training models end-to-end, enabling models to perceive traffic similarly to human drivers [115, 116]. Leveraging created real-world datasets like DR(eye)VE [8] and BDD-A [7], several deep neural networks have been developed to predict drivers' gaze maps on a pixel-wise level, such as those in [8, 7, 117, 118, 15]. For instance, the DR(eye)VE model [8] employs a multi-branch architecture focusing on color, motion, and semantics, while the BDD-A model [7] utilizes AlexNet [119] features processed through convolutional layers and a convolutional LSTM to predict gaze maps. Attention models are also being used for predicting driver saliency maps to aid braking decisions, as seen in [120]. Additionally, networks like ML-Net [9] and PiCANet [10], known for general saliency prediction, are well-performing. ML-Net combines features from various CNN levels for saliency prediction, while PiCANet is a contextual attention network that selects informative context locations pixel-wise for more accurate saliency maps.

Apart from these gaze map prediction networks, other models extend to foresee additional driving-relevant areas. Deng et al. [121] utilize a convolutional-deconvolutional neural network (CDNN) trained with eye tracker data from multiple individuals, whereas Pal et al. [15] suggest incorporating distance-based and pedestrian intent-guided seman-

tic information into the ground-truth gaze maps. This semantic enhancement is used to train models, thereby augmenting them with semantic knowledge.

2.3 Human factors in Explainable AI

In this section, XAI works with a focus on human subjects in the cycle will be introduced. Concretely, the first section discusses the importance and current methods of human-grounded evaluation of XAI methods. Then, prior work considering human individual needs (expertise) is discussed, aiming at human-centric XAI methods.

2.3.1 Functional-grounded Evaluation

With the growing number of attribution methods, various scholars have presented desiderata that explanations should fulfill. [122] consider two subcategories in this field, namely *human-grounded* metrics relying on human judgment and *functional-grounded* metrics. The latter do not require a human-generated ground truth that can be hard or even impossible to obtain. Metrics of this type frequently rely on the idea that if the most important part of the image is changed, the output probability of the given black-box model should also change in return. Examples include the Sensitivity-n measure proposed by [123] and the infidelity and max-sensitivity metrics by [124]. [125] and [126] also propose to perturb the pixels in the input image according to the importance scores. However, [13] show that the perturbation introduces artifacts and results in a distribution shift, putting these no-retraining approaches in question. They propose the Remove and Retrain (ROAR) framework with an extensive model retraining step to adapt to the distribution shift. Therefore, we distinguish between evaluation methods with *retraining* and *no-retraining* approaches.

Only few papers have used and compared different evaluation strategies for attribution methods and a sound theoretical explanation for the differences between them is still missing. [127] assesses different baselines for feature attribution applying the Integrated Gradient method [11]. They also observe that changing the hyperparameter settings can lead to varying results. [43] draw the same conclusion for attributions on tabular data. [42] compute the consistency among different, no-retraining evaluation strategies and report an alarmingly low agreement. In this work, we conduct a rigorous analysis of reasons for existing inconsistency and provide a solution to reduce it, which is not studied in previous works. Moreover, our solution also reduces high computational costs caused by retraining.

2.3.2 Human-grounded Evaluation

AI's success story has not excluded high-stakes decision-making tasks like medical diagnosis [128, 129, 130, 131, 132], credit scoring [133, 134, 135, 136, 137], jurisprudence [138, 139, 140, 141] or recruiting and hiring decisions [142, 143, 144, 145], influenced by data-driven algorithms. Nonetheless, the operational mechanisms and decision-making

methods of contemporary AI systems often remain opaque, leading to their characterization as a "black box"'. The utilization of such opaque models in critical safety areas, like public health or finance, poses a significant challenge [146]. This stems from the imperative need for AI systems that are both transparent and trustworthy, catering to the requirements of professionals (for enhanced understanding of the system's operations) and users (for dependable reliance on the model's decisions).

While a huge number of model explanations are available, the question of how to evaluate their quality is still an open research question, and, hence, has been extensively studied in recent years. Evaluating and comparing different methods of explanation in XAI research is challenging due to the multidisciplinary nature of interpretability and explainability [44, 122, 147]. This evaluation is categorized into human-grounded measures, involving human subjects and functionally-grounded metrics, which do not require human involvement [122, 44]. There is a growing interest in developing automatic evaluation methods for explanations, as detailed in a comprehensive review focused on functionally-grounded evaluation methods [44]. The inherent human-centric nature of explainability has led to a recognized need for human-centered evaluations in XAI research [122, 40].

Various studies contribute to advocating the usage of human-grounded evaluation. Chromik and Schuessler [50] propose a taxonomy for XAI evaluations that involve human subjects. Mohseni et al. [51] categorize human-related evaluation metrics into four groups: mental model, user trust, human-AI task performance, and explanation usefulness and satisfaction. Hoffman [49] focuses on psychometric evaluations, proposing a conceptual model for the XAI process and identifying key components for evaluation. The broader application of XAI, aimed at supporting decision-making and benefiting end-users, is discussed in [148], including studies on collaborative human-AI decision-making. Ferreira and Monteiro [149] examine the user experience of XAI applications, exploring who uses XAI and in what context. Liao et al. [40] focus on user studies in XAI revealing the pitfalls of existing XAI methods, emphasizing the important role of humans in XAI development. Doshi-Velez and Kim [122] highlight the need for sophisticatedly designed human-subject experiments to reduce confounding factors, further underlining the complexity of human involvement in XAI evaluation.

2.3.3 User-centric XAI design

XAI has acknowledged the importance of human involvement in understanding explanations, which has led to more frequent use of human-centered approaches for evaluating explanation methods, as discussed in the last section. In addition to these assessments, some methods have incorporated human elements into the creation of explanations [150, 55, 54, 56, 57]. These studies aim to explain reinforcement learning policies that employ theories from cognitive science for constructing models of the human user. The most relevant user-centric XAI method among these approaches is the Bayesian Teaching framework proposed in [58]. It is applied in image classification, which chooses explanations by considering users as Bayesian agents. However, this approach does not account for the variance in users' thought processes or their prior knowledge.

In contrast, the proposed method in Chapter 6 addresses the distinct explanatory needs of various users. This design is mainly shaped by insights from *active learning*. The concept of active learning focuses on optimizing model accuracy with minimal labeling effort [151, 152]. This approach is particularly beneficial in situations where labeled training data is scarce. Active learning has found applications beyond mere classification, extending to areas like sequence labeling [153] and image semantic segmentation [154]. While the objective of active learning is to enhance model training, its methodologies offer significant insights for XAI, which is geared towards teaching humans about AI models.

2.4 Current Research Gap

After reviewing state-of-the-art works for each topic, this section summarizes the main messages from previous works and the existing current research gaps, which will be addressed in this dissertation.

Incorporating Human Attention. Previous works utilize attention modules that are inspired by human attention mechanisms to improve the model performance. Many works also compare human gaze-based attention with machine attention and conclude that machine attention is different from human attention. However, the following two research questions have not been fully addressed yet:

- Does human gaze-based attention discover more effective features for solving the visual task than a model does?
- How can we incorporate human attention in the model training procedure to improve model perception capability?

These research questions are studied in Chapter 3.

Predicting Human Intention. This part focuses on two applications in the autonomous driving domain: to predict driver maneuvers and to predict objects that drivers intend to interact with. For the former application, previous works utilize mainly videos of drivers but only leverage manually encoded information from the video outside of the cabin. These methods are not practical and can be improved by distilling more useful information from the outside videos. Therefore, Chapter 4 (Section 4.2) aims to propose:

- A model that utilizes driver and driving scene monitoring videos to precisely anticipate driver maneuver intentions.

For the latter challenge, current methods can only anticipate driver attention and the computation costs are high. Therefore, Chapter 4 (Section 4.3) aims at:

- A model that can predict "where" and "what" the driver is focusing on in a resource-efficient manner.

Enhancing Human Comprehension. As introduced in the last section, human subjects are considered in previous works when evaluating the effectiveness of different XAI methods and including human reasoning factors to inform model explanation generation. However, there is no consensus for evaluating XAI methods, as there is no ground truth for model explanations. Chapter 5 therefore bridges the current research gap to a fair evaluation of XAI methods via:

- Theoretical analysis of the bias term in the automatic evaluation and corresponding mitigation solution;
- Guidelines in conducting human-grounded evaluations to avoid pitfalls.

Effort has been made to generate model explanations based on human reasoning. However, the state-of-the-art XAI methods do not consider different user needs. Chapter 6 aims to highlight the importance of considering individual human expertise by:

- Developing a novel framework for tailored explanations for explaining decisions made by image classification models.

Part I

Incorporating Human Attention

AI models process input data to extract important features, similar to human perceptual abilities. This dissertation begins by exploring the initial step in decision-making, focusing on the design of AI models that are inspired by human attention mechanisms.

This part is adapted from the work that was published in BMVC 2021 [28]:

- Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021)

  Human Attention in Fine-grained Classification

  In *2021 British Machine Vision Conference (BMVC)*

Motivation. As discussed in Section 1.2.1, humans utilize a top-down attention mechanism during problem-solving, i.e. when operating in a goal-driven manner. Inspired by this, [155] designs a top-down attention module for models and shows its effectiveness in automated image caption generation and visual question answering. However, it is not clear whether humans are able to discover more efficient features for solving these tasks than models. The motivation of this work is to study the features discovered by human attention in challenging classification tasks. Moreover, integrating human attention knowledge into models is a non-trivial problem. This work also aims at designing a novel algorithm to effectively integrate this knowledge into models and thus improve the model performance.

Principal Methodology. To capture human attention, an eye-tracker is used to record eye movements, which are then used to serve as a basis to discover important features through an image comparison game on a fine-grained classification task. The collected human gaze data is processed and transformed into *saliency maps*. To verify the effectiveness of human saliency maps, an insertion evaluation framework is used. Moreover, human saliency maps are also compared to model post-hoc attention saliency maps in this experiment. To effectively incorporate human attention knowledge into classification models, two novel approaches are introduced in this work: Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN). GAT uses human-focused areas to augment the data and guides the model to focus more on these effective features. KFN uses an architecture that integrates the feature embeddings learned from human saliency maps to the original features, which combines the features from both branches to inform final decisions.

Main Findings. Experimental results highlight the efficacy of human attention in identifying features that are essential for classification: using just 5% of the image where humans focused on and masking the rest yielded model performance with an accuracy of 81%, while using the entire image achieves an accuracy of 85%. These features are also more effective than the ones discovered by the model post-hoc attention. For instance, the model trained with model attention-modified images only reaches an accuracy of around 70%.

Using the proposed approach, GAT and KFN, the model performance is improved significantly on two datasets. For instance, when applying GAT and KFN on vanilla

ResNet-50 on the bird fine-grained classification dataset, the performance is improved from 85.6% to 88.7%. Remarkably, GAT is a data augmentation strategy and it can easily plug into any other complex model architectures, further improving the model performance. On the chest X-ray diagnosis task, using the proposed approach, the model performance is improved by 4.38%, compared to the state-of-the-art method. These results show that, by leveraging human attention, model performance can be improved as it is able to find critical features for making accurate decisions. This research not only sheds light on the effective role of human attention in challenging classification tasks, but also paves the way for future studies on integrating human perception knowledge into computer vision models.

My contributions. I led this project by introducing an innovative research concept that integrates human attention into a computer vision model. My leadership extended to lead the data collection, programming the model framework, and conducting a comprehensive analysis of the results derived from the implemented model. Additionally, I took charge of authoring the manuscript.

3 Human Gaze-based Attention in Classification

3.1 Introduction

Human attention (HA), which can be represented by human gaze, offers insights into our actions and choices [156]. Similarly, many computer vision systems utilize human gaze data to identify key objects for task completion [31, 34, 70]. A common practice in visually representing human attention in these systems is to apply a Gaussian blur to focal points, creating a feature map, often referred to as a *saliency* map [157], [35] (refer to Figure 3.1). In parallel, post-hoc attention of a network, or model explanation, attempts to pinpoint critical areas that influence neural network decisions [90, 88, 85, 158, 89, 159]. Saliency maps are a common tool for visualizing both human and model attention, facilitating the examination of their similarities and differences. Previous studies have indicated that humans and models often focus on different areas when executing the same task [96, 97]. However, it remains uncertain whether a feature identified by human attention is more effective in task resolution. Our research aims to bridge this gap, hypothesizing that (1) human attention zeroes in on crucial features for task completion, such as fine-grained classification, and (2) incorporating human attention can enhance model performance in these tasks.

To test the first hypothesis, we represent human attention through saliency maps and compare the focus areas of human and model attention (model explanation), demonstrating that human attention often highlights more discriminative regions for classification. To validate the second hypothesis, we introduce two methods that leverage key features identified by human gaze: Gaze Augmentation Training for refining classifiers and a Knowledge Fusion Network for integrating human attention insights into models.

This study presents the contributions as follows:

- We have gathered human gaze data for the CUB-200-2011 (CUB) dataset and enriched it with human attention insights. This enhanced dataset is named CUB-GHA (Gaze-based Human Attention). We further establish the effectiveness of human gaze data in identifying discriminative features within this unique dataset.
- We introduce two novel approaches to integrate human attention into classification tasks: Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN). These methods are designed to leverage human attention knowledge for enhanced classification accuracy.
- Our methods are thoroughly evaluated on two challenging datasets: the CUB-GHA dataset for fine-grained bird species classification, and the CXR-Eye dataset

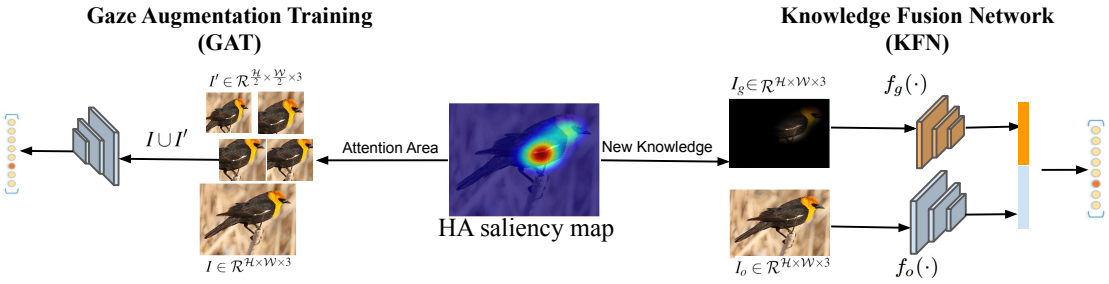


Figure 3.1: Methodology overview highlights two primary processes. Firstly, the HA saliency map is employed to pinpoint areas of focus, which are then incorporated to improve the training dataset in the Gaze Augmentation Training (Left). Secondly, this HA saliency map serves as an additional information channel, which is integrated with the existing image data in the Knowledge Fusion Network (Right).

which includes chest radiography images and radiologist gaze data. Our research demonstrates that incorporating human attention into classification models can significantly improve their performance, setting new benchmarks in various classification tasks.

3.2 CUB-GHA Dataset

This section outlines the methodology for gathering gaze data and subsequently examines the impact of machine explanations and human focus on the detailed classification model. We utilize the CUB-200-2011 dataset [160], containing 11,788 images across 200 bird species, to acquire gaze data. This dataset includes diverse annotations such as image attributes, body part locations, and bird descriptions. Our process results in a modified version with enhanced human gaze data, referred to as CUB-GHA. The choice of the CUB dataset is driven by two reasons: First, the subtle distinctions between similar bird species are primarily in their localized and compositional features, which human gaze can accurately capture. For example, differentiating between species based on minor variations like throat color is a more specific task compared to contrasting broadly dissimilar animals like a bear and a horse (illustrated in Figure 3.2). Second, the CUB dataset's extensive use in various computer vision applications, including detailed classification [78, 79, 161], zero-shot learning [162, 163, 164, 84], and explainable AI [165, 93, 166], makes CUB-GHA a potentially valuable asset for investigating the influence of human attention in these areas.

3.2.1 Gaze Data Collection

As shown in [70], when viewing two closely related classes, humans tend to focus on features that distinguish one class from the other. In our research, we implement an image comparison game, similar to [70], where participants are motivated to concentrate

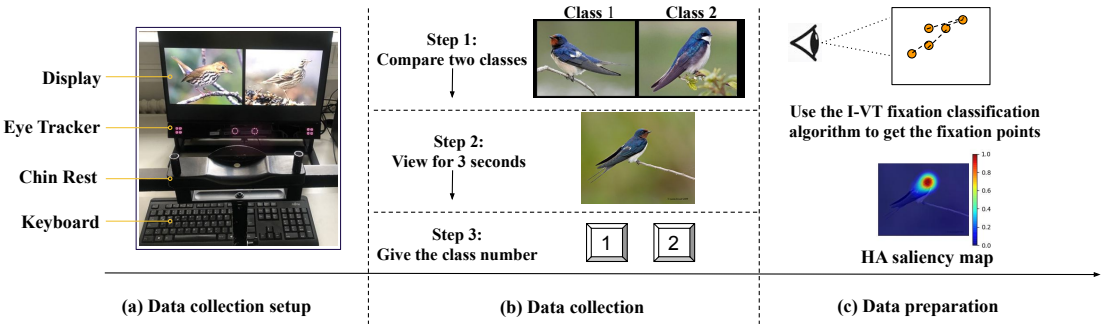


Figure 3.2: **(a)** Eye Tracker Configuration: A Tobii Spectrum eye-tracker is utilized, capable of recording gaze patterns at a swift 1200 Hz frequency. **(b)** Data Collection: The first step provides a diagrammatic representation of the task where participants view images of two distinct species. In the second step, an image of one randomly chosen species is displayed for gaze tracking. To make the process engaging for participants, they are asked to identify the species in the third step. **(c)** Data Preparation: Gaussian-based techniques are utilized to visually depict human attention through saliency maps.

on these distinguishing features while comparing two similar images from distinct categories. The task is intentionally made difficult by selecting two very similar classes for each comparison pair, aiming to yield more powerful insights.

Figure 3.2 provides an overview of our data collection process. Part (a) of the figure illustrates the experimental setup, which includes an image of the eye-tracking device (Tobii Spectrum Eye Tracker) operating at 1200 Hz, with the chin rest and the display screen. The display has a resolution of 1920 $\times$ 1080. The chin rest plays a crucial role in ensuring accurate tracking of eye movements. Each image displayed during the experiment is resized to fit the screen and is centrally positioned. The typical distance maintained between the participant's nose and the screen is around $60cm$. The comparison task, divided into three steps, is depicted in part (b) in Figure 3.2. In step 1, two images are simultaneously presented to the participants, each representing a different bird class from the CUB dataset; for example, images might be of the Barn Swallow and Tree Swallow. These pairs are carefully chosen from the same sub-classes, and their visual similarity is manually verified by different individuals to ensure the comparison task is not overly simple. Participants are given the flexibility to view these images for an unrestricted duration. In step 2 of the experiment, when participants indicate readiness to proceed with the classification task, they are shown an image from one of the two bird classes featured in the CUB dataset. The task for the participant is to identify the class to which the displayed image belongs. To focus the participant's attention directly on the classification task and minimize unrelated exploratory gaze behavior, each image is presented for only three seconds.

Each data collection session consists of presenting one image from each of the 200 classes in the dataset, resulting in a total of 200 images reviewed per session. To en-

sure a comprehensive evaluation, every image in the CUB dataset is examined by five different participants. The study involves 25 subjects, comprising 19 males and 6 females, with an average age of 27.64 years (standard deviation of 4.15 years). Although not all participants engage in an equal number of sessions or view the same instances, the experiment is structured to ensure that each participant is exposed to all classes within each session they attend. Notably, all the participants are novices in the domain, possessing no specialized knowledge about bird species.

3.2.2 Human Attention Saliency Map Generation

The collected gaze data undergoes preprocessing to pinpoint fixation locations, achieved through the use of the Velocity-Threshold Identification (I-VT) algorithm [167]. The fixations identified in the dataset provide not only coordinate information but also the duration of each fixation. Utilizing these details, we create saliency maps that represent human gaze, as depicted in Figure 3.2 (c).

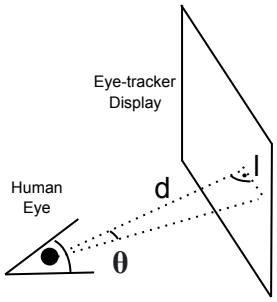


Figure 3.3: Illustration of a person viewing an image on an eye-tracking monitor.

Concretely, Figure 3.3 depicts a person observing an image through an eye-tracker display. As described in the paper, each fixation point is converted into a Gaussian distribution $N(\mu, , \sigma^2)$ on the HA saliency map, with σ being 75 pixels, corresponding to the display's resolution. We determine the standard deviation σ using the following method. In our experimental setup, the observer's eye is 60 cm away from the eye-tracker display, and the visual angle θ is fixed at 2°, in line with [168]. Consequently, $l = \tan 2° \cdot d = 21\ mm$. Given the display's dimensions (530 mm in width) and its resolution (1920 pixels), we deduce that an extent of 21 mm on the display translates to roughly 75 pixels. This value (75 pixels) is adopted as the standard deviation, with the image being adjusted to the display's resolution (1920 × 1080). Post-processing, the saliency map is resized back to its original dimensions. The duration of each fixation is then used to weight its corresponding Gaussian distribution. The generated saliency map is depicted as a grayscale image.

3.3 Comparison between Human and Post-hoc Model Attention

In this section, we test the theory that *HA identifies key areas for detailed classification.* Post-hoc model attention can be extracted with model explanation (ME) techniques, as discussed in Section 2.1. Using the same image and task, both HA and ME pinpoint essential regions for decision-making in humans and models respectively. Therefore, we compare HA against four MEs from a trained classifier (vanilla ResNet-50 [169]) which achieves a classification accuracy of 85.58% on CUB. This comparison confirms that HA successfully highlights characteristics that more effectively distinguish one bird species from another. The four ME used are Class Activations Maps (CAM) [88], Gradient-based CAM (Grad-CAM) [85], InputXGradient (IxG) [89], and IntegratedGradients (IG) [158].

To quantitatively assess HA and ME, we employ the keep and retrain (KAR) method (proposed in [159]) to determine whether the significant areas identified by HA and ME contribute to the model's decision-making process. The specific method is outlined as follows: we start with an input image I that exists within the space $\mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 3}$, alongside an importance estimation map A, which exists in the space $\mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$. Here, H and W denote the height and width of the input image, respectively. The map A might be either the HA or ME saliency map. A mask M is created in the space $\mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$ to select specific pixels from I. Initially, we rearrange A into A^R in descending order based on the attention values. Following this, we convert A into a binary format by designating the top p percent of pixels in A^R as one, while the rest are set to zero:

$$M(x,y) = \begin{cases} 1.0, & \text{if } (x,y) \in P \\ 0.0, & \text{otherwise} \end{cases},$$

where P are the indices of top ranked p percent pixels. The mask M is applied to the image I in both training and testing phases, resulting in a modified image $I' = M \odot I$. This process ensures that only the top p percent of key features are visible to the network. Then, a new model is trained with this modified dataset, and its test accuracy is evaluated. The objective is to determine the significance of the features identified by A (a model explanation or human attention saliency map) in the classification process. An effective estimation by A would represent vital features using a minimal pixel count. Hence, achieving higher accuracy with fewer pixels implies greater importance of these features. The underlying reasoning is that highly important pixels should contain class-specific details; thus, adding more pixels of lesser importance will not significantly enhance model performance. When a saliency map effectively identifies informative features as crucial for classification, there will be a swift increase in accuracy at the onset of pixel insertion. In other words, a greater Area Under the Curve (AUC) reflects a more accurate estimation of feature significance. The new dataset is created with an insertion percentage p = [5,10,15,20,15,30,50,70,90].

Figure 3.4 (top) displays both the KAR curves and the AUC scores for each method, while the bottom part of the figure presents qualitative saliency maps for HA and the four MEs for a specific image. It is observed that HA and MEs highlight different areas

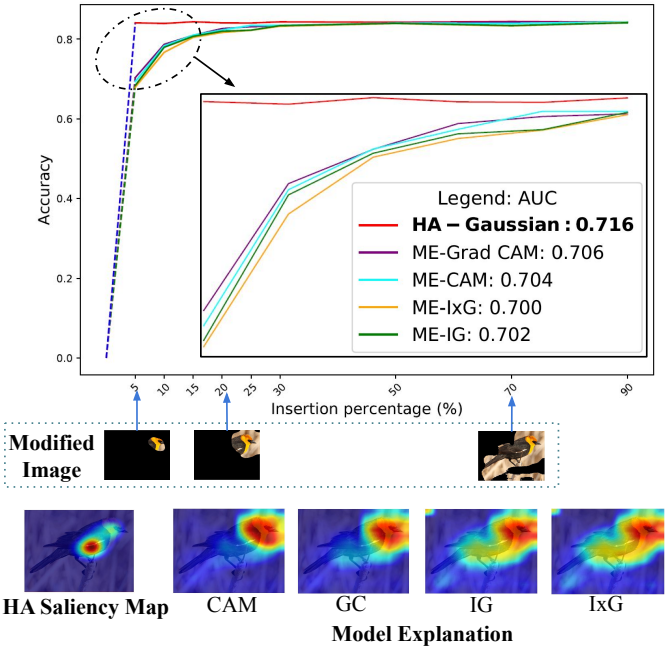


Figure 3.4: Comparison between HA and ME in identifying distinct features. **Top:** It shows the test accuracy on altered datasets utilizing various saliency maps. The horizontal axis represents the percentage of insertion, while the vertical axis indicates the accuracy on the test set. The Area Under the Curve (AUC) for each line is detailed in the enlarged image. **Middle:** This part presents images modified with Grad-CAM as a representative example. **Bottom:** This section visually represents HA and the four different MEs.

of the image: humans prioritize the white feathers on the black wing, whereas the model deems the yellow head as the most crucial feature (refer to the original image in Figure 3.1). HA is more effective in identifying relevant and significant features for the fine-grained classification model than MEs. For instance, HA achieves an AUC score of 0.716, surpassing Grad-CAM (0.706) and IG (0.702). When only 5% of the important pixels are revealed, the model using HA-modified images attains an 81% accuracy, significantly higher than the approximately 70% accuracy achieved by the model using ME-modified images.

Furthermore, we perform a quantitative analysis to compare the similarities between HA and MEs, utilizing a range of metrics in Table 3.1. These include Kullback-Leibler divergence (KL-D), correlation coefficient (CC), and similarity (SIM) - commonly employed in image comparison studies [170]; rank-correlation (Rank-Co) as proposed in [96]; the shuffled AUC metric (sAUC) for evaluating individual pixels in saliency maps; and information gain (IG), which assesses performance against a standard [170, 61]. The comparison shows that CAM and Grad-CAM are similar, with Grad-CAM scoring 0.565 on CC and 1.242 on KL-D, compared to CAM's 0.563 and 1.248. Moreover, IG and IxG show comparable results, with IG scoring 0.699 versus IxG's 0.694 on CC, and 1.318

versus 1.310 on KL-D. These findings are supported by qualitative data as well. Across these metrics, Grad-CAM consistently appears most akin to HA, leading in all six metrics. This aligns with KAR findings that place Grad-CAM at the forefront regarding performance among all MEs.

	KL-D ↓	CC ↑	SIM ↑	Rank-Co ↑	sAUC ↑	IG ↑
CAM	1.248	0.563	0.399	**0.761**	0.460	0.938
Grad-CAM	**1.242**	**0.565**	**0.415**	**0.761**	**0.508**	**1.376**
IG	1.318	0.546	0.361	0.699	0.436	0.921
IxG	1.310	0.543	0.375	0.694	0.461	1.001

Table 3.1: Similarity comparison between MEs and HA saliency map. (↓: the lower the better; ↑: the higher the better.)

3.4 Human Attention Integration Strategy

This section outlines our method of integrating gaze data to enhance classification results. This is achieved either by enriching the training data with gaze (GAT) or by utilizing it as an additional source of information (KFN). A depiction of this framework is presented in Figure 3.1.

3.4.1 Gaze Augmentation Training

Driven by the belief that our model should focus on key areas of an image, as indicated by human attention (HA), we improved the model's response to these areas by incorporating them into our training, as shown in Figure 3.1 (left). To create k augmented images from the original image $I \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 3}$ (with $\mathcal{H}$ and $\mathcal{W}$ denoting the image's width and height), we use a sliding window technique to identify regions of human attention. This involves a window of size (w, h) moving across the HA map $A \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$ from the top-left to the bottom-right corner, advancing by stride size s in both directions. The areas under the window are ranked based on average pixel values, and the top k areas are selected to create cropped images. These images are then resized to half the width and height of I, resulting in $I' \in \mathcal{R}^{\frac{\mathcal{H}}{2}\times\frac{\mathcal{W}}{2}\times 3}$, following the approach in [77, 78, 71] where the focus areas are reduced in size. I' retains the same label y as I. To capture diverse regions, we vary the window sizes and apply non-maximum suppression. The training set is thus expanded to include both I and I'. The model is trained on this augmented dataset using cross-entropy loss. Note, however, that the Gaze Attention Transformer (GAT) only requires human gaze data during training, as it processes only original images during testing.

3.4.2 Knowledge Fusion Network

Illustrated in Figure 3.1 (right), our KFN model consists of a dual-branch network combining the insights from HA and the native image characteristics. The first branch, designated as the image knowledge branch, processes the original images $I_o \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 3}$, with $\mathcal{H}$ and $\mathcal{W}$ denoting the image's height and width respectively. A CNN backbone $f_o(\cdot)$ is employed to derive the image feature $f_o(I_o) \in \mathcal{R}^{D_o}$ from I_o, where D_o indicates the feature channel's dimension. The second branch, termed the HA knowledge branch, integrates the gaze attributes of the image. Here, the HA is combined with the original image, denoted by $I_g = I_o \odot A$, where $A \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$ represents the HA saliency map. This process assigns weights to image pixels based on HA saliency maps, highlighting areas of human attention more prominently. I_g encodes crucial visual elements for classification purposes. An additional CNN backbone $f_g(\cdot)$ extracts the gaze feature as $f_g(I_g) \in \mathcal{R}^{D_g}$. Then, these gaze and original image features are merged to create the combined feature $f(I_o, I_g) \in \mathcal{R}^{(D_o+D_g)}$. This fusion of HA in a multiclass classification context explores HA's capacity to enhance image classifier efficacy. The training of the network employs the cross-entropy loss.

3.5 Experiment

This part begins with an overview of the datasets and the specifics of the implementation. Following that, the outcomes of our GAT and KFN are presented. Our methods are tested for their broad applicability on two datasets, namely CUB-GHA and CXR-Eye, which stands for Eye Gaze Data for Chest X-rays [2].

3.5.1 Implementation details

The CUB-GHA dataset contains a total of 11788 images, split into 5994 for training and 5794 for validation, as detailed in [160]. In every image, eye gaze data from 5 individuals is featured. The CXR-Eye dataset, on the other hand, consists of 1083 chest X-ray images, each accompanied by gaze data from a radiologist during standard radiology reviews [2]. This dataset's primary objective is to determine, based on the chest X-ray image, whether the individual has pneumonia, congestive heart failure (CHF), or is in normal health. Additionally, the eye gaze information is presented in a saliency map format, and each image is tagged with one of three possible labels. The choice of this dataset is driven by its uniqueness as a human gaze dataset in the medical field. We assert that integrating human attention in critical applications, e.g., computer-aided diagnosis can enhance user acceptance and trust.

In our study conducted on the CUB dataset, we first adjust the dimensions of the input images to 448×448 (by cropping after resizing the smaller edge to 448) and apply a random horizontal flip during training. The SGD optimizer is utilized [171], starting with a learning rate of 0.001. For the CXR dataset, we resize the input images to 224×224 and also incorporate a random horizontal flip in the training phase. Here, the Adam optimizer is employed [172], with an initial learning rate set at 0.0005. The

	Small	Medium	Large
CUB-GHA	(123,134) (134,123) (123,123) (134,134)	(174,190) (190,174) (174,174) (190,190)	(246,264) (269,246)
CXR-Eye	(87,95) (95,87) (95,95) (87,87)	(123,135) (135,123) (123,123) (135,135)	(180,190) (190,180)

Table 3.2: Sliding window size used in GAT.

CXR-Eye dataset, being comparatively smaller, is subjected to 5-fold cross-validation, and the mean accuracy across the five validation sets is presented as the ultimate score. All tests are conducted over 100 epochs, using a single NVIDIA GeForce RTX 3090. The learning rate is reduced by a factor of 0.1 every 50 epochs.

In the experiments, we use ResNet-50 [169] and EffiecientNet-b5 [173] pretrained on ImageNet as backbones on CUB and CXR, respectively. In the GAT approach, the original image is segmented using three different groups of window dimensions: large, medium, and small. Within each group, a sliding window technique generates k augmented images per training set image. Specifically, for large-scale windows, k is set to 2 for large, 3 for medium, and 4 for small scale, resulting in a total of 9 augmented images. Furthermore, when integrating GAT with KFN, the GAT-trained classifier serves as the backbone of the KFN, which is then fine-tuned over only 20 epochs.

3.5.2 Evaluation on CUB-GHA

Window sizes in GAT. Table 3.2 lists the dimensions of concrete sliding windows (w, h) applied in the GAT experiments for each dataset. The sliding window sizes for the CUB-GHA dataset are derived from the average dimensions of bird bounding boxes when images are resized to 448×448: a width of 246 and a height of 269. These dimensions are used for the large scale windows. The medium scale windows are determined by multiplying these dimensions by $\frac{\sqrt{2}}{2}$, resulting in window sizes of 174 by 190, which halves the area of the bounding box. The small scale uses a factor of 0.5. For the CXR-Eye dataset, factors of 0.8 and 0.85 are applied to the resized image size of 224×224 to obtain large window sizes of 180 and 190. The medium window sizes employ factors of 0.55 and 0.6, while the small window sizes are scaled from the medium sizes using a factor of $\frac{\sqrt{2}}{2}$. The purpose behind varying sliding window sizes is to capture different discriminative regions for classification. An intersection over union (IoU) threshold of 0.25 is set in non-maximum suppression to ensure diversity in the cropped areas. Table 3.3 presents an ablation study on the impact of different numbers of cropped areas (k) in augmentation training, as shown in two datasets. The notation (2,2,2) implies that two cropped areas from each window scale contribute to the augmentation training set. The chosen configuration of (2,3,4) demonstrates relatively superior results on both datasets, as depicted in Figure 3.5, which illustrates the augmentation images using this setting across the three window scales.

Ablation studies. In our study, we assess the impact of GAT and KFN on classification by conducting an ablation study using the CUB dataset. A ResNet-50 model trained

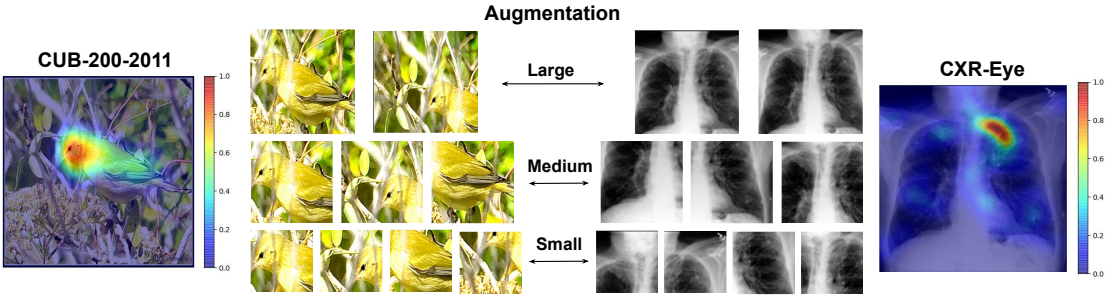


Figure 3.5: Illustration of cropped images used in the GAT. **Left and Right:** Saliency maps from HA applied for augmentation on CUB-GHA and CXR-Eye. **Middle:** Images cropped at three different scales (large, medium, and small).

(L,M,S)	CUB (%)	CXR (%)
(2,2,2)	87.50	71.03
(2,3,2)	88.06	71.58
(2,3,3)	88.00	71.86
(2,3,4)	88.00	72.21

Table 3.3: Accuracy (%) of applying various window size configurations on CUB-GHA and CXR-Eye. The left side displays the count of windows utilized in large, medium, and small sizes.

with cross-entropy loss serves as the baseline, while its variations include incorporating GAT and KFN modules. The data in Table 3.4 reveals significant enhancements in classification precision due to both GAT and KFN. Specifically, GAT, when combined with Human Attention (HA), boosts the base model's accuracy from 85.58% to 88%, highlighting the importance of human gaze in identifying key features for classification. Similarly, employing HA with KFN increases accuracy from 85.58% to 86.99%, confirming KFN's effective integration of human attention knowledge. To further validate the effectiveness of, we replace it with model-generated saliency maps using Grad-CAM [85] and IG [158] in both GAT and KFN. Here, HA outperforms these alternatives, with KFN (HA) reaching an accuracy of 86.99%, compared to 85.66% with KFN (IG), underscoring the unique insights offered by human gaze that the model alone cannot replicate. Additionally, combining GAT and KFN yields better results than using either independently.

Comparison with state-of-the-art. In our study, we compare our proposed modules with other SOTA baselines. To ensure a fair comparison, we use the ResNet-50 as the baseline and set the input resolution at 448×448 for all methods. First, we evaluate our GAT against other data augmentation techniques such as MixUp [174], CutMix [175], and SnapMix [176], as shown in Table 3.5 (top). Unlike these methods, our GAT

Method		Acc.
ResNet-50 [169]		85.58
GAT	Grad-CAM [85]	87.68
	IG [158]	87.73
	HA	88.00
KFN	Grad-CAM [85]	85.04
	IG [158]	85.66
	HA	86.99
GAT+KFN	HA	**88.66**

Table 3.4: Ablations study of GAT and KFN on CUB. "Acc." denotes the accuracy in %.

does not create synthetic images. MixUp linearly blends two images and their labels, while the others swap sections of one image with parts from another. Conversely, GAT merely enlarges the dataset using cropped images, thus adding minimal computational load to the training of the classifier. Among these approaches, training a ResNet-50 with GAT surpasses other advanced augmentation methods, reaching an accuracy of 88%. Furthermore, this enhanced training base can be seamlessly integrated with other frameworks for improved results; for example, when we combined it with our KFN, it yielded superior outcomes.

Method	Acc.
MixUp [174]	86.23
CutMix [175]	86.15
SnapMix [176]	87.75
Ours (GAT)	**88.00**
OSME+MAMC[72]	86.30
TASN [79]	87.90
API [80]	87.70
ACNet [81]	88.10
Ours (KFN+GAT)	**88.66**

Table 3.5: Comparison with the state-of-the-art methods on CUB. **Top:** Comparison of GAT with data augmentation methods. **Bottom:** Comparison of GAT+KFN with attention-based models.

In Table 3.5 (bottom), our KFN+GAT network is compared with attention-based techniques on CUB. Selected for their efficiency and significance in attention modules, these methods include OSME+MAMC [72], TASN [79], API [80], and ACNet [81]. These approaches use attention modules to extract key features from the network's intermediate results, whereas our method incorporates HA directly. For example, [72, 81] employ multiple layers on top of the residual block output for region feature extraction; API

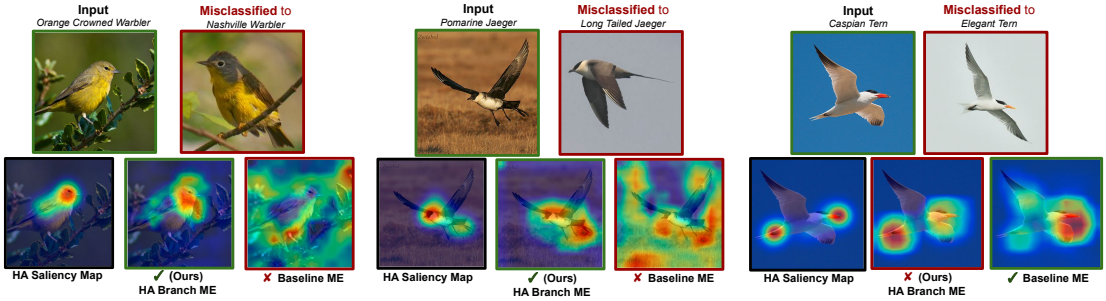


Figure 3.6: Illustration of model explanations using HA. Two improved examples and one failure example of our model are shown. For each of these cases, we present the input alongside the misclassified categories: HA saliency map, the explanation of our model, and the explanation of the baseline model.

[80] emulates human comparison processes, similar to our data collection approach, for learning distinct feature representations. Surpassing all SOTA models, our full network achieves an 88.66% in accuracy, higher than the attention networks API (87.70%) and ACNet (88.10%). This underscores the efficacy of our KFN and GAT, demonstrating that incorporating human gaze can enhance model performance in classification tasks.

Method	S3N [177]	S3N + GAT (Ours)	CrossX [178]	CrossX + GAT (Ours)	MMAL [74]	MMAL + GAT (Ours)
Accuracy	87.95%	88.91%	87.70%	88.51%	89.25%	**89.53**%

Table 3.6: Combining our GAT model with the state-of-the-art methods on CUB.

Qualitative results. Figure 3.6 illustrates two instances where our model shows improvement compared to a vanilla ResNet-50, and one instance where it incorrectly classifies. In the first example, the baseline model mistakenly identifies an Orange Crowned Warbler as a Nashville Warbler due to the focus on the yellow belly, a common feature. However, our model distinguishes the species by the throat color: The Orange Crowned Warbler has a purely yellow throat, in contrast to the Nashville Warbler's throat, which blends gray and yellow. The second case highlights the tail as the key feature. Here, our model accurately identifies the tail, unlike the baseline model, which confuses the background for the tail. Additionally, our model's explanation is more concise and reflects human saliency maps more closely. The third case demonstrates a limitation of our model, where it mistakenly classifies a Caspian Tern as an Elegant Tern, focusing on the feet rather than the beak. This error arises despite our model's alignment with the human focus, as it overemphasizes the foot color, a crucial feature in differentiating a Caspian Tern from either a Common or an Arctic Tern.

Figure 3.7: Illustration of the influence of using HA in model explanation. **Left to Right:** the original Chest X-ray image; HA saliency map; Model explanation of the Image Branch (w/o HA knowledge) and Model explanation of the HA Branch.

3.5.3 Evaluation on CXR-Eye

Comparison with state-of-the-art. In the paper on CXR-Eye [2], the authors employ the Efficient-b5 [173] for classification purposes. They utilize randomly generated splits for training, validation, and testing datasets. To ensure a fair evaluation, we re-evaluate their model using the same 5-fold cross-validation approach and average the accuracies from five validations to determine the evaluation outcome. This baseline approach yields a 70.97% accuracy rate. Incorporating our GAT leads to a slight increase in performance, achieving 71.86%, while the addition of the KFN further enhances accuracy by 3.45%, reaching 74.42%. Combining both GAT and KFN in the full model results in a notable improvement, with an accuracy of 75.35%, surpassing the vanilla Efficient-b5 by 4.38%. Among the enhancements, KFN had a more significant impact on the CUB model than GAT. The potential reason for this difference is how the gaze data is collected.

In the CXR-Eye, the eye movement data of the radiologist is gathered during their analysis process. Observing Figure 3.7 (second column), it is obvious that the radiologist's gaze is dispersed across numerous points (indicated by the light blue region). While these points might be crucial for diagnostic purposes, GAT specifically identifies the zones where the radiologist's attention is more prolonged. Moreover, KFN enhances overall effectiveness by incorporating information from all these relevant areas, thus resulting in a more substantial improvement.

Qualitative results. To assess the impact of HA integration in the network, we examine the model explanations (using Grad-CAM [85]) for each component of KFN, with qualitative outcomes presented in Figure 3.7. The HA branch appears to align closely with human focus, unlike the image branch, which targets different areas. For instance, in the upper example, the human gaze focuses on the left, reflected by the HA branch, while the image branch favors the right side. The image branch in the second example fixates on an incorrect region, but the HA branch adjusts the focus correctly. Thus,

KFN outperforms models relying solely on images. More importantly, the incorporation of gaze data enhances model trustworthiness and acceptance, especially in fields like medical diagnostics, by ensuring model decisions are consistent with human gaze behaviors.

3.6 Conclusion

In this study, we explored the role of human gaze in classification tasks using the CUB and CXR datasets. We introduced a new gaze dataset, CUB-GHA, to demonstrate that human attention is directed toward key areas in detailed classification tasks. Our research proposed the Gaze Augmentation Training and Knowledge Fusion Network, incorporating human gaze insights into the network, significantly enhancing classification accuracy on both datasets. This finding supported the hypothesis that human attention can guide models in identifying unique features for different classification tasks.

Additionally, this research provided a valuable resource for the community: the CUB-GHA dataset enriched with human gaze data. This dataset complements existing datasets with detailed annotations (like textual explanations, attributes, and bounding boxes) and can be used in various applications where human-AI interaction involves the human gaze.

Part II

Predicting Human Intention

The previous part of the dissertation delves into the incorporation of human attention knowledge into models to enhance their performance, underscoring the efficacy of AI model design informed by human factors. This part considers another human factor, namely human intention, and explores the potential of AI in assisting humans, specifically through predicting human intentions.

This part is adapted from two works that appeared in ITSC 2020 [179] and PACMHCI 2022 [45], respectively:

- Rong, Y., Akata, Z., & Kasneci, E. (2020)

  Driver intention anticipation based on in-cabin and driving scene monitoring

  In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*

- Rong, Y., Kassautzki, N.-R., Fuhl, W., & Kasneci, E. (2022)

  Where and what: Driver attention-based object detection

  In *Proceedings of the ACM on Human-Computer Interaction (PACMHCI)*

In the following section, I will summarize the motivation, principal methodology, main findings, and my contributions to each of the two papers.

Driver intention anticipation based on in-cabin and driving scene monitoring

Motivation. As discussed in Section 1.2.2, more efficient human-AI collaboration can be fostered if AI models can foresee human intentions. A very good application use case for this is in the area of high-level autonomous driving, where there is a growing need for an AI model to serve as a co-pilot [180]. To address these important practical use-cases, this part employs driver intention prediction as the case study.

Driver intention can be inferred from their actions and the surrounding traffic information recorded in monitoring videos. Section 4.2 is motivated to predict human maneuver intention based on videos of in-cabin and driving scene monitoring, while prior works mainly focus on using driver monitoring videos. This work aims at improving driver intention prediction using dynamic traffic information. For this task, the dataset Brain4cars [3] is used, which contains in-cabin and outside videos.

Principal methodology. To effectively predict driver intentions based on videos, Section 4.2 proposes a novel method to use outside videos to inform driver intention prediction, because traffic scenes can offer insights into the motion of the car. More specifically, it proposes a Convolutional-LSTM (ConvLSTM)-based auto-encoder to extract motion features from the out-cabin videos. Another branch is trained to extract driver behavior features from videos. Then, a classifier is trained on features from both in- and outside of the cabin jointly for maneuver intention anticipation.

Main findings. Evaluation results (Section 4.2) show that the inside and outside monitoring videos have complementary information. Therefore, the proposed framework is able to achieve more precise prediction than using any of the video sources only. More specifically, using inside video only achieves an accuracy of 77.4% and F_1-score of 75.5%, and using outside video only reaches 60.9% and 66.4% in accuracy and F_1-score, respectively. The proposed method achieves state-of-the-art performance with an accuracy of 84.0% and F_1-score of 84.3%. These findings demonstrate the effectiveness of the proposed algorithms in predicting driver intention: the proposed method leverages motion features derived from the optical flow, and human action feature embeddings extracted by a 3D CNN.

My contributions. This paper was a project under my leadership. Specifically, in the ITSC 2020 paper [179], I identified the prevailing research gap and introduced the use of ConvLSTM architecture for extracting features from external video data. This effort involved developing the code and conducting analyses of the experimental results.

Where and what: Driver attention-based object detection

Motivation. This work is beyond driver actions in Section 4.2. In Section 4.3, this work detects the objects of the driver's intention, i.e., the objects that the driver focuses on and plans to interact with. Human drivers utilize their gaze-based attention to identify crucial objects and to make further reactions. The analysis of gaze data has become increasingly relevant in enhancing autonomous driving systems. Prior studies have mainly focused on determining "where" drivers look, neglecting "what" specific objects capture their attention. Conversely, while standard object detectors can identify all objects in a driving scene, they do not provide insights into the driver's intentions. This work is motivated to bridge this research gap by the precise detection of critical objects within the driver's focus.

Principal methodology. To predict objects of the driver's intentions, Section 4.3 combines a gaze prediction module with an object detection module, allowing for the precise detection of critical objects within the driver's focus. Concretely, the gaze prediction module estimates where the driver is looking by predicting saliency maps, while the object detection module identifies all objects within the traffic. Attention-based objects are then recognized and presented to users, guided by the combination of the saliency map and the objects detected. Another novelty is that the predicted saliency map is grid-based rather than pixel-based. In this way, predicted saliency maps can be used to precisely detect the objects, without adding too much computational burden to the whole framework.

Main findings. Section 4.3 presents the evaluation results of the proposed driver-attention-based object detection framework. The results address the advantage of the proposed attention prediction module compared to other driver attention prediction

methods in object detection precision and computation costs. For instance, the SOTA model achieves 0.86 in AUC and 73.8% in F_1-score, but the model needs 47.2M parameters and 92.30 GFLOPs to compute the attention saliency map. However, our method achieves competitive results in object detection with 0.85 in AUC and 72.8% in F_1-score with much less computational resources, i.e., 7.5M parameters in the network and 17.0 GFLOPs. Similar to the first work, these findings demonstrate the effectiveness of the proposed algorithms. It introduces an effective attention module to predict the objects within human attention. These methods can be further adapted in a wide range of applications that utilize video input.

My contributions. In this PACMHCI paper [45], I originated the research idea and developed the methodology to address the problem. I personally coded the framework, led the collection and analysis of the main results, and was responsible for writing both manuscripts.

4 Driver Intention Prediction

4.1 Introduction

On the halfway to autonomous driving vehicles, it is therefore necessary to provide already existing Advanced Driver Assistance Systems (ADAS) the functionality for collaboration with the human driver in the most efficient way, for example, to alert the driver in case of a dangerous maneuver. To achieve such an effective collaboration, many researchers focused on detecting the maneuver intention of the driver before execution.

Recently, multiple research has concentrated on identifying a driver's intent to maneuver prior to their execution. Datasets such as Brain4cars [101] and the Honda Research Institute Driving Dataset (HDD) [102] have been specifically developed to study and understand driving behaviors. The HDD system, as detailed in [102], employs high-resolution video cameras, together with GPS, LiDAR sensor signals, and vehicle CAN-Bus data to capture detailed traffic scenes. Brain4cars [101] collects videos both from within and outside a vehicle, incorporating GPS and vehicle dynamics information with the video data. These videos offer insights into various patterns of driver maneuvers and overall road traffic behavior. The significant amount of information conveyed through these frames has been extensively explored in research, particularly in the context of predicting driver intentions based on video analysis of drivers glancing towards side mirrors. Studies utilizing the Brain4cars dataset, including [101, 3, 103, 104, 105], have successfully demonstrated maneuver prediction.

Much of the prior research in predicting driver maneuvers is mainly based on video data obtained from monitoring the driver. Studies have consistently demonstrated that observing a driver's behavior, particularly their eye movement, is not just useful for recognizing activities [106, 107], but also plays an important role in ensuring safe transition behaviors in partially automated vehicles [65]. In addition, video clips capturing the driver are employed to derive various features, such as the positioning of the head [101, 3, 103, 105]. However, these studies typically involve a manual process of converting traffic information into a four-element vector. This vector includes two Boolean indicators to show the presence of lanes on either side of the car, a Boolean value indicating the proximity of an intersection or turn within 15 meters, and a final element representing the vehicle's current speed. Consequently, the external information captured in videos is not further analyzed.

The outside video from the road scene ought to be highly revealing, offering insights that the internal video fails to deliver. Therefore, Section 4.2 focuses on (1) efficiently garnering vehicular movement data from traffic videos, thereby enhancing outcomes that were previously reliant on a singular video stream; (2) introducing a comprehensive approach that forgoes the need for manual coding data; and (3) maintaining a minimal-

ist model design (fewer parameters) to ensure its suitability for mobile platforms with limited resources.

Beyond intention predictions based on human actions, objects of the driver's attention can also indicate the driver's intentions, as demonstrated in Figure 1.4. To detect these objects, effective models should be able to perform "gaze-object mapping," which involves two tasks: predicting where the driver is looking and associating that gaze with specific objects. Predicting a driver's gaze is feasible, especially when an eye tracker is not available, or the vehicle operates at a higher autonomy level without a human driver. For example, research by Pomarjanschi et al. [181] demonstrates that highlighting critical objects like pedestrians on a head-up display can decrease collision incidents. In this scenario, a model that identifies these important objects acts as a "co-driver," providing alerts to assist the actual driver. For completely self-driving vehicles, it is crucial to quickly and accurately identify relevant objects for decision-making and explanation purposes [182]. There is increasing interest in research on predicting where human drivers focus their gaze [8, 7, 121]. These studies generate pixel-level attention maps, but they don't convey the semantic significance of the observed attention; that is, the models indicate *where* drivers look, but not *what* objects are in those areas.

To address the existing gap in research between predicting driver gaze and detecting semantic objects in autonomous driving, this work in Section 4.3 introduces a dual-focused approach: (1) predicting both the location and the nature of objects drivers observe, and (2) developing a model that is computationally efficient due to the limited resources available in self-driving vehicles. We have developed an innovative framework centered on efficient, attention-based object detection aligned with human driver gaze. This framework not only generates attention saliency maps at the pixel level but also identifies objects within these attention zones, as depicted in Figure 4.6. Initially, a feature encoder processes the input image's information. Subsequently, these features are utilized for simultaneous gaze prediction and object detection. Our primary objective is to accurately identify high-level (object) information rather than achieving pixel-level precision in saliency map predictions. Therefore, we employ a grid-based approach to identify salient areas, reducing computational demands while ensuring robust performance in crucial object detection tasks.

In the following sections, details of each proposed method will be introduced.

4.2 Driver Intention Prediction based on Videos

As highlighted in the last section, various works have focused on predicting driver maneuvers [101, 3, 103, 104, 105]. Yet, none have successfully predicted driver intentions by integrating data from both inside and outside vehicle video feeds, given the intricate nature of road traffic, which complicates the creation of explicit features. Consequently, several research efforts, including [101, 3, 103, 105], have resorted to using manually encoded feature vectors. Alternatively, attempts at training CNNs with external video data in an end-to-end manner have been less than optimal [104], attributed to the insufficient availability of on-road video data pertinent to maneuver anticipation for training such

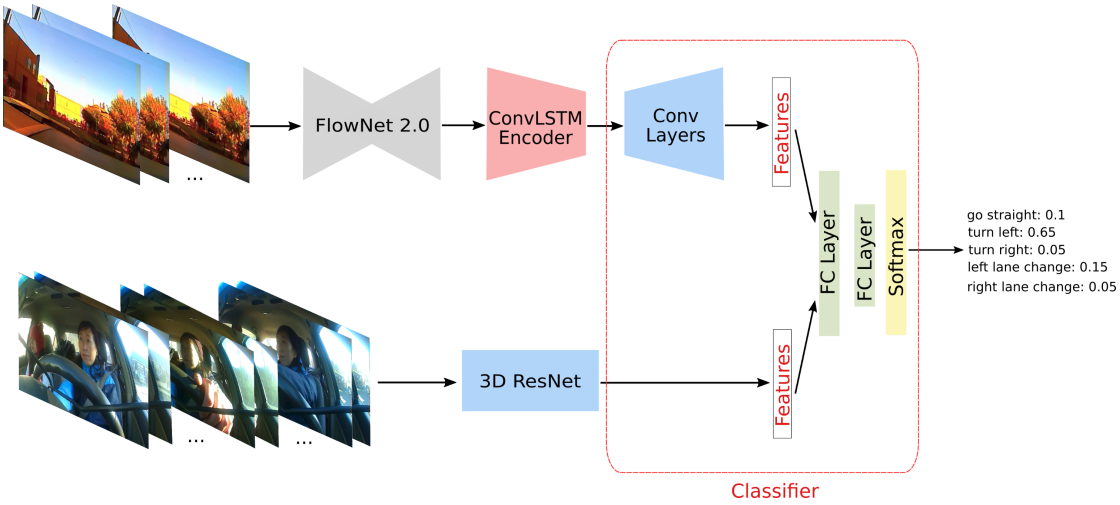


Figure 4.1: The overview of our framework. The upper branch depicts the feature extraction from out-cabin videos: FlowNet 2.0 [6] extracts the optical flow from the consecutive frames; then the traffic motion is captured by a ConvLSTM-based encoder. The bottom branch represents the feature extraction from in-cabin videos based on the 3D ResNet-50 network. The red frame at the end refers to the classifier, where a decoder (marked as "Conv Layers") for outside features is integrated. This novel classifier architecture allows features from inside and outside of the cabin to be considered jointly.

deep networks. This section introduces an approach that combines outside video with in-cabin driver video features for enhanced prediction of driving intentions. Figure 4.1 illustrates the framework architecture.

4.2.1 Driver Maneuver Prediction Framework

Future Frame Prediction. Utilizing the ConvLSTM framework [183], we develop a network structured for encoder-decoder-based training, focusing on motion prediction and feature extraction. Thanks to its convolutional nature, this framework effectively addresses issues in forecasting spatio-temporal sequences [36]. An illustration of this architecture is presented in Figure 4.2. In this structure, $h_{i,j}$ represents the hidden state and $c_{i,j}$ the cell state, where i corresponds to the specific time step and j to the layer level. It is important to note that all states at $i = 0$ are initially set by the network.

The input comprises a set of five optical flow images, denoted as X_i where i is less than 5 and belongs to the set of integers $\mathbb{Z}$. We selected the number five for the input size to ensure a consistent sampling across a duration ranging from one second (30 frames) to five seconds (150 frames). In this context, "consistent" implies that the spacing L between each input image is the same. The decoder's output is the anticipated frame for a future point in time, specifically L frames ahead. Unique to our design, the decoder uses a point-wise convolutional layer, distinguishing our model from prior approaches in [36, 184]. This configuration enables the encoder to efficiently compress motion data

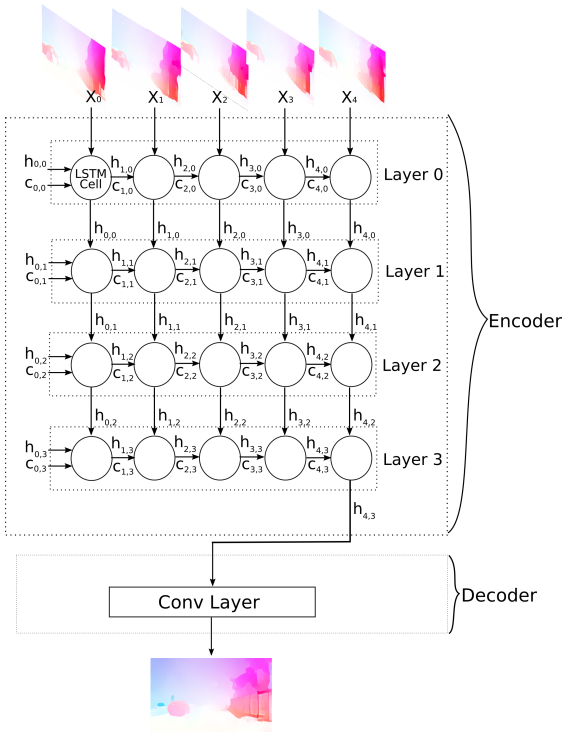


Figure 4.2: Architecture of the proposed future motion prediction module.

from the five-frame input, which is crucial for predicting future movements. Hence, the encoder functions primarily as a motion feature extractor, which leads to a reduced emphasis on the decoder's role.

The network architecture details are presented in Table 4.1. The third column lists the output size for each layer, which is a four-dimensional measurement: time step as the first dimension, followed by the number of channels, and finally the input image's height and width. Each ConvLSTM cell processes a single frame at each time step, resulting in the first dimension being reduced to one post the input layer. It is also important to note that the encoder's output yields the necessary feature for predicting maneuvers.

Feature Fusion. The method presented utilizes two input sources: inside and outside videos, as depicted in Figure 4.1. In processing traffic videos, FlowNet 2.0 first converts raw video frames into optical flow images. Then, these images are given into the ConvLSTM encoder. This encoder generates a 3D feature (32×112×176) that undergoes further processing through several convolutional blocks (Conv-Block) before fusion. Another pathway employs a 3D ResNet-50 for processing driver videos, maintaining the core structure of the original network mentioned in [185]. To enhance model performance and avoid overfitting, a dropout layer follows the average pooling layer at the end. The feature extracted here is the input to the final fully-connected (FC) layer of ResNet-50,

Layer	Kernel Size / Stride	Output size
Input		5×3×h×w
Layer 0	(3,3)/(1,1)	1×128×h×w
Layer 1	(3,3)/(1,1)	1×64×h×w
Layer 2	(3,3)/(1,1)	1×64×h×w
Layer 3	(3,3)/(1,1)	1×32×h×w
Conv	(1,1)/(1,1)	1×3×h×w

Table 4.1: The convolution information about the future motion prediction module.

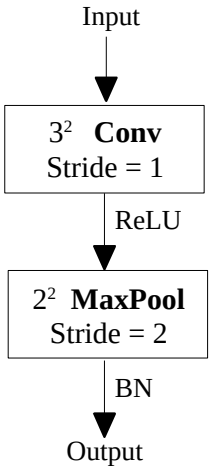


Figure 4.3: The architecture inside the Conv-Block.

represented as a 2048-dimensional vector. The ResNet-50 receives a 16-frame video clip as input.

The novel aspect of the classifier presented is its decoder that processes outside features, which is trained jointly with the features from inside videos. This is detailed in Table 4.2. For interpreting outside motion, a Conv-Block is utilized. The internal configuration of a Conv-Block is depicted in Figure 4.3. Here, "ReLU" denotes the activation function, while "BN" stands for Batch Normalization layer. Between the final two FC layers, both a ReLU and a BN layer are incorporated. The output dimension following each layer is indicated in the table's third column. Finally, N_{cls} signifies the total number of classes, which amounts to five in this specific instance.

4.2.2 Experimental Results

Dataset. The Brain4Cars dataset [101] comprises videos that capture driver actions (resolution of 1088px × 1920px at 25 frames per second) and external environmental

Feature	Layer	Output size
Outside	Conv-Block 0	64×37×59
	Conv-Block 1	128×12×20
	Conv-Block 2	256×4×7
	Conv-Block 3	512×1×2
	Concatenate	3072×1
Both	FC 0	3072×2048
Both	FC 1	2048× N_{cls}
Both	Softmax	N_{cls}

Table 4.2: The architecture of the proposed classifier, which considers joint features from in- and outside videos. The first column indicates the feature source, the second column shows the name of the layer, and the third column is the output size after the layer. The features are combined in the "Concatenate" layer.

video length [s]	> 4	> 3	> 2	> 1	> 0
samples	490	542	563	573	585

Table 4.3: The number of the valid samples relatively to the video length.

footage (resolution of 480px × 720px at 30 frames per second), both recorded concurrently. This dataset categorizes five types of driving maneuvers: *proceeding straight, shifting to the left lane, turning left, moving to the right lane, and turning right.*

This dataset specifically focuses on the driver's behavior before a maneuver, meaning that the actual maneuver does not take place within the video duration. In our research, we emphasize the early detection abilities of our models. For this purpose, we consider each second as a critical juncture. During model assessment, we utilize video frames leading up to a specific time step T, where T belongs to the set (-5, -4, -3, -2, -1), indicating the seconds before the maneuver occurs. The "−" symbol denotes the time in seconds preceding the maneuver. Shorter videos correspondingly depict a briefer period prior to the start of a maneuver. Owing to varying video lengths, the amount of data available for early prediction analysis differs. Additionally, any instances lacking simultaneous internal and external footage are deemed invalid and excluded from our analysis. Table 4.3 details the number of valid video samples available for training our comprehensive framework, with respect to the duration (s) before a maneuver begins.

Out-cabin Motion Extraction. To ensure the generalization, the training incorporated temporal augmentation. This involved randomly selecting and trimming a 5-frame sequence as input for the network, with the desired output being the L-th frame following the sequence's final frame. Spatially, the frames are downscaled to 112 × 176 dimensions,

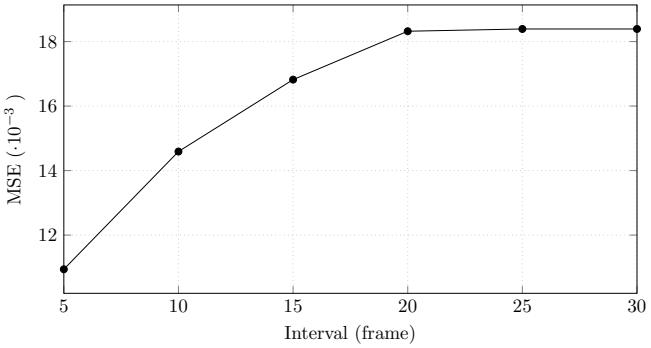


Figure 4.4: MSE for different interval values.

maintaining the original aspect ratio. For optimization, we utilized Mean Square Error (MSE) as the loss function and Stochastic Gradient Descent (SGD) as the optimizer, setting the weight decay at 0.001 and momentum at 0.9. The training regimen spanned 60 epochs, employing a learning rate of 0.1.

In the evaluation, we focus on determining the model's capacity for future prediction. Specifically, our assessment involves using our model for time frames ranging from $L \in (5, 10, 15, 20, 25, 30)$ frames. In this context, the model's decoder generates predictions for the motion in the L-th frame following the final input frame. Consequently, a greater L value indicates a longer prediction into the future, with the upper limit set at 30 frames (equivalent to 150 frames or a 5-second video, which is the maximum duration in our dataset). Intervals shorter than 5 frames (less than 0.33 seconds) are not considered due to their brevity. The primary frame of interest for evaluation is the video final frame, and the Mean Squared Error (MSE) serves as the benchmark for assessment. The Figure 4.4 illustrates the average MSE across various intervals.

Note that we amplify the MSE value by a factor of 1000 for clarity in demonstrating differences. The findings indicate a challenge for the model in accurately forecasting frames that are significantly ahead of time. Specifically, the model learning efficacy diminishes when tasked with predicting beyond a 20-frame interval, equivalent to 0.67 seconds. To ensure more accurate motion feature representation, we select a model configuration with an L value of 5.

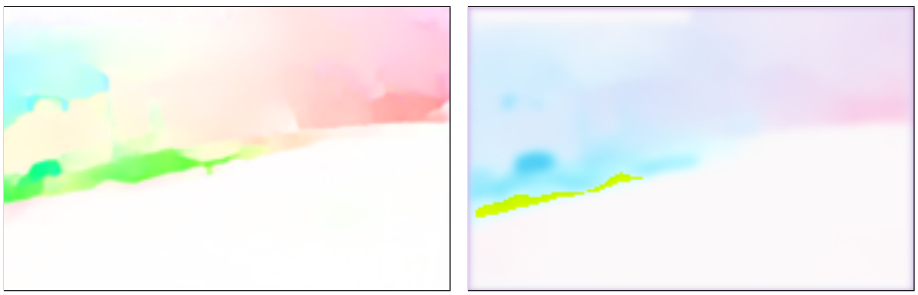
Upon fixing the interval L at 5, we assess our model over various durations within the video. In detail, the frames fed into the model all fall within the timeframe preceding T (with T taking values from the set -4, -3, -2, -1, 0), and the target frame is identified as the final frame in each one-second segment. For a comprehensive evaluation, we employ three evaluation metrics: Mean Squared Error (MSE), Structural Similarity (SSIM) index, and Peak Signal-to-Noise Ratio (PSNR). The predictive performance is detailed in Table 4.4. Notably, for PSNR and SSIM, higher values indicate superior performance. The results across five different folds are presented as "Average (Avg) ± Standard Error (SE)".

The findings indicate that optimal prediction of maneuvers is attained using video data captured 4 to 5 seconds prior to the maneuver itself. Thus, motion changes are

prediction at [s]	MSE ($\cdot 10^{-3}$)	SSIM	PSNR
-4	9.13 ± 0.42	0.909 ± 0.001	21.77 ± 0.16
-3	9.42 ± 0.40	0.906 ± 0.002	21.49 ± 0.10
-2	10.75 ± 0.61	0.904 ± 0.002	21.35 ± 0.18
-1	9.97 ± 0.22	0.900 ± 0.001	21.27 ± 0.05
0	10.73 ± 0.46	0.898 ± 0.002	21.08 ± 0.10

Table 4.4: Results of future motion prediction.

not massive earlier on before $-3s$. In scenarios with considerable motion variations, such as during a car turn, the encoder struggles to fully capture the extent of these changes. Notably, in the final three seconds leading up to a maneuver, external motion changes become increasingly evident. Although motion continues to evolve between $2s$ and $1s$ before the maneuver, the changes are less pronounced compared to adjacent time intervals. Overall, critical traffic motion changes are detectable within a three-second window preceding the maneuver. Using the features extracted from the outside videos by the ConvLSTM-encoder alone can also produce a prediction among five classes. The results are presented in Table 4.5, whereas a comparison to related approaches is provided in Table 4.6.



(a) Target image. (b) Predicted image.

Figure 4.5: The comparison of target and the predicted image.

In-cabin Action Recognition. The 3D ResNet-50 is utilized for inside feature extraction due to its proven effectiveness in recognizing human actions, as detailed by Hara et al. [185]. Although end-to-end training typically necessitates a substantial dataset, which Brain4cars lacks, we adapt by employing a Kinetics-pretrained 3D ResNet-50 model [185] and subsequently fine-tuning it using the Brain4cars internal video dataset.

To mitigate overfitting, spatial and temporal data augmentation techniques were implemented. Spatially, we use random cropping (primarily focusing on the driver's side),

scaling, and horizontal flipping. It is important to note that when the augmentation affects directionality (left/right), the corresponding labels are adjusted. Temporally, we uniformly extract short segments from each second. These segments form a 16-frame sequence serving as the input for the 3D ResNet-50, with an input resolution of 112×112. Additionally, an extra dropout layer is incorporated before the final FC layer during the training phase. We set a dropout rate to 0.5 and utilize cross-entropy loss as our loss function. The training spans 60 epochs, starting with a learning rate of 0.1 and decreasing by a factor of 0.1 after the 30th and 50th epochs. The chosen optimizer is SGD with a momentum of 0.9 and a weight decay of 0.001. For evaluation purposes, we use frames captured at the end of each second up to time T ($T \in (-4, -3, -2, -1, 0)$) to compile the 16-frame input required by the 3D ResNet.

The primary component of the 3D ResNet-50, trained for this purpose, serves as the feature extraction mechanism. Features extracted just before the ultimate fully connected (FC) layer are inputted into the final classification system. The outcomes using only this module (the inside video) for classification are detailed in Table 4.5, and a comparative analysis with the SOTA method is presented in Table 4.6.

Feature Fusion. The ConvLSTM model and the 3D ResNet-50 model are trained independently, leading to the extraction of features from both inside and outside the video through these two distinct modules. The feature derived from the outside of the video forms a volume with dimensions of $32 \times 112 \times 176$, while the feature from inside the video is represented as a vector of 2048 elements. The evaluation process takes various time periods into account, similar to the approach used in both modules.

The metrics used to evaluate performance are accuracy and the F_1-score. This F_1-score incorporates both the precision (Pr) and recall (Re) metrics of a classifier, as delineated in Equation (4.1). The term n denotes the total number of classes, while Ω represents the entire set of classes identifiable by our model. This set includes four specific maneuvers plus an additional category for "no maneuver." The variable TP_i signifies the count of correctly identified instances for class i. Furthermore, P_i and N_i refer to the quantities of samples predicted as class i and those actually labeled as class i, respectively.

$$\begin{aligned} Pr &= \frac{1}{n} \sum_{i \in \Omega} \frac{TP_i}{P_i} \\ Re &= \frac{1}{n} \sum_{i \in \Omega} \frac{TP_i}{N_i} \\ F_1 &= \frac{2 \cdot Pr \cdot Re}{Pr + Re} \end{aligned} \tag{4.1}$$

Table 4.5 presents the precision and F_1 scores in percentages, for various intervals leading up to a maneuver, utilizing diverse data inputs. The proximity to the initiation of the maneuver enhances both the accuracy and F_1 scores, regardless of the data source variance. Commonly, the initial phase of any maneuver (or the absence thereof) is characterized by linear progression. Consequently, the longer the observation duration by the model, the more precise its judgments become. These outcomes suggest that early

Inside video	Time period	Acc (%)	F_1 (%)
	[-5,-4]	56.49 ± 0.02	48.19 ± 0.03
	[-5,-3]	63.63 ± 0.02	58.46 ± 0.02
	[-5,-2]	70.48 ± 0.02	68.63 ± 0.03
	[-5,-1]	75.73 ± 0.01	73.09 ± 0.01
	[-5,0]	77.40 ± 0.02	75.49 ± 0.02
Outside video	Time period	Acc (%)	F_1 (%)
	[-5,-4]	44.08 ± 0.01	38.91 ± 0.03
	[-5,-3]	44.22 ± 0.01	38.75 ± 0.01
	[-5,-2]	50.43 ± 0.01	46.98 ± 0.01
	[-5,-1]	59.53 ± 0.01	62.37 ± 0.01
	[-5,0]	60.87 ± 0.01	66.38 ± 0.03
In- & outside	Time period	Acc (%)	F_1 (%)
	[-5,-4]	59.13 ± 0.02	53.35 ± 0.02
	[-5,-3]	64.93 ± 0.02	60.33 ± 0.01
	[-5,-2]	72.07 ± 0.02	70.56 ± 0.02
	[-5,-1]	79.92 ± 0.02	78.90 ± 0.01
	[-5,0]	83.98 ± 0.01	84.30 ± 0.01

Table 4.5: The results of using the proposed framework with different input data sources. The results of five folds are shown in the form: "Avg ± SE".

detection of maneuvers is feasible. For instance, when employing dual video inputs, the model accurately forecasts 71.72% of maneuvers two seconds before they begin..

Utilizing both video sources across various time periods yields the most effective outcomes. Relying solely on external videos leads to inferior results compared to using the other two sources. The subpar performance of external data can be attributed to the auto-encoder's limitation of offering only the motion feature of a forthcoming frame. In contrast, the internal feature encapsulates data over an extended duration. Furthermore, it is observed that significant motion typically happens within the three seconds preceding maneuvers. This is particularly evident in the period from -4 to -2, where there is a notable enhancement in both accuracy and F_1.

In Table 4.6, we compare our findings with those presented in [104], as both studies employ end-to-end training and explore effectiveness using three distinct data sources. Our comparison includes model accuracy, F_1 scores, and the parameter count. All reported results are based on a zero time-to-maneuver scenario and are validated through a 5-fold cross-validation process.

Our method surpasses the approach in [104], except using inside video only. It is because that we choose the 3D ResNet-50 over the 3D ResNet-101 utilized in [104]. The 3D ResNet-101 contains about double the parameters compared to our chosen 3D ResNet-

Method	Data Source	Acc (%)	F_1 (%)	Param.(M)
[104]	inside only	83.1 ± 2.5	81.7 ± 2.6	85.26+m
	outside only	53.2 ± 0.5	43.4 ± 0.9	85.26+m
	in-&out-side	75.5 ± 2.4	73.2 ± 2.2	170.52+m
our	inside only	77.40 ± 0.02	75.49 ± 0.02	46.22
	outside only	60.87 ± 0.01	66.38 ± 0.03	5.41+m
	in-&outside	$\mathbf{83.98} \pm 0.01$	$\mathbf{84.30} \pm 0.01$	**57.92**+m

Table 4.6: Comparison of our proposed framework with other method. The results of five folds are shown in the form: "Avg ± SE". In order to show a clear difference, we use "m" to represent the number of parameters in FlowNet2.0, which is a common module in both methods.

50. Our decision to opt for a smaller ResNet model was driven by the need to prevent overfitting issues when tuning a significantly large network with a limited dataset. Additionally, we require a model with lower resource consumption, considering its suitability for automotive applications. Our methodology demonstrates superior performance over the prior model, with considerably fewer parameters, through a dual-stream input approach. It attains an average accuracy of 83.98% and an average F_1 score of 84.30% across five folds, exceeding the previous model by 8.48 percentage points in accuracy and 11.1 percentage points in F_1. Focusing exclusively on exterior videos, our model outperforms the earlier one by 7.67 percentage points in accuracy and 22.98 percentage points in F_1. Our model is efficient in extracting valuable features from outside videos using fewer parameters. More importantly, unlike the previous model, ours does not encounter performance deterioration due to external videos. This suggests that inside and outside videos have complementary information.

4.3 Driver Attention-based Object Detection

Existing driver gaze prediction models utilize features derived from deep neural networks, typically employed in image classification or object recognition tasks, such as AlexNet [119] or VGG [186]. These features are then processed through decoding modules to generate detailed, pixel-specific saliency maps. Our proposed method, illustrated in Figure 4.7, aims to determine the objects that capture a driver's attention, utilizing a grid-based approach for saliency map prediction. This technique involves the concurrent operation of an object detector and an attention predictor, both utilizing the same image features in a manner that conserves computational resources. This section will introduce the attention-based object detection framework, as detailed in Section 4.3.1, which

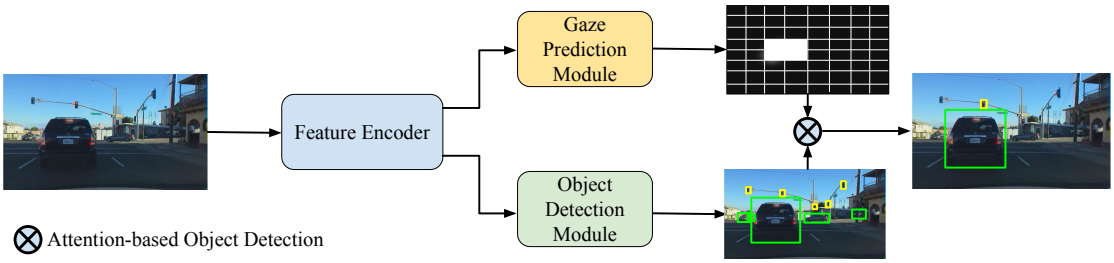


Figure 4.6: Overview of our proposed critical object detection framework. The **feature encoder** extracts features from the input image. The **gaze prediction module** predicts driver attention in a grid-based saliency map and the **object detection module** detects all the objects in the traffic using extracted features. The **attention-based objects** are detected and returned to users based on the predicted saliency map and detected objects.

contains the gaze prediction module and the object detection algorithm, among other components. In Section 4.3.2, the specific architecture of the model will be discussed.

Contributions in this section can be summarized as follows: (1) We propose a framework to predict objects that human drivers pay attention to while driving. (2) Our proposed grid-based attention prediction module is very flexible and can be incorporated with different object detection models. (3) We evaluate our model on two datasets, BDD-A and DR(eye)VE, showing that our model is computationally more efficient and achieves comparable performance compared to other state-of-the-art driver attention prediction models.

4.3.1 Algorithm Details

The described framework is structured in the following manner: An RGB image from driving scenes, denoted as X, belongs to the space $\mathcal{R}^{3\times H\times W}$, with H and W representing the image height and width, respectively. An image feature encoder, $\mathcal{E}(\cdot)$, processes X to produce a feature v. This feature, v, exists in the space $\mathcal{R}^{c_v\times h_v\times w_v}$, where h_v, w_v, and c_v refer to the height, width, and channel count of the feature map. The gaze prediction module, $\mathcal{G}(\cdot)$, takes v as input and outputs a grid-vector $\hat{y} = \mathcal{G}(v)$. Following this, $\hat{y}$ is transformed by $\mathcal{T}(\cdot)$, resulting in a 2D saliency map $\hat{M}$ in the space $\mathcal{R}^{H\times W}$. The object detection module $\mathcal{D}(\cdot)$ identifies a series of objects within the image, represented as $\mathbf{O} = \mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_n}$, where each $\mathbf{o_i}$ contains data on the bounding box and class of the object, and n is the count of detected objects. The framework employs an attention-based object detection operation $\bigotimes$, which, utilizing both $\hat{M}$ and $\mathbf{O}$, identifies a subset of focused objects $\mathbf{O^f}$, expressed as $\hat{M} \bigotimes \mathbf{O} = \mathbf{O^f}$, where the size of $\mathbf{O^f}$ is at most n. Figure 4.7 illustrates the various components of this framework.

Gaze prediction module. To minimize computational demands, we suggest predicting the gaze saliency map using a grid-based approach, effectively transforming the saliency map creation into a multi-label prediction task. Specifically, we convert the original

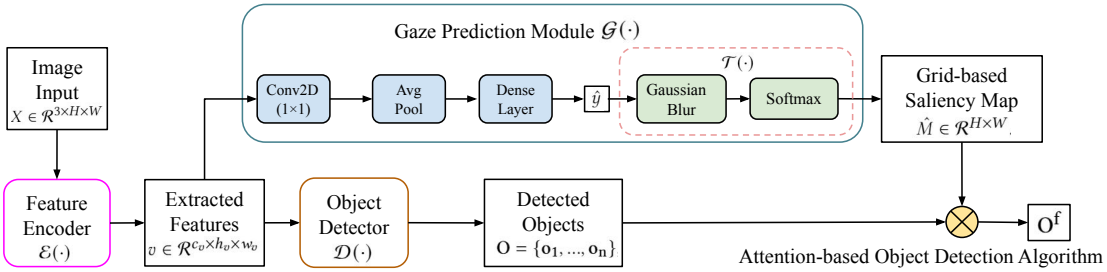


Figure 4.7: Overview of our proposed driver attention-based object detection framework.

saliency map $M \in \mathcal{R}^{H\times W}$ into a grid-vector $y \in \mathcal{R}^{n\cdot m}$, with n and m representing the number of grid cells vertically and horizontally. Each element in the grid-vector y holds a binary value, indicating whether a specific region in the gaze map is focused (1) or not (0). To derive the grid-vector y from the saliency map M, we follow these steps: (1) Transform M into M' by binarizing it at 15% of the maximum pixel value (values above this threshold become 1, others 0). (2) Calculate the probability of focus for each grid cell in y as $p = \frac{\sum M'_j}{\sum M'}$, where $\sum M'_j$ is the total pixel values in the j-th grid cell, and $\sum M'$ is the total of all pixels. (3) Set the region's entry to 1 if its focus probability exceeds the threshold $\frac{1}{n\cdot m}$, or to 0 otherwise. Figure 4.8 illustrates this process.

In the context of a grid configuration with dimensions n and m, we utilize the encoded representation $v = \mathcal{E}(X)$ alongside the grid-vector y, which is derived from the actual saliency map M. For training the gaze prediction component $\mathcal{G}(\cdot)$, the method of binary cross-entropy loss is employed:

$$L(\hat{y}, y) = -\frac{1}{K}\sum_{i=1}^{K} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot (1 - log(\hat{y}_i)) \tag{4.2}$$

where $\hat{y} = \mathcal{G}(v)$, with the variable K signifying the product of n and m, which denotes the total count of grid cells.

To create a 2D saliency map, the process involves generating $\hat{M} = \mathcal{T}(\hat{y})$. In detail, each element of $\hat{y}$ corresponds to a specific cell in the 2D grid (refer to Figure 4.8). These cells are populated with their respective values from $\hat{y}$. The dimensions of each cell are given by $\frac{H}{n} \times \frac{W}{m}$, resulting in the formation of an $n \times m$ 2D matrix. Following this, a Gaussian blur and softmax operation are employed to refine the 2D matrix, which serves as the estimated saliency map $\hat{M}$. The top section of Figure 4.7 illustrates the method used for generating a grid-based saliency map.

Attention-based object detection. The object detector $\mathcal{D}(\cdot)$ processes the input v and identifies all object information $\mathbf{O}$, which includes their classes and bounding boxes. This detector works in tandem with our feature encoder $\mathcal{E}(\cdot)$, together constituting a complete object detection network. For effective training of this object detector, it is essential to have a comprehensive image dataset, richly annotated with both bounding boxes and class information. We leverage the pre-trained parameters of existing, well-established

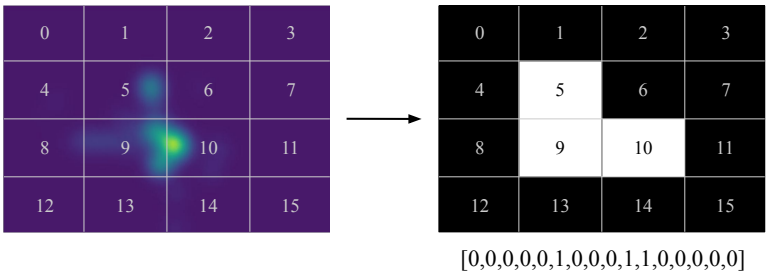


Figure 4.8: Illustration of transforming a saliency map into a grid-vector. The used grid here is 4×4. Grid cells 5, 9, and 10 reach the threshold, therefore the grid-vector y for the saliency map M is $[0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0]$.

object detection models like YOLOv5 [187] for our $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$. The architectural design specifics will be elaborated in the subsequent section. It is important to note that our approach does not necessitate additional training for $\mathcal{E}(\cdot)$ or $\mathcal{D}(\cdot)$, which significantly accelerates the training process of our framework. Regarding the attention-based object detection operation $\bigotimes$, it functions as follows: for each detected object $\mathbf{o_i}$ in $\mathbf{O}$, the maximum pixel value within its bounding box on the saliency map $\hat{M}$ is used as the likelihood of the object $\mathbf{o_i}$ being the focus. We can set a threshold Th to ascertain if $\mathbf{o_i}$ is focused on by drivers. This threshold Th is adjustable based on user requirements for various metrics, such as precision or recall.

4.3.2 Model Details

In this framework, we employ three pretrained object detection frameworks as our feature encoder $\mathcal{E}(\cdot)$: YOLOv5 [187], Gaussian YOLOv3 [188], and CenterTrack [189]. This approach assesses the effectiveness and versatility of our gaze prediction method. In detail, we incorporate certain layers from YOLOv5 (specifically the small-sized version, release v5.0) up to but not including the final CSP-Bottleneck (Cross Stage Partial [190]) layer in the neck structure (PANet [191]). The rest of the YOLOv5 model (namely, the detector layer) serves as our object detector $\mathcal{D}(\cdot)$. Likewise, a portion of the YOLOv3 network (the first 81 layers) is used as $\mathcal{E}(\cdot)$, and we utilize the "keypoint heatmaps" from each category in CenterTrack [189]. The specific dimensions of the extracted feature v and the output dimensions after each layer in the gaze prediction module are detailed in Table 4.7. When using YOLO architectures, the convolutional layer with a 1×1 kernel size reduces the input channels to 16, whereas it narrows them down to a single channel with CenterTrack features. An average pooling layer is applied to lessen the computational load on the dense layer by diminishing the width and height of the feature maps. Before entering the dense layer, all features are converted into vector forms. The dense layer concludes with a sigmoid activation function, producing an output $\hat{y} \in \mathcal{R}^{n \cdot m}$.

Feature Encoder $\mathcal{E}(\cdot)$		Gaze Prediction $\mathcal{G}(\cdot)$		
Backbone	v	Conv	Avg Pooling	Dense Layer
YOLOv5 [187]	$512 \times 12 \times 20$	$16 \times 12 \times 20$	$16 \times 6 \times 10$	number of grid cells
Gaussian YOLOv3 [188]	$1024 \times 13 \times 13$	$16 \times 13 \times 13$	$16 \times 7 \times 7$	number of grid cells
CenterTrack [189]	$80 \times 72 \times 128$	$1 \times 72 \times 128$	$1 \times 18 \times 32$	number of grid cells

Table 4.7: Network architecture details when using different object detectors. Column "Feature Encoder" shows the used backbone for extracting feature v and the dimension of v. Column "Gaze Prediction" demonstrates the dimension of output after each layer.

4.3.3 Implementation Details

Datasets. Experiments are conducted on two datasets, BDD-A and DR(eye)VE. The **BDD-A** dataset, as detailed in [7], comprises 1426 ten-second videos recorded in busy areas with many objects on the roads. The dataset contains 926 training videos, 200 for validation, and 300 for testing. From these, three frames per second are extracted, and after discarding invalid gaze maps, the training set contains 30158 frames, with 6695 and 9831 frames in the validation and test sets, respectively. Table 4.8 presents the data for the objects that garnered the most attention in the test set. Per frame, an average of 7.99 cars are present ("Total"), but only 3.39 typically catch the driver's gaze ("Focused"). While each frame typically includes 0.94 traffic lights, drivers generally notice just 0.18, focusing mainly on those relevant to their driving direction. Overall, each frame contains about 10.53 objects, with roughly 40% (4.21 objects) capturing the driver's focus, making the accurate detection of these objects a challenging task.

Object	**Person**	**Bicycle**	**Car**	**Motorcycle**	**Bus**	**Truck**
Total	0.78	0.03	7.99	0.03	0.18	0.48
Focused	0.24	0.02	3.39	0.01	0.11	0.25
Object	**Traffic light**	**Fire Hydrant**	**Stop Sign**	**Parking Meter**	**Bench**	***Sum***
Total	0.94	0.02	0.05	0.004	0.002	10.53
Focused	0.18	0.002	0.008	-	-	4.21

Table 4.8: Traffic-related class analysis on BDD-A test set: The values in the table show the average number of objects in one video frame. "Total" means detected objects while "focused" means attended objects by the human driver. "-" refers to a number smaller than 0.001. "Sum" includes also non-traffic objects.

The DR(eye)VE dataset, as detailed in [192], comprises 74 videos. From its test set, we select five videos (specifically numbers 66, 67, 68, 70, and 72) for analysis. These videos are chosen for their diversity in aspects such as time, driver, landscape, and weather conditions. Each of the videos has a duration of 5 minutes, with a frame rate of 25 FPS, leading to a total of 7500 frames per video. After discarding frames with invalid gaze map data, the resulting test set consisted of 37,270 frames. We use a pretrained YOLOv5 network to these videos, and the findings are presented in Table 4.9. In comparison with the BDD-A dataset (Table 4.8), the DR(eye)VE dataset features a more uniform

environment with less variety of objects on the roads. The average count of objects per frame is 3.24, with 39% of these objects being the focus of drivers' attention, a percentage comparable to that in the BDD-A dataset.

Object	**Person**	**Bicycle**	**Car**	**Motorcycle**	**Bus**	**Truck**
Total	0.07	0.009	2.35	0.003	0.026	0.09
Focused	0.02	0.004	1.06	-	0.01	0.04
Object	**Traffic light**	**Fire Hydrant**	**Stop Sign**	**Parking Meter**	**Bench**	***Sum***
Total	0.46	-	0.02	0.005	0.003	3.24
Focused	0.07	-	0.002	0.003	-	1.26

Table 4.9: Traffic-related class analysis on DR(eye)VE dataset (test set): The value is the average number of objects in each video frame. "Total" means detected objects while "focused" means attended objects by the human driver. "-" refers to the number smaller than 0.001. "Sum" also includes non-traffic objects.

Evaluation metrics. Our evaluation of the models is conducted across three dimensions: object detection (at the object level), saliency map creation (at the pixel level), and the consumption of computational resources. For the appraisal of the gaze map quality, we use Kullback–Leibler divergence (D_{KL}) and Pearson's Correlation Coefficient (CC), in line with prior studies [7, 8, 15]. The saliency maps, both predicted and ground truth, are resized to 36×64, maintaining the original aspect ratio, following the approach in [7]. To ensure a fair comparison, we normalize the size of saliency maps from various models to 36×64, as recommended by Xia et al. [7].

In our object detection assessment, the ground-truth "focused" objects are identified by applying our attention-based object detection to all objects detected using the YOLOv5 model and the ground-truth gaze saliency maps, $M \bigotimes \mathbf{O}$. Here, the highest value within the object's bounding area is taken as the probability. An object is deemed "focused on" if this probability exceeds 15%, a threshold set to exclude objects less likely than a random choice (typically ten objects per frame, as detailed in table 4.8). For evaluation purposes, each object is treated as a binary classification task: whether it is the driver's focus or not. The metrics utilized for this evaluation are Area Under the ROC Curve (AUC), precision, recall, F_1 score, and accuracy. Lastly, to quantitatively assess and compare the computational demands of our models, we take into account the number of trainable parameters and the floating-point operation per second (GFLOPs) required by the networks.

Training details. This work utilizes a single NVIDIA CUDA RTX A4000 GPU to perform all experiments. Our gaze prediction module underwent training on the BDD-A training set for 40 epochs, employing the Adam optimizer [193], and underwent validation on the corresponding validation set. We initiate the learning rate at 0.01, reducing it by a factor of 0.1 after each set of 10 epochs. Both the feature encoder and the object

	Object-level					Pixel-level	
	AUC	*Prec (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	D_{KL}	*CC*
2×2	0.58	43.86	88.97	58.75	50.05	2.35	0.18
4×4	0.76	52.43	**91.50**	66.66	63.40	1.61	0.41
8×8	0.84	57.87	89.16	70.18	69.71	1.27	0.55
16×16	**0.85**	71.98	73.31	**72.64**	77.92	1.15	0.60
32×32	**0.85**	**75.47**	68.79	71.97	**78.58**	**1.13**	**0.62**

Table 4.10: Comparison of using different grid settings on object- and pixel-level performance (Th=0.5). For all metrics except D_{KL}, a higher value indicates better performance. The best result is marked in bold.

detector are pretrained[1], and no additional fine-tuning in the object detection phase is conducted.

4.3.4 Evaluation on BDD-A

Different Grids. In our gaze prediction module, we experiment with various grid configurations, ranging from a 2×2 arrangement ($n = m = 2$) to a more complex 32×32 grid ($n = m = 32$), with each step doubling in size. The backbone for these experiments is consistently YOLOv5. A comparative analysis of these different grid sizes is detailed in Table 4.10. The term "Pixel-level" denotes the assessment of the saliency map using D_{KL} and CC metrics, while "Object-level" refers to the effectiveness of attention-based object detection. To ensure a fair comparison across different settings, we maintain a detection threshold Th of 0.5 for identifying attended areas. This comparative study indicates that finer grids tend to yield better performance. However, it is noteworthy that the improvement when upgrading from 16×16 to 32×32 grids is marginal, with the AUC values being nearly identical. Consequently, for the sake of computational efficiency, the 16×16 grid configuration is selected for all subsequent experiments.

Different thresholds. The influence of varying Th on attention-based object detection is summarized in Table 4.11. Our findings indicate that a reduced Th enhances the recall score, whereas an increased Th elevates the precision score. Optimal F_1 score is obtained at a Th value of 0.4, and the highest accuracy is achieved with Th at 0.6. Adjusting Th to 0.5 results in commendable performance, registering an F_1 score of 72.64% and accuracy of 77.92%. Th serves as a configurable hyperparameter, allowing users to tailor it based on the specific needs of their applications. For instance, setting a higher Th is advisable for applications where high precision is crucial.

[1]Links for pretrained parameters are available for YOLOv5 at `https://github.com/ultralytics/yolov5`, YOLOv3 at `https://github.com/motokimura/PyTorch_Gaussian_YOLOv3`, and CenterTrack at `https://github.com/xingyizhou/CenterTrack`.

	Prec	Recall	F_1	Acc
0.3	63.76	**83.33**	72.24	74.39
0.4	68.11	78.36	**72.88**	76.68
0.5	71.98	73.31	72.64	77.92
0.6	75.81	68.09	71.74	**78.55**
0.7	**79.61**	62.04	69.73	78.47

Table 4.11: Comparison of different Th using 16×16 grids on attention-based object detection. Results are shown in % and for all metrics, a higher value indicates better performance. The best result is marked in bold.

Comparison with the state-of-the-art. We compare our algorithm with three newly proposed models, namely YOLOv5, Gaussian YOLOv3, and CenterTrack, and four established saliency models: BDD-A [7], DR(eye)VE [8], ML-Net [9], and PiCANet [10][2].

Our evaluation covers three metrics: object detection, the creation of gaze saliency maps, and the cost in terms of resources. We employ the YOLOv5 object detector for identifying objects in images, followed by the application of our attention-focused object detection method $\bigotimes$, which utilizes saliency maps produced by each model. The term "Baseline" denotes the mean saliency map from the BDD-A training set, as depicted in Figure 4.10 (b). To impartially assess object-level scores like precision, recall, F_1, and accuracy, dependent on the threshold Th, we determine for each model the optimal Th that maximizes the true positive rate (TPR) and minimizes the false positive rate (FPR). This involves generating a ROC curve (Receiver Operating Characteristic) for each model using the BDD-A test set and identifying the Th corresponding to the closest point to (0,1) on the curve, calculated as $argmax(\sqrt{TPR \cdot (1 - FPR)})$. The ROC curves and specific Th values for each model are presented in Figure 4.9. The outcomes of these comparisons across various models are detailed in Table 4.12.

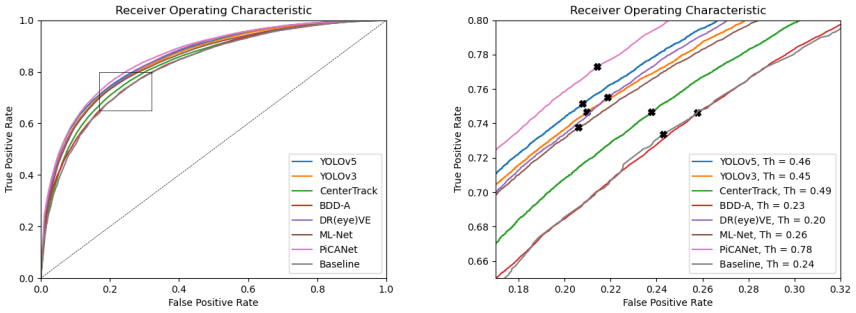


Figure 4.9: ROC curves and computed thresholds on the BDD-A. On the right, the curves are zoomed in and the points that belong to the computed thresholds are marked.

[2]Each model received training using the BDD-A dataset. We accessed the BDD-A model's trained parameters from `https://github.com/pascalxia/driver_attention_prediction`, while the others were obtained from `https://sites.google.com/eng.ucsd.edu/sage-net`.

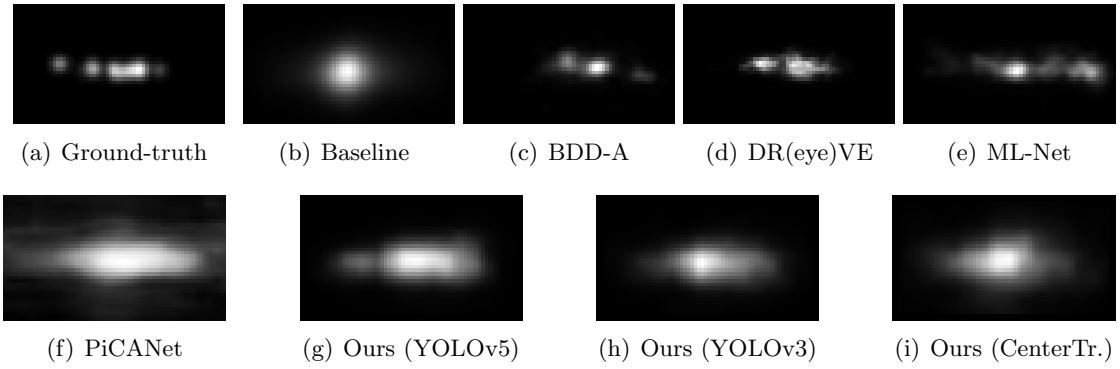
	Object-level					Pixel-level		Resource	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	D_{KL}	*CC*	*Param.(M)*	*GFLOPs*
Baseline	0.82	66.10	74.22	69.92	74.47	1.51	0.47	0.0	0.0
BDD-A [7] †	0.82	66.00	74.33	69.92	74.43	1.52 (1.24)	0.57 (0.59)	3.75	21.18
DR(eye)VE [8] ‡	0.85	70.04	74.94	72.41	77.16	1.82 (1.28)	0.57 (0.58)	13.52	92.30
ML-Net [9]†	0.84	70.48	73.75	72.08	77.15	1.47 (1.10)	0.60 (0.64)	15.45	630.38
PiCANet [10]†	0.86	70.23	77.67	73.76	77.91	1.69 (1.11)	0.50 (0.64)	47.22	108.08
Ours (CenterTrack)*	0.83	68.93	72.83	70.83	76.01	1.32	0.56	19.97	28.57
Ours (YOLOv3)*	0.85	70.25	74.72	72.41	77.24	1.20	0.59	62.18	33.06
Ours (YOLOv5)*	0.85	70.54	75.30	72.84	77.55	1.15	0.60	7.52	17.0

Table 4.12: Comparison with other gaze models on the BDD-A dataset. On object-level, all models are evaluated with detected objects of YOLOv5. Our three models use 16×16 grids. Pixel-level values in brackets are the results reported from the original work [7, 15]. * indicates that the backbone is pretrained on COCO [16], † on ImageNet [17] and ‡ on UCF101 [18]. The resource required for the gaze prediction is listed in the last column.

The AUC scores indicate that our dual YOLO models are on par with other models at an object level, with PiCANet being marginally superior. Despite not being designed for creating pixel-level saliency maps, the metrics of D_{KL} and CC reveal that our YOLOv5-based model, having a D_{KL} value of 1.15 and CC of 0.60, is comparable to its counterparts at the pixel level under our test conditions. Regarding object detection, our YOLO-based pair achieves an AUC of 0.85, which is slightly below PiCANet's 0.86. However, they outperform other models in F_1 scores and accuracy.

Additionally, our gaze prediction model utilizes the same backbone (feature encoder) as the object detection network, requiring only an additional dense layer, thus reducing computational demands. For example, our YOLOv5-based model needs a total of 7.52M parameters, with merely 0.25M being extra for gaze prediction. This leads to the same computational load as a standard YOLOv5 network (17.0 GFLOPs). Overall, the merit of our system lies in its ability to perform gaze prediction with minimal additional computational resources or parameters compared to those required for object detection. In contrast, other models require a separate object detection network to identify attention-centric objects within their existing architectures. For a fair evaluation, we only list the resource requirements for each model saliency prediction in Table 4.12. To match a similar level of object detection performance, for instance, DR(eye)VE necessitates 13.52M parameters and 92.30 GFLOPs just for saliency mapping, which exceeds the requirements of our YOLOv5 framework for both object detection and saliency map computation.

Qualitative results. The qualitative outcomes of saliency map predictions using various models are illustrated in Figure 4.10. In our approach, we employ the architectures of YOLOv5, YOLOv3, and CenterTrack. It is observed that models such as BDD-A, DR(eye)VE, and ML-Net offer more accurate and focused attention predictions. Nevertheless, BDD-A and ML-Net incorrectly emphasize a minor region on the right side instead of the left side. Conversely, our predictions (g) and (h) concentrate on both the



(a) Ground-truth (b) Baseline (c) BDD-A (d) DR(eye)VE (e) ML-Net

(f) PiCANet (g) Ours (YOLOv5) (h) Ours (YOLOv3) (i) Ours (CenterTr.)

Figure 4.10: Comparison of predicted driver attention saliency maps using different models. (a) Ground-truth driver attention map; (b) The baseline saliency map (center-bias); (c-f) Predictions using models [7, 8, 9, 10]; (g-i) Predictions using our framework with different backbones.

center and the right side. Despite our predictions being grid-based, they exhibit finer detail compared to those produced by PiCANet.

Figure 4.11 illustrates an example of attention-based object detection using various models, where objects are enclosed in bounding boxes. This framework is from a video depicting a vehicle approaching and overtaking other vehicles waiting in the right lane at a crossroad. A comparison between (i) and (a) reveals that while the human driver notices several objects, they do not focus on all. Our models, incorporating features from YOLOv5 and CenterTrack, identify all waiting vehicles as objects of driver focus (shown in (b) and (d)), aligning with the actual situation depicted in (a). The BDD-A model, however, highlights a car in the opposite lane and a church clock, overlooking a distant waiting car. Additionally, the consistent prediction of gaze towards the vanishing point is a notable issue in driving saliency models. This case demonstrates that our model does not habitually predict the vanishing point on the road, in contrast to DR(eye)VE, ML-Net, and PiCANet, which often identify the object near the central point as critical.

We include an analysis of two instances where our YOLOv5-based model does not perform as expected in Figure 4.12. The first scenario involves a vehicle overtaking two cyclists by moving from the left to the middle lane. The model accurately identifies the vehicles ahead and the cyclists. However, it mistakenly flags cars parked beyond the cyclists as critical, diverging from the actual situation. This case highlights the impact of attention-based object detection: important elements like the approaching vehicles and cyclists are recognized, but cars parked further away in a different lane are not. In the second example, a vehicle approaches a crossroad with a changing traffic light. The model detects the vehicle ahead slowing down and a car parked to the right. Additionally, it marks a cyclist on the right as critical. This error also indicates that shows that the predictions of our model are not limited to the center part of an image.

(a) Ground-truth driver attention (b) Ours (YOLOv5) (c) Ours (YOLOv3)

(d) Ours (CenterTrack) (e) BDD-A (f) DR(eye)VE

(g) ML-Net (h) PiCANet (i) All objects

Figure 4.11: Comparison of attention-based object detection using different models. (a) Ground-truth attention; (b-d) Predictions using our framework with different backbones; (e-h) Predictions using models [7, 8, 9, 10]; (i) Object detection without driver attention.

4.3.5 Evaluation on DR(eye)VE

Comparison with the state-of-the-art. We conduct an evaluation using the DR(eye)VE dataset to validate the generalization ability of our model, without any additional training. The evaluation involves running the YOLOv5 model on 16×16 grids, benchmarking it against the performance metrics of DR(eye)VE, BDD-A, ML-Net, and PiCANet. Similar to the approach in the BDD-A experiments, we derive the threshold values from the ROC curves as detailed in Figure 4.13. We conduct an object-level assessment using metrics such as AUC, precision, recall, F_1, and accuracy, as well as a pixel-level evaluation using D_{KL} and CC. The findings are detailed in Table 4.13. In our experimental framework, the bottom-up models, ML-Net and PiCANet, outperformed the top-down networks, namely DR(eye)VE and BDD-A. Notably, our model and PiCANet achieve the highest scores at the object level ($AUC = 0.88$) and surpass all other models at the pixel level ($D_{KL} = 1.78$, $CC = 0.51$). Achieving good performance on DR(eye)VE shows that our model is not limited to the BDD-A dataset.

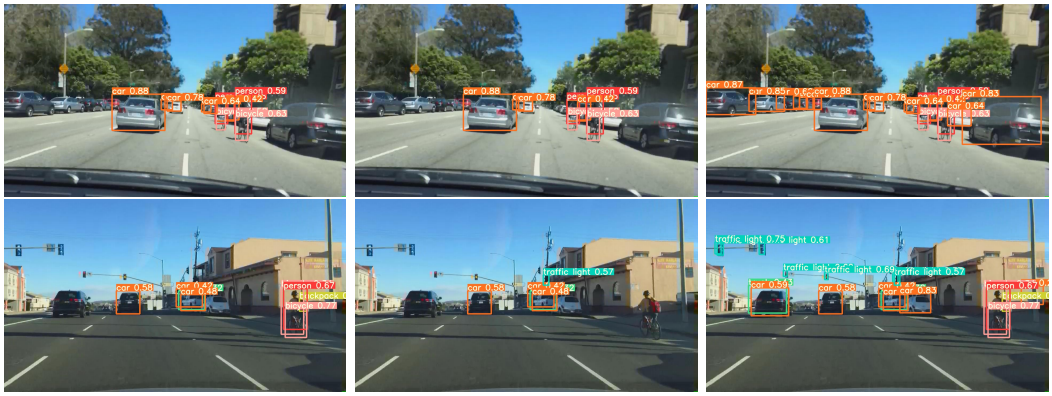


Figure 4.12: Comparison of our prediction, ground-truth in attention-based object detection and not using attention-based object detection on BDD-A test set. (Failed cases.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

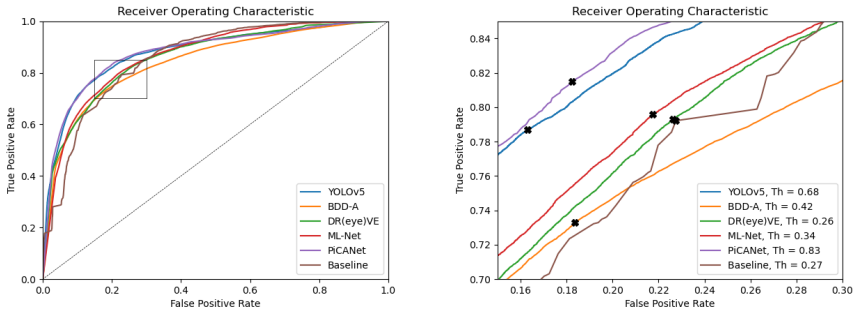


Figure 4.13: ROC curves and computed thresholds on the DR(eye)VE. On the right, the curves are zoomed in and the points that belong to the computed thresholds are marked.

Qualitative Results. Figure 4.14 presents two illustrative cases from the DR(eye)VE dataset, demonstrating the functionality of our attention-centric object prediction model. The top row demonstrates frames from a video where the driver navigates a leftward bend. Here, our model (on the left) successfully identifies a cyclist ahead of the car and another vehicle poised to enter traffic from the right. It omits distant cars from its focus area, aligning well with the actual ground-truth data (shown in the middle). The lower row features a scenario where the driver intends to make a left turn. In this instance, our model (on the left) anticipates the presence of vehicles and traffic signals straight ahead, while the ground-truth (in the middle) also acknowledges a car making a left turn. This scenario emphasizes the complexities of accurately predicting driver focus, particularly when influenced by varying driving objectives [194].

	Object-level					Pixel-level	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	*KL*	*CC*
Baseline	0.86	65.18	77.79	70.93	77.94	2.00	0.40
BDD-A [7] †	0.84	71.63	73.34	72.48	78.38	2.07	0.46
DR(eye)VE [8] ‡	0.86	68.90	79.39	73.77	78.09	2.79	0.47
ML-Net [9]†	0.87	69.74	79.73	74.40	78.71	2.17	0.45
PiCANet [10]†	0.88	73.90	81.48	77.50	81.64	2.36	0.41
Ours (YOLOv5)*	0.88	75.33	78.73	76.99	81.74	1.78	0.51

Table 4.13: Comparison with other gaze models on DR(eye)VE dataset. On object-level, all models are evaluated with detected objects of YOLOv5. Our models uses 16×16 grids. * indicates that the backbone is pretrained on COCO [16], † on ImageNet [17] and ‡ on UCF101 [18].

4.3.6 Discussion

Modelling with LSTM-Layer. In expanding our model for video-based forecasting, we incorporate an LSTM-layer (Long Short-Term Memory [195]) with a hidden state size of 256 ahead of the dense layer within the gaze prediction network. This network processes eight-frame video clips as input. Employing the same setup as outlined in the previous section (namely, 16×16 grids with a threshold Th of 0.5), we evaluate our modified architecture and obtain specific outcomes on the BDD-A dataset:

Object Detection: $AUC = 0.85$, Precision = 73.13%, Recall = 70.44%,
F_1 score = 71.76%, Accuracy = 77.83%

Saliency Prediction: $D_{KL} = 1.17$, $CC = 0.60$

The outcomes presented above are comparable to those of our model that lacks the LSTM-layer, with both registering an $AUC = 0.85$ and $CC = 0.60$. Variations in the sequence length, ranging from 2 to 16, do not significantly impact the model performance, as demonstrated in Table 4.14 when adding one LSTM layer with the hidden size 256 before the dense layer of our YOLOv5-based 16×16 grids model. All sequence lengths achieve very similar results. In a similar manner, [7] found that the inclusion of LSTM layers did not enhance performance in driver gaze prediction and instead led to central biases in the predictions.

Figure 4.15 presents a comparative analysis of two sets of predicted gaze maps: those utilizing an LSTM module (displayed in the middle) versus those without it (shown on the left), alongside the ground-truth maps (on the right). This LSTM module is composed of a single layer with a hidden size of 256 and processes input sequences of length 8. Observations indicate that incorporating the LSTM module improves the accuracy in predicting the central region of the gaze maps. However, this enhancement presents both benefits and drawbacks, resulting in an unchanged AUC value of 0.85.

To summarize, increasing the number of frames does not increase the information gain. A potential explanation for the observed bias is that using an LSTM layer ignores the spatial information, since the extracted features given to the LSTM layer are reshaped to vectors. In future research, we aim to examine the incorporation of modules that con-

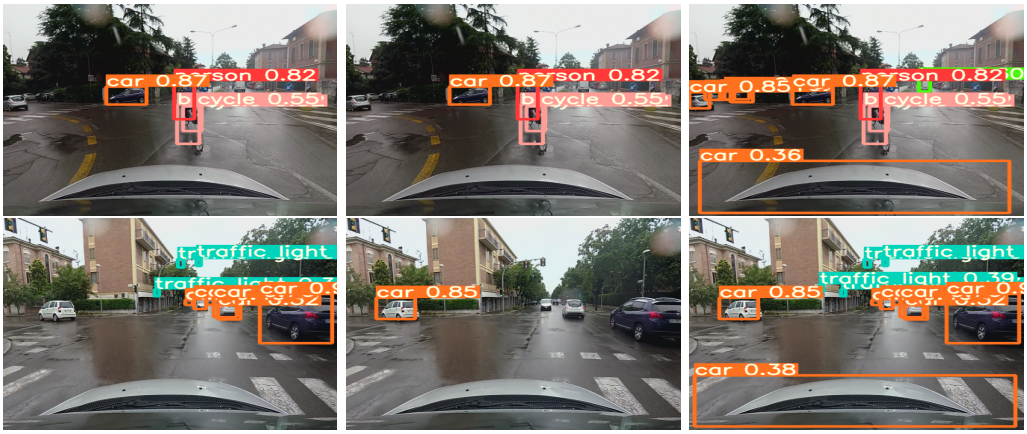


Figure 4.14: Comparison of our prediction, ground-truth in attention-based object detection and not using attention-based object detection on the DR(eye)VE test set ($Th = 0.4$ to better illustrate the wrongly predicted attention region in the failed case). (The second line is a failed case.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

	Object-level					**Pixel-level**	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	*KL*	*CC*
2	0.85	72.40	72.68	72.54	78.00	1.16	0.60
4	0.85	72.58	73.02	72.80	78.18	1.16	0.60
6	0.85	72.52	73.04	72.78	78.16	1.18	0.60
8	0.85	73.13	70.44	71.76	77.83	1.17	0.60
16	0.85	71.84	73.39	72.61	77.86	1.18	0.60

Table 4.14: Comparison of different input sequence lengths when using one LSTM layer. Our model uses the 16×16 grids. For all metrics except D_{KL}, a higher value indicates better performance. ($Th = 0.5$)

tain temporal aspects, like the convolutional LSTM (convLSTM) [36]. The convLSTM is proficient in assimilating the temporal data from each spatial area and forecasting new attributes for these areas, drawing on historical motion data within them. For instance, studies [7, 179] have shown that convLSTM effectively gathers spatial-temporal data, which is crucial for predicting driver attention and actions. Additionally, we are considering the application of 3D CNN to extract spatial-temporal features. An example of this is [8], which utilizes 3D convolutional layers to process a series of frames for anticipating the focus of a driver.

Limitations and future work. A current constraint in ongoing projects is the inherent central bias present in all existing model predictions. As noted, this bias originates from the nature of the ground-truth data. Given that human drivers predominantly focus on the center of the street, this results in significantly skewed data: for instance,

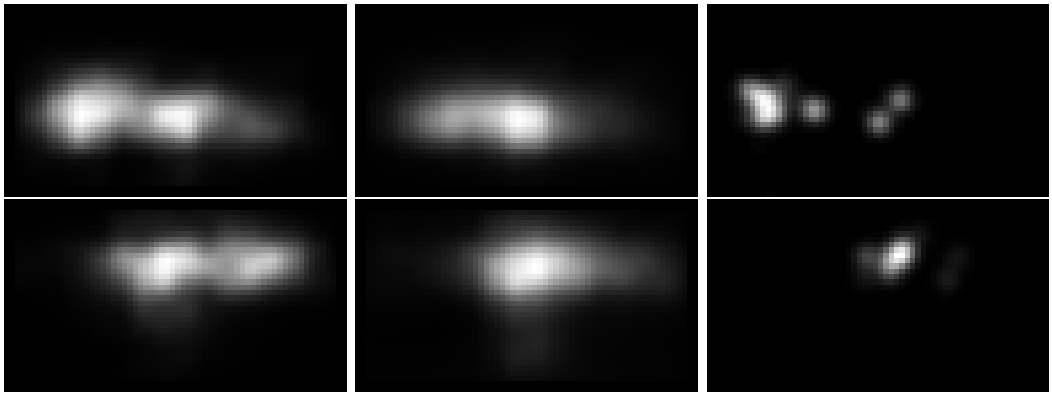


Figure 4.15: Comparison of predicted gaze maps without and with LSTM and ground-truth **Left**: Our prediction without LSTM; **Middle**: Our prediction with LSTM; **Right**: Ground-truth.

74.2% of all objects of interest in BDD-A are located within the central bias zone, as illustrated in the baseline of Figure 4.10. Such central bias is a reflection of typical human behavior and is further accentuated in the saliency models developed by Kümmerer et al. [196, 35]. While our model does recognize objects in the peripheral areas of the scene, as our qualitative examples demonstrate, there is a noticeable bias for the center. The model F_1 score is 81.7% in the central region, contrasting with a mere 34.8% F_1 score in the peripheral areas. PiCANet, outperforming all other models, shows better F_1 scores both outside (44.0%) and inside (82.7%) the central area. However, its performance is better in the central region. Our goal is to enhance the model's predictive accuracy in the peripheral areas while maintaining its strong performance in the central region. In autonomous driving scenarios, it is also crucial to evaluate the model's generalization capabilities using varied datasets, not just limited to gaze map data. Considering drivers use peripheral vision and do not always concentrate on every important object around them, incorporating datasets that also emphasize objects based on semantic relevance (e.g., [15]) could broaden the model's utility for identifying task-specific objects.

In the conducted experiments, models are developed using saliency maps created from the gaze of drivers. These maps highlight salient elements corresponding to areas where objects of interest for the task are likely to be found, indicating top-down features [197]. These elements are identified through visual data from camera imagery. The aspect of driving tasks could be further enriched by integrating additional input data, as human mechanisms for choosing top-down features necessitate a comprehensive grasp of the task beyond mere visual cues. Specifically, factors like road conditions (external) and the driver's own goals (internal), such as their destination, influence the driver's focus and gaze patterns, along with traffic information. Regrettably, the existing dataset for training our model lacks this extra input. Future research aims to include GPS and Lidar sensor data, offering deeper task-related insights for more accurately predicting where drivers direct their attention.

4.4 Conclusion

In this chapter, we introduce two novel models that address different types of driver intentions. The first study in Section 4.2 targets driver maneuver prediction. In this study, we proposed a method that integrates both inside and outside cabin motion features to predict driver maneuver intentions. We employed a ConvLSTM-based auto-encoder to capture the motion of traffic videos. The motion details captured were then processed through a classifier fusing inside and outside movement features. Our approach was trained in an end-to-end manner, avoiding reliance on manually encoded or hand-crafted features. The results demonstrated that our dual-input system significantly outperformed existing methods that utilize a single data source. Furthermore, our validation confirmed that videos from both inside and outside the cabin provide important and complementary insights.

Section 4.3 presents an innovative approach for identifying objects that human intends to interact with within driving contexts. The approach involved predicting saliency maps that indicate driver attention and identifying objects within these maps. Both saliency map prediction and object detection utilized the same underlying structure (feature encoder), with the saliency maps being generated in a grid format. This method ensured considerable computational efficiency. Extensive testing on two datasets related to driver attention, namely BDD-A and DR(eye)VE, demonstrated that our approach surpassed other baselines in attention saliency map prediction and object detection with reduced computational demand.

Part III

Enhancing Human Comprehension

The previous two parts of this dissertation delve into AI models that take into account human attention and intentions, aligning with the human factors of *perception* and *response* procedures. This section shifts focus to another human factor: *reasoning*. In designing AI models, the integration of XAI technologies is commonly employed to reflect the reasoning process.

This part is adapted from two papers that were published in ICML 2022 [198] and TPAMI 2023 [41], respectively, and a third work that will be published in AAAI 2024 [199]:

- Rong, Y., Leemann, T., Borisov, V., Kaneci, G., & Kasneci, E. (2022)

  A Consistent and Efficient Evaluation Strategy for Attribution Methods

  In *Proceedings of the 39th International Conference on Machine Learning (ICML)*

- Rong, Y., Leemann, T., Nguyen, T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2023)

  Towards Human-centered Explainable AI: User Studies for Model Explanations

  IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)

- Rong, Y., Qian, P., Unhelkar, V., & Kasneci, E. (2023)

  I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise

  Pre-print. (To appear at the 38th Annual AAAI Conference on Artificial Intelligence (AAAI)).

In the following section, I will summarize the motivation, principal methodology, main findings, and my contributions to each of the three papers.

A Consistent and Efficient Evaluation Strategy for Attribution Methods

Motivation. The decision-making mechanisms of current AI systems are complex and thus opaque, leading to their characterization as "black boxes." XAI is the key to unboxing these models, as discussed in Section 1.2.3. This work (in Chapter 5) studies the popularly used automated evaluation metrics, as they often yield *inconsistent* results, thereby complicating the benchmarking process for model explanations. Therefore, it is important to thoroughly analyze the biases existing in the current benchmarking and to mitigate the bias to provide a fair evaluation.

Principle methodology. This work studies the bias in mechanisms underlying evaluation strategies based on perturbation by conducting a rigorous *information-theoretic* analysis, which is explained in Section 5.2. It formally reveals that results can be significantly confounded. Concretely, the bias term originates from a masking operation in the evaluation process. Beyond discovering the bias term, a mitigation solution is proposed and thus contributes to a novel evaluation strategy. The mitigation solution

utilizes linear imputation in the masking operation and alleviates information leakage through masking.

Main findings. Section 5.2 presents both theoretical and experimental analyses of the bias term, referred to as Class Information Leakage. The mitigation solution named Noisy Linear Imputation is able to significantly decrease the inconsistency in the evaluation results, which occurs due to a different feature removal order. For instance, the common imputation method of using a "fixed value" results in a Spearman Rank correlation of -0.01 among different orders. Our proposed solution significantly improves the correlation score to 0.61. Moreover, the mitigation is further generalized to a novel evaluation strategy, which can efficiently evaluate model explanations. Compared to the previous evaluation strategies requiring retraining, our strategy saves 99% of the computational costs.

My contributions. I led the development and research efforts for this work. My role contains identifying research gaps within the existing literature and devising innovative solutions to address these gaps. Specifically, I introduced the application of mutual information theory in our ICML 2022 work [198]. My responsibilities extended to implementing the methodologies, which included coding the frameworks and establishing baselines, as well as conducting experiments to compare our methods against other state-of-the-art approaches. Moreover, I also undertook the tasks of manuscript writing and creation of illustrations.

Towards Human-centered Explainable AI: User Studies for Model Explanations

Motivation. This work highlights the role of XAI technologies in enhancing human comprehension of opaque AI models. While the aim of model explanations is to facilitate human understanding, there are not that many XAI studies that adequately consider *human subjects* in their studies [44]. As user studies are gaining attention in the XAI research community, there is no consensus in conducting these user studies. Therefore, in this work, our goal is to provide practical guidelines to benefit practitioners and researchers in XAI.

Principle methodology. This work (explained in Section 5.3) performs an extensive literature review to provide the guidelines for user study design in XAI. The analysis method first classifies research papers based on the measured factors, such as user understanding or user trust. Then, papers within each category are concisely summarized to extract key insights, which are instrumental in developing the guidelines.

Main findings. According to 97 studied papers, user study design is generally divided into three main steps: before, during, and after user study. At every stage, a *summary card* presents considerations for researchers to be mindful of and aim to avoid common

pitfalls as well as guarantee potentially unbiased assessment outcomes. Details of the guidelines can be found in Figure 5.11.

My contributions. I led the development for this work appeared in TPAMI 2023 [41]. My role included identifying research gaps within the existing literature and conducted the literature review. Concretely, I proposed a data-driven literature analysis for this work, and I coded the framework for the analysis. Beyond leading the literature review, I was responsible for the manuscript writing, and the creation of illustrative plots to effectively communicate our findings.

I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise

Motivation. This work addresses the importance of considering a human factor – reasoning – in designing model explanations. More specifically, current SOTA model explanations are *one-size-fits-all* solutions and some explanations may not be useful for users. Therefore, this work is motivated to consider human reasoning for making decisions and to include it in generating model explanations. In this manner, the model explanation is tailored to the *individual* user's needs, which can better help each user in understanding the AI model.

Principle methodology. To consider the human factor in designing model explanations, this work (detailed in Chapter 6) proposes a framework named I-CEE that provides image classification explanations tailored to *user expertise*, i.e., I-CEE models the informativeness of the example images to depend on user expertise. Concretely, there are two main components in I-CEE: User Expertise Estimation and Selection Strategy. As humans often use "concept-based thinking" in classifying images, the User Expertise Estimation is designed based on a concept discovery algorithm and trained following the annotator model in active learning. In this manner, a user's expertise in applying task-relevant concepts can be simulated. Selection Strategy is informed by the educational psychology concept "Hypercorrection Effect", aiming at selecting the most informative examples based on the estimated user expertise.

Main findings. Evaluation results in Chapter 6 show that the proposed framework I-CEE is able to achieve the SOTA performance in *simulatability*, i.e., users' ability to predict the model's decisions, indicating the power of generated explanations in improving human comprehension. Concretely, evaluations with the simulated users show that I-CEE outperforms other baseline models, especially on two realistic datasets, as illustrated in Figure 6.5(b,c). A user study with $N = 100$ human subjects is conducted to address the effectiveness of I-CEE in practice. For instance, I-CEE significantly improves the simulatability score by 11.5% with $p = 0.007$ compared to the baseline. These results highlight the importance of considering personalization via user expertise in XAI.

My contributions. I am the leading author of the paper appeared at AAAI 2024 [199]. I discovered the research gaps and initialized the innovative solutions to address these gaps. Moreover, I developed the simulated user diagram for this work and I implemented the methodologies in scripts, which included establishing baselines, as well as conducting experiments to compare our methods against other state-of-the-art approaches. Last but not least, I wrote the majority of the manuscript and created illustrations and figures to facilitate efficient scientific communication.

5 Evaluating Model Explanations

5.1 Introduction

XAI focuses on enhancing the interpretability and transparency of AI systems. This approach emphasizes the significance of human stakeholders in AI development, as underscored in key studies [200, 201]. Despite the availability of numerous model explanations, the challenge of evaluating their quality transparently remains unresolved, attracting considerable research attention in recent years. A well-recognized classification of XAI evaluation strategies categorizes them into three distinct types: functionally-grounded, application-grounded, and human-grounded evaluation [122]. Functionally-grounded evaluations, which can be conducted in an automated way and do not necessitate human involvement, contrast with the latter two types that involve human participants and are more resource-intensive to implement.

Many functionally-grounded metrics (automatic metrics) have been developed to assess XAI algorithms, as reviewed in [44]. However, a notable challenge is the difficulty in comparing these diverse automatic evaluation measures, as there is no ground-truth in explanation in real-world applications. Common methods for evaluating and comparing various attribution techniques typically involve an ablation strategy. This involves altering the input features, such as image pixels, which are identified as either the most or least significant. In this context, altering pixels deemed highly significant should lead to a reduction in the accuracy of predictions, whereas modifying those considered insignificant should not significantly impact the predicted outcomes. These techniques are designed to gauge the *fidelity* of explanations [42], that is, the extent to which the explanations accurately represent the predictions made by the underlying model. Fidelity assessed using a single data instance is referred to as local fidelity, whereas global fidelity is evaluated across the entire dataset [42].

The sensitivity of evaluation outcomes to parameters such as the perturbation function and sequence is significant. The choice of sequence, either *most relevant pixels first* or *least relevant pixels first*, can lead to starkly divergent results in removal strategies. Local attribution methods, for example, seem to perform well in one order but may perform poorly in the other[42, 43, 13]. This inconsistency presents a challenge for researchers attempting to objectively compare different attribution methods, with the sources of such inconsistencies remaining unclear. Additionally, some methods necessitate a retraining step for the global fidelity assessment, a process often deemed impractically costly [13, 42]. Section 5.2 first provides a thorough analysis of the bias existing in the current evaluation strategy and then proposes a solution to overcome these problems. These two drawbacks and our improvements are illustrated in Figure 5.1.

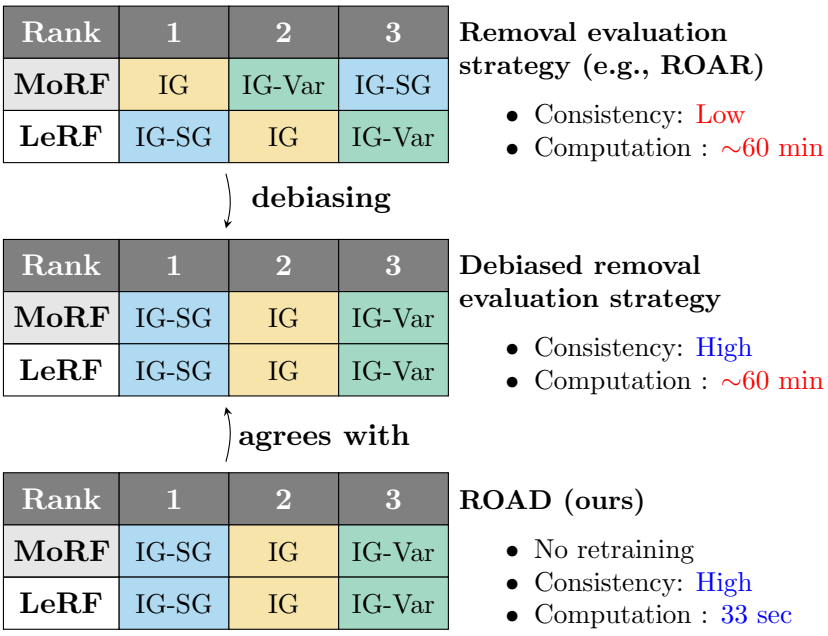


Figure 5.1: Comparison between previous removal and retraining evaluation strategies (**Top**) and ours (**Bottom**). Previously, rankings of different attribution methods, Integrated Gradients (IG) [11] and its two variants SmoothGrad (IG-SG) [12], SmoothGrad^2 (IG-SQ) [13], are highly inconsistent with respect to hyperparameters such as the removal orders Most Relevant First (MoRF) and Least Relevant First (LeRF). Our ROAD strategy achieves a consistent ranking using only 1% of the previously required resources.

One limitation of automated metrics is their lack of certainty in accurately reflecting human preferences [46, 47]. Therefore, in the context of XAI, particularly when transitioning to real-world applications, conducting user studies becomes essential to validate broader assumptions about the effectiveness of explanations [48]. Regrettably, a mere fraction (approximately 20%) of XAI evaluation studies incorporate human participants [44]. While there are initiatives to create taxonomies and define or understand the various aspects of human-centric evaluations [49, 50, 51], a systematic discussion on the recent user studies and their outcomes is still pending. Additionally, Yang et al. [52] observe that the field of XAI is developing in isolation, with distinct treatments across various disciplines (such as machine learning and HCI). Therefore, providing effective guidance for XAI user study design is critical to ensuring that both algorithm and application developers in XAI adequately address the genuine requirements of users. The objective of Section 5.3 is to bridge this research gap in current XAI user study methodology by offering practical guidelines for user studies, derived from a thorough and methodical review of the literature.

C	Class label random variable
I	Mutual information
$\mathcal{I}$	Imputation operator
$\boldsymbol{M}$	Binary mask in $\{0,1\}^d$
$\mathcal{M}$	Mask selection operator (takes out relevant features)
$\boldsymbol{x}$	Input features in $\mathbb{R}^d$
$\boldsymbol{x}_l$	Low importance features only in $\mathbb{R}^{d-k}$
$\boldsymbol{x}'_l$	Imputed low importance features in $\mathbb{R}^d$

Table 5.1: Notation used in this section.

5.2 Bias in Automatic Evaluation

The objective of this section is to address the shortcomings of the existing evaluation strategy thereby enhancing the consistency of the evaluation. We introduce an innovative strategy to counteract the biases arising from confounders, which contribute to inconsistencies. Moreover, our findings indicate that in a debiased environment, omitting retraining does not notably alter the outcomes. Figure 5.1 illustrates our robust evaluation strategy ROAD (RemOve And Debias) compared to the previous work ROAR (ReMOve And Retrain) [13], which are not limited by computational resource constraints, is important for the community.

5.2.1 Retraining Evaluation Strategies

Our approach considers a pixel removal strategy, where pixels are successively replaced by imputed values. In line with existing research [42, 125], we explore two different orders of removal: **MoRF** (Most Relevant First) and **LeRF** (Least Relevant First). The former starts with the most important pixels, while the latter begins with the less significant ones. We introduce a detailed definition of MoRF with retraining, known as the ROAR benchmark [13], which will be the focus of our study. In this section, we exclusively discuss the MoRF for our analysis. Nevertheless, a similar examination of LeRF can be conducted without much additional effort.

For simplifying our derivations, we outline the process through a sequence of independently analyzable operations. A classifier $f : \mathbb{R}^d \to \{1, \dots, c\}$ maps inputs $\boldsymbol{x} \in \mathbb{R}^d$ to labels $C \in \{1, \dots, c\}$, where c is the number of classes. The purpose of a feature attribution explanation for the prediction is to allocate an importance value to each input dimension. In the MoRF scenario, the features are ranked based on their importance in descending order. Following this, the top k most significant features for each instance are identified for removal, where k is incrementally increased from 0 to d during the benchmark. However, for the moment we consider only one fixed value of k. Therefore, we can model the explanation $\boldsymbol{e}_k$ as a choice of features via a binary mask $\boldsymbol{M} = \boldsymbol{e}_k(f, \boldsymbol{x}) \in \{0,1\}^d$, where the value is set to one if the feature is in the top-k and zero otherwise. In addition, we consider $\mathcal{M}_l : \{0,1\}^d \times \mathbb{R}^d \to \mathbb{R}^{d-k}$ as the operation for

selecting the least important dimensions as indicated by the mask, and $\boldsymbol{x}_l = \mathcal{M}_l(\boldsymbol{M}, \boldsymbol{x})$ as the vector comprising only the unselected features. We assume that in $\boldsymbol{x}_l$, the features maintain their original input order, meaning they are arranged ascendingly by their initial input indices. This setup enables us to examine the information flow in the feature *mask* $\boldsymbol{M}$ and the feature *values* $\boldsymbol{x}_l$ separately.

The ROAR strategy evaluates the performance of a retrained classifier f' on perturbed samples $\boldsymbol{x}'_l \coloneqq \mathcal{I}_l(\boldsymbol{M}, \boldsymbol{x_l})$, where $\mathcal{I}_l : \{0,1\}^d \times \mathbb{R}^{d-k} \to \mathbb{R}^d$ acts as an imputation operator that redistributes all inputs in the vector $\boldsymbol{x}_l$ to their original positions and sets the rest with a certain value. Specifically, in the context of zero imputation $\boldsymbol{x}'_l = \mathcal{I}_l(\boldsymbol{M}, \mathcal{M}_l(\boldsymbol{M}, \boldsymbol{x})) = (1 - \boldsymbol{M}) \odot \boldsymbol{x}$, meaning that top-$k$ features are discarded. For a better evaluation result, a rapid drop in accuracy with an increase in k is desirable, as this indicates the effective removal of the most important features.

Analysis with Mutual Information. In retraining-based pixel removal strategies, we use Mutual Information (MI) as an indicator for potential accuracy, since a higher MI typically correlates with increased accuracy. [45] (Appendix A.2) provides proof of the relation between classification accuracy and mutual information. During MoRF retraining, the metric $I(\boldsymbol{x}'_l; C)$ becomes crucial as it measures the remaining information in less significant features, thereby influencing the achievable accuracy, which is the focus of our evaluation. A reduction in mutual information $I(\boldsymbol{x}'_l; C)$ often leads to a significant decrease in accuracy, yielding impressive MoRF benchmark results:

$$\downarrow I(\boldsymbol{x}'_l; C) \quad \Rightarrow \quad \uparrow \text{MoRF benchmark.}$$

Hence, in the MoRF framework, lower mutual information between $\boldsymbol{x}'_l$ and C is preferable. Conversely, in LeRF, a higher accuracy, hence higher $I(\boldsymbol{x}'_l; C)$, is desired.

5.2.2 Bias Analysis

Bias: Class Information Leakage. This section aims to demonstrate that leaking class information solely through the mask's shape is readily achievable, and this can significantly alter the evaluation score. To do this, we first distinguish the impact of the mask from the feature values. Our analysis is based on the multi-information $I(C; \boldsymbol{x}'_l; \boldsymbol{M})$, as defined by [202]:

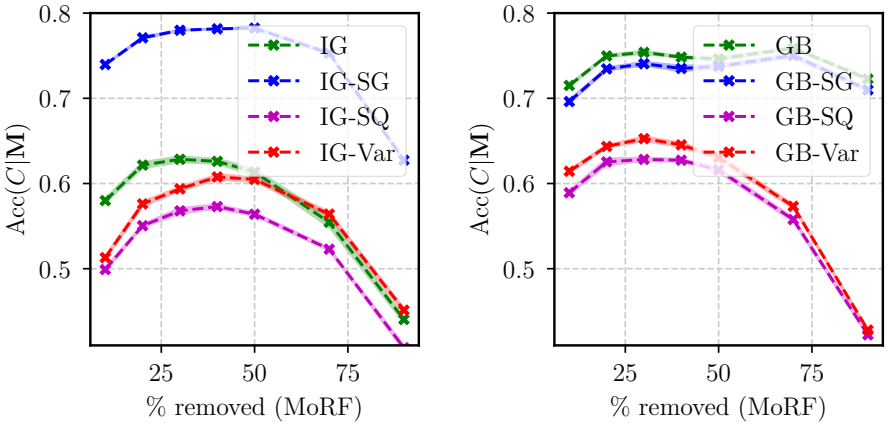
$$I(C; \boldsymbol{x}'_l; \boldsymbol{M}) = I(C; \boldsymbol{x}'_l | \boldsymbol{M}) - I(C; \boldsymbol{x}'_l) \tag{5.1}$$

$$I(C; \boldsymbol{x}'_l; \boldsymbol{M}) = I(C; \boldsymbol{M} | \boldsymbol{x}'_l) - I(C; \boldsymbol{M}). \tag{5.2}$$

Setting eq. (5.1) and eq. (5.2) equal, we arrive at the identity:

$$\underbrace{I(\boldsymbol{x}'_l; C)}_{\text{Eval. Outcome}} = \underbrace{I(C; \boldsymbol{x}'_l | \boldsymbol{M})}_{\text{Feature Info.}} + \underbrace{I(C; \boldsymbol{M})}_{\text{Mask Info.}} - \underbrace{I(C; \boldsymbol{M} | \boldsymbol{x}'_l)}_{\text{Mitigator}}. \tag{5.3}$$

The first term denoted as "Feature Information" represents the information about the class embedded within the features (excluding the mask), which we aim to measure. Conversely, the "Mask Information" term illustrates the significant influence class-specific



(a) Integrated Gradients(IG) (b) Guided Backpropagation(GB)

Figure 5.2: Accuracy of a trained classifier only using the binary masks $\boldsymbol{M}$ without feature values as input on the CIFAR-10 data set. Binary masks $\boldsymbol{M}$ were computed for different variants of IG and GB. Only the masks contain enough information to reach an accuracy of almost up to 80 % (compared to 85 % with full images) highlighting that the feature values do not play an important role in the evaluation. This underlines the necessity to compensate for this confounder.

information within the mask can exert on the outcome. However, this effect can be offset by the "Mitigator" term. The Mitigator becomes entirely indiscernible when the mask can be flawlessly deduced from the imputed image $\boldsymbol{x}'_l$. This leads to an uncompensated outcome known as Class Information Leakage.

Extent of Class Information Leakage. To validate the bias term in practice, we conducted experiments on the CIFAR-10 dataset [203]. We employed the same attribution techniques as described in [13]: Integrated Gradients (IG) [11] and Guided Backpropagation (GB) [204] were used as foundational explanations. Additionally, we applied three ensemble approaches for each method: SmoothGrad (SG) [12], SmoothGrad^2 (SQ) [13], and VarGrad (Var) [14].

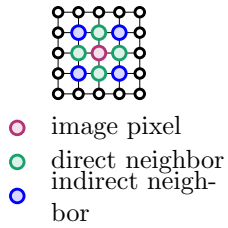
Our experimental findings indicate that when employing fixed value imputation using the global mean, the explanatory masks inadvertently disclose class information. This conclusion is reached through a two-step process: Firstly, we demonstrate that the Mask Information $I(C; \boldsymbol{M})$ exhibits an exceptionally high value. Secondly, we confirm that the Mitigator is small, suggesting that class information infiltrates the evaluation outcome via $I(C; \boldsymbol{M})$.

To evaluate the class-related information in the mask, we train a ResNet-18 [169] that exclusively uses binary masks $\boldsymbol{M}$ (excluding pixel values $\boldsymbol{x}_l$) for class prediction. As previously discussed, the accuracy of a classifier can serve as a substitute for measuring Mutual Information (MI), which is impractically costly for high-dimensional data. The

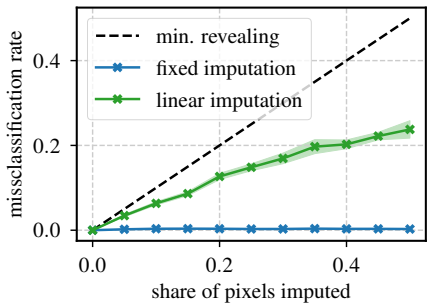
curves[1] are depicted in Figure 5.2. Remarkably, using only the mask leads to high-accuracy curves, peaking at nearly 80% for IG-SG, just slightly lower than the accuracy achieved with the full inputs. This suggests that the Mask Information $I(C; \boldsymbol{M})$ is nearly as substantial as the Evaluation Outcome $I(C; \boldsymbol{x}'_l)$.

To demonstrate that the Mitigator is nearly non-existent, leading to leakage of class information, we test whether it is easy to infer M from $\boldsymbol{x}'_l$. In the case of fixed value imputation, the inverse is feasible: Assign a value of 0 to every pixel within the mask if its corresponding image pixel matches the filling value (which must be deduced from the distribution). For more robust validation, we utilize a three-layer convolutional network to predict whether each pixel is original or imputed. As illustrated by Figure 5.3(e) (blue curve), the error rate with fixed value imputation is nearly zero, indicating the network's high accuracy in identifying imputed pixels. Our analysis suggests that using fixed value imputation, the impact of the Mitigator is negligibly small.

(a) Original (b) Fixed Imp. (c) Noisy Lin. Imp.



(d) Graphical model used to derive our imputation



(e) Misclassification rate of the imputation predictor for different shares of randomly imputed pixels (CIFAR-10)

Figure 5.3: The considered imputation operators. When 50 % of the original image (a) are removed, they can either be imputed by a fixed value (b) or by our proposed Noisy Linear strategy (c,d). Training of an imputation predictor (e) shows that it is much harder to tell which pixels are original and which were imputed when using our proposed imputation model. This is closer to the optimal, minimally revealing imputation (black). Hence, by using imputed samples of this kind, Class Information Leakage is reduced.

[1] Standard Errors are denoted by shaded areas in all figures, but they are typically barely noticeable due to their small size.

5.2.3 Debiasing Evaluation Strategy

Reduction of Class Information Leakage. To mitigate the bias, we follow an intuitive approach: In cases where we cannot ensure the absence of class data within the mask itself, our objective is to prevent this class data from being transmitted into the imputed images. Thus, we ensure that the mask employed is not easily recognized from the imputed image. Our goal is to achieve $I(\boldsymbol{x}'_l; \boldsymbol{M}) = 0$, meaning the mask is unrelated to the imputed vector, enabling us to disentangle the influences. Unfortunately, this is not always feasible: If both are influenced by the class label, they must inevitably share some information about the class. Therefore, we demand $I(\boldsymbol{x}'_l; \boldsymbol{M}) \approx 0$, and $I(\boldsymbol{x}'_l; \boldsymbol{M}|C) = 0$. I.e., $I(\boldsymbol{x}'_l; \boldsymbol{M}) - I(\boldsymbol{x}'_l; \boldsymbol{M}|C) \approx 0$. As $I(C; \boldsymbol{M}) - I(C; \boldsymbol{M}|\boldsymbol{x}'_l) = I(\boldsymbol{x}'_l; \boldsymbol{M}) - I(\boldsymbol{x}'_l; \boldsymbol{M}|C)$, $I(C; \boldsymbol{M}) \approx I(C; \boldsymbol{M}|\boldsymbol{x}'_l)$, suggesting Mitigator successfully compensate the Mask Information term.

Debiasing with Noisy Linear Imputation. In an effort to mitigate Class Information Leakage, we propose an improved imputation operation, which does not reveal M from the imputed image, i.e., $I(\boldsymbol{x}'_l; \boldsymbol{M}) \approx 0$. Beyond this requirement, we encounter other challenges in practice: (1) The imputation approach must be exceptionally efficient, considering the necessity to apply it to every image in the dataset. (2) It is desirable to have few hyper-parameters to avoid introducing other confounding factors.

We propose a novel approach named *Noisy Linear Imputation*, achieving objectives mentioned above. Our method tackles key challenges inherent in current approaches. Essentially, our aim is to develop subtler imputations that are less detectable and yield a reduced $I(\boldsymbol{x}'_l; \boldsymbol{M})$. Toward this goal, we posit that each pixel's value can be effectively estimated using a weighted average of its neighboring pixels (see Figure 5.3(d)), given the strong correlation among image pixels[2]:

$$\begin{aligned}\boldsymbol{x}_{i,j} = {} & w_d \left(\boldsymbol{x}_{i,j+1} + \boldsymbol{x}_{i,j-1} + \boldsymbol{x}_{i+1,j} + \boldsymbol{x}_{i-1,j}\right) \\ & + w_i \left(\boldsymbol{x}_{i+1,j+1} + \boldsymbol{x}_{i-1,j+1} + \boldsymbol{x}_{i+1,j-1} + \boldsymbol{x}_{i-1,j-1}\right),\end{aligned}$$

where w_d, w_i represent constant coefficients for direct and indirect diagonal neighbors, respectively, an equation system emerges from formulating a single equation for each removed pixel. We incorporate the values of known pixels directly into this system, treating only the removed pixels as unknown variables. This interconnects the equations when neighboring pixels are removed, preventing them from being solved independently. Despite this, the system remains sparse and is thus solvable efficiently, even with a substantial number of missing pixels. In determining the weights for the linear interpolation's neighbor weights, we take cues from the graph's structure (refer to Figure 5.3(d)): Indirect neighbors are twice as far from the original node in the graph compared to direct neighbors. Therefore, direct neighbors are assigned a weight double that of diagonal neighbors. To maintain a weighted interpolation where weights sum up to 1, we set $w_d{=}\frac{1}{6}$ and $w_i{=}\frac{1}{12}$. Additionally, a slight random noise ($\sigma = 0.1$) is added to the solution, preventing the model from learning the linear dependency.

[2]Specifically, the correlation coefficients are ρ=0.89 for direct neighbors and ρ=0.82 for indirect neighbors in CIFAR-10

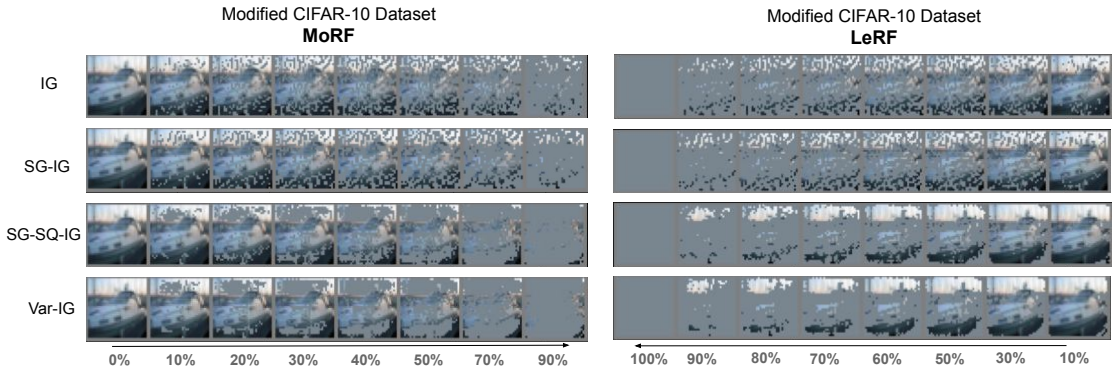


Figure 5.4: Illustration of modified data set in MoRF/LeRF and fixed value imputation settings. **Left**: Modifications in the MoRF framework. **Right**: Modifications in the LeRF framework. **Top to Bottom**: Modifications using Integrated Gradient (IG) [11] and three ensemble variants of IG: SmoothGrad (SG-IG) [12], SmoothGrad^2 (SG-SQ-IG) [13], and VarGrad (Var-IG) [14]. The percentage of pixels that are removed or kept is given at the bottom.

Figure 5.3 (top) illustrates an exemplar of an imputed sample. The imputed version shown in Figure 5.3(c) makes inferring the mask much more challenging than the version with fixed-value imputation in Figure 5.3(b). For validation, we train the imputation predictor anew and present the outcomes in Figure 5.3(e). Our method is verified to approximate the ideal, Minimally Revealing Imputation, more closely. There are more advanced imputation methods available, such as those based on Generative Adversarial Networks (GANs), including the Generative Adversarial Imputation Nets (GAIN) introduced by [205]. Nonetheless, our approach marks a significant advancement in efficiency and effectiveness, especially as it circumvents the necessity of training a GAN model. For thoroughness, we also conduct further experiments with GAN-based imputation, detailed in Section 5.2.5.

5.2.4 Experiments

Having established the effectiveness of our Noisy Linear Imputation, this section demonstrates additional practical benefits. This approach is termed ROAD (RemOve And Debias). The experiments in this section were performed using CIFAR-10 and the eight specified attribution techniques. Moreover, the Food-101 dataset [206], comprising high-resolution images, was employed to test the generalizability of our method, which can be found in Appendix D.1. Over 1000 models were trained from the ground up on imputed data utilizing the outlined strategies, explanations, and removal percentages.

Implementation details. A vanilla ResNet-18 model [169] is utilized for training on the CIFAR-10 dataset, and various explanations are derived from this model. This model undergoes training with an initial learning rate of 0.01, employing the SGD optimizer [207]. The learning rate is reduced by 0.1 following 25 epochs, and the training continues

Retrain MoRF vs. LeRF		No-Retrain MoRF vs. LeRF	
fixed	lin	fixed	lin
-0.01$_{\pm 0.01}$	**0.61**$_{\pm 0.01}$	0.01$_{\pm 0.00}$	**0.58**$_{\pm 0.01}$

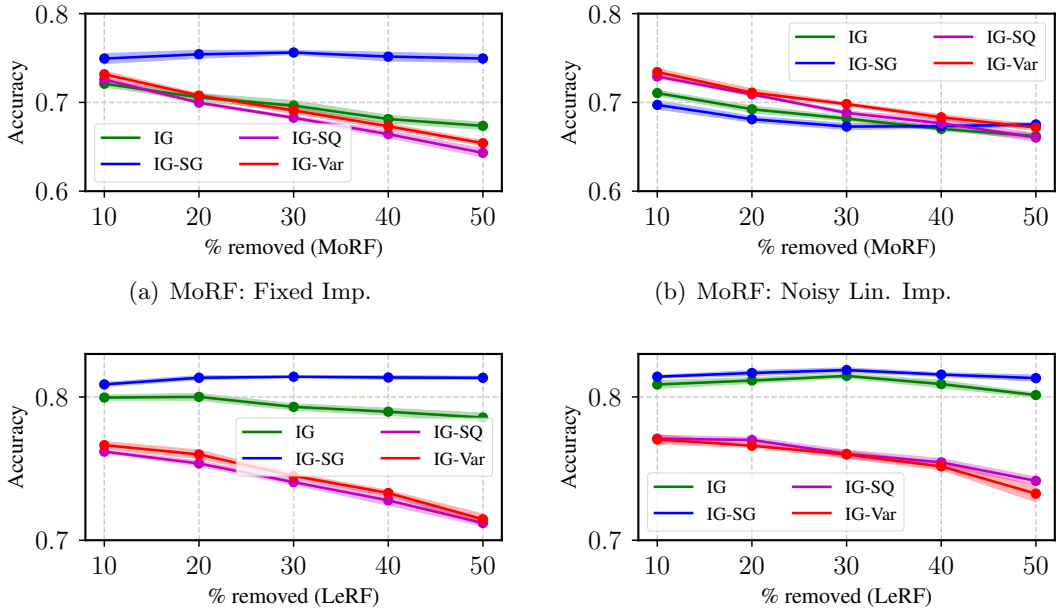
Table 5.2: Spearman rank correlation between evaluation strategies on **CIFAR-10**. There is almost no agreement between MoRF and LeRF when using fixed imputation (as in previous works). When using our imputation ("lin"), consistency across MoRF and LeRF orders increases drastically.

for a total of 40 epochs on a single GPU. This training results in the model attaining a test accuracy of 84.5% (which aligns with the results reported in [42]). For attribution, we adhere to the settings described in [13]. The baseline explanations include Integrated Gradient (IG) [11] and Guided Backprop (GB) [204]. We further incorporate three ensemble methods: SmoothGrad (SG) [12], SmoothGrad^2 (SG-SQ) [13], and VarGrad (Var) [14]. Each explanation technique is applied to the dataset, modifying it with pixel fractions $\eta = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9]$. Figure 5.4 displays the altered images, showcasing four distinct explanations from the GB family in both MoRF and LeRF sequences, with a fixed mean value imputation approach.

Consistency under Removal Orders. For evaluation methodologies less susceptible to hyperparameter configurations and conducive to stable ranking, our focus is on the uniformity of evaluation outcomes across varied deletion sequences such as MoRF and LeRF. Illustrated in Figure 5.5 are the curves derived from the "Retrain" approach. For enhanced clarity, only a subset of four curves is presented, specifically those representing attribution methods using IG with retraining, where up to 50% of pixels are eliminated. The comprehensive set of curves for IG and its variants, along with GB and its variants, is included in Appendix D.2.

The outcomes applying the standard fixed value imputation are depicted in Figure 5.5(a) and Figure 5.5(c), while the Noisy Linear Imputation results are shown in Figure 5.5(b) and Figure 5.5(d). In the context of MoRF, an initial steep decline signifies a more effective attribution method, whereas a gradual decline is preferable in LeRF. Therefore, under fixed imputation, MoRF's variants are IG, IG-Var, IG-SQ, IG-SG, and for LeRF, it is IG-SG, IG, IG-SQ, IG-Var. For example, IG-SG ranks lowest in MoRF but highest in LeRF. The use of Noisy Linear Imputation eliminates this disparity. In MoRF, the ranking is IG-SG, IG, IG-SQ, IG-Var, identical to LeRF.

We perform a quantitative analysis of the uniformity across all eight attribution techniques, both with and without retraining. Specifically, we rank our explanation methods (1=best, 8=worst) based on the proportion of perturbed pixels. Subsequently, we determine the Spearman Rank correlation among different evaluation methods. As illustrated in Table 5.2, the correlation score for fixed value imputation is -0.01 with retraining and 0.01 without it, suggesting a lack of consistency in the rankings. However, with the



(a) MoRF: Fixed Imp.

(b) MoRF: Noisy Lin. Imp.

(c) LeRF: Fixed Imp.

(d) LeRF: Noisy Lin. Imp.

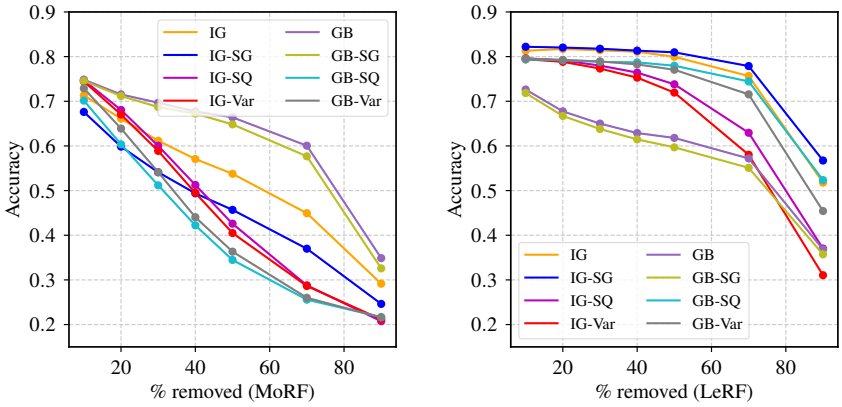
Figure 5.5: Consistency comparison using fixed value vs. Noisy Linear Imputation. The higher accuracy is better in LeRF, while the lower is better in MoRF. Comparing (a) and (c), fixed value imputation gives different rankings in MoRF and LeRF orders: IG-SG is the best in LeRF but the worst in MoRF. Comparing (b) and (d), Noisy Linear Imputation changes the outcome considerably and yields a consistent ranking in MoRF and LeRF.

introduction of our Noisy Linear Imputation, there is a significant shift: The correlation scores rise to 0.61 and 0.58 with and without retraining, respectively. This suggests that information leakage might play a significant role in the observed inconsistencies.

Efficiency. Employing Noisy Linear Imputation effectively narrows the gap observed in evaluations conducted with and without retraining. This improvement is mainly due to the decrease in distribution shift, thanks to the utilization of an approach akin to *Minimally Revealing Imputation*. If the imputation of all pixels were ideal, the images produced would not deviate from the expected distribution. Our focus is on comparing different attribution methods. Hence, we calculate the Spearman correlation for rankings derived with and without retraining, as presented in Table 5.3. The consistency in order between the "Retrain" and "No-Retrain" scenarios, both employing Noisy Linear Imputation, is evident. The rank correlation is 0.84 for MoRF and 0.94 for LeRF. This uniformity in ranking leads to the insight that the disparity between "No-Retrain" and "Retrain" scenarios is minimal when using Noisy Linear Imputation. Consequently, we infer that omitting the retraining phase does not significantly alter the outcomes. For a complete evaluation result, please refer to Appendix D.2.

MoRF		LeRF	
Retain vs. No-Retr.		Retain vs. No-Retr.	
fixed	lin	fixed	lin
$0.15_{\pm 0.01}$	$\mathbf{0.84}_{\pm 0.01}$	$0.09_{\pm 0.01}$	$\mathbf{0.94}_{\pm 0.01}$

Table 5.3: Spearman rank correlation between evaluation with and without retraining. Our Noisy Linear Imputation ("lin") also results only in marginal differences between "Retrain" and "No-Retrain". We conclude that the retraining step is no longer necessary.



(a) MoRF (b) LeRF

Figure 5.6: Evaluation results in MoRF (a) and LeRF (b) using our ROAD framework.

5.2.5 Discussion: GAN Imputation

We use Generative Adversarial Imputation Nets (GAIN) proposed by Yoon et al. [205] as an imputation operator. We first trained a GAIN model on the CIFAR-10 dataset. During this process, we conduct a hyperparameter tuning specifically for the GAIN model, maintaining the default parameters as suggested in [208]. We concentrate on optimizing two parameters: `alpha` (denoted as α), a factor influencing the reconstruction loss of non-imputed pixels in the GAN, and `hint_rate` (abbreviated as hr), which aids the Discriminator by providing hints to balance task difficulty. The training spans 100 epochs, leading to stabilized Mean Squared Errors (MSEs) and Frechet Inception Distances (FIDs). We employ MSE against the original pixels as a metric to evaluate the generative capabilities of the model. While Kachuee et al. note the effectiveness of lower values for both parameters, they do not specify exact figures. We broaden the range for α up to 100 and conduct a thorough search. The performance outcomes of the GAIN models on CIFAR-10 are detailed in Table 5.5. For our experiments, we select the most effective setup with $\alpha = 100$ and $hr = 0.01$. In Figure 5.7, the imputation outcomes for a CIFAR-10 image (a) using three distinct methods are presented. The GAN imputation

Strategy	Retrain		No-Retrain	
	fixed†	lin	fixed	lin⋆
Time	3903±117 s	4686±2 s	18.0±0.1 s	33.3±0.1 s
Relative	100 %	120 %	0.5 %	0.9 %

Table 5.4: Mean runtime (5 runs) for evaluating a single explanation method (IG). † refers to ROAR, and ⋆ to our ROAD.

	α=0.1	α=1	α=10	α=100
hr=0.01	0.0131	0.0164	0.0090	**0.0085**
hr=0.1	0.0113	0.0133	0.0131	0.0101
hr=0.3	0.0172	0.0183	0.0151	0.0127
hr=0.9	0.0303	0.0484	0.0379	0.0088

Table 5.5: Mean-Squared-Errors for GAIN on CIFAR-10 using different hyperparameter choices.

approach (d) appears to produce the most realistic imputed image, surpassing both the fixed value (b) and noisy linear (c) imputations. Despite its inability to flawlessly replicate the original, such as the noisy background and altered body color, discerning the masked area in (d) remains challenging. An expertly trained imputation predictor confirms that the GAN method is the most aligned with the ideal, Minimally Revealing Imputation.

Nonetheless, GAN imputation has its limitations. It might add extraneous elements absent in the original, like the new patterns on the dog's body in (d). Its effectiveness dwindles significantly when many pixels are missing (see Figure 5.8). Furthermore, fine-tuning its hyperparameters is both time-consuming and costly, detracting from the model's efficiency and simplicity. In contrast, our Noisy Linear imputation avoids these issues and is more efficient in implementation as shown in Table 5.6. Therefore, given these considerations, Noisy Linear Imputation is recommended for use in our evaluation framework.

Strategy	Retrain			No-Retrain		
	fixed†	lin	gan	fixed	lin⋆	gan
Time	3903±117 s	4686±2 s	6421±74 s	18.0±0.1 s	33.3±0.1 s	35.0±0.1 s
Relative	100 %	120 %	164 %	0.5 %	0.9 %	0.9 %

Table 5.6: Mean runtime (5 runs) for evaluating a single explanation method (IG) on three imputation operators. † refers to ROAR, and ⋆ to our ROAD.

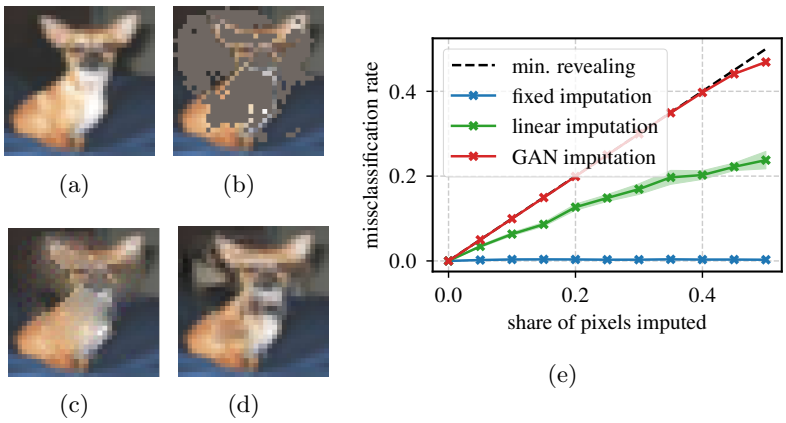
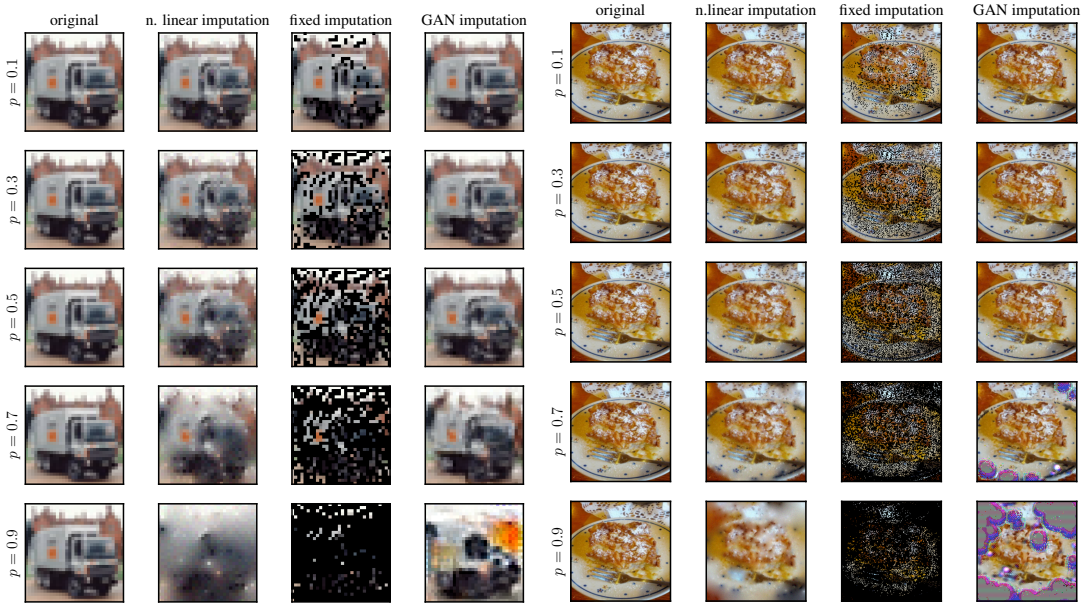


Figure 5.7: The considered imputation operators. When 30 % of the original image (a) are removed, they can either be completed by a fixed value (b) or by our proposed Noisy Linear imputation (c) or GAN imputation (d). Training of an imputation predictor (e).



(a) Sample from CIFAR-10 (b) Sample from Food-101

Figure 5.8: Sample images from CIFAR-10 and Food-101 imputed with the three methods considered in this work for different percentages.

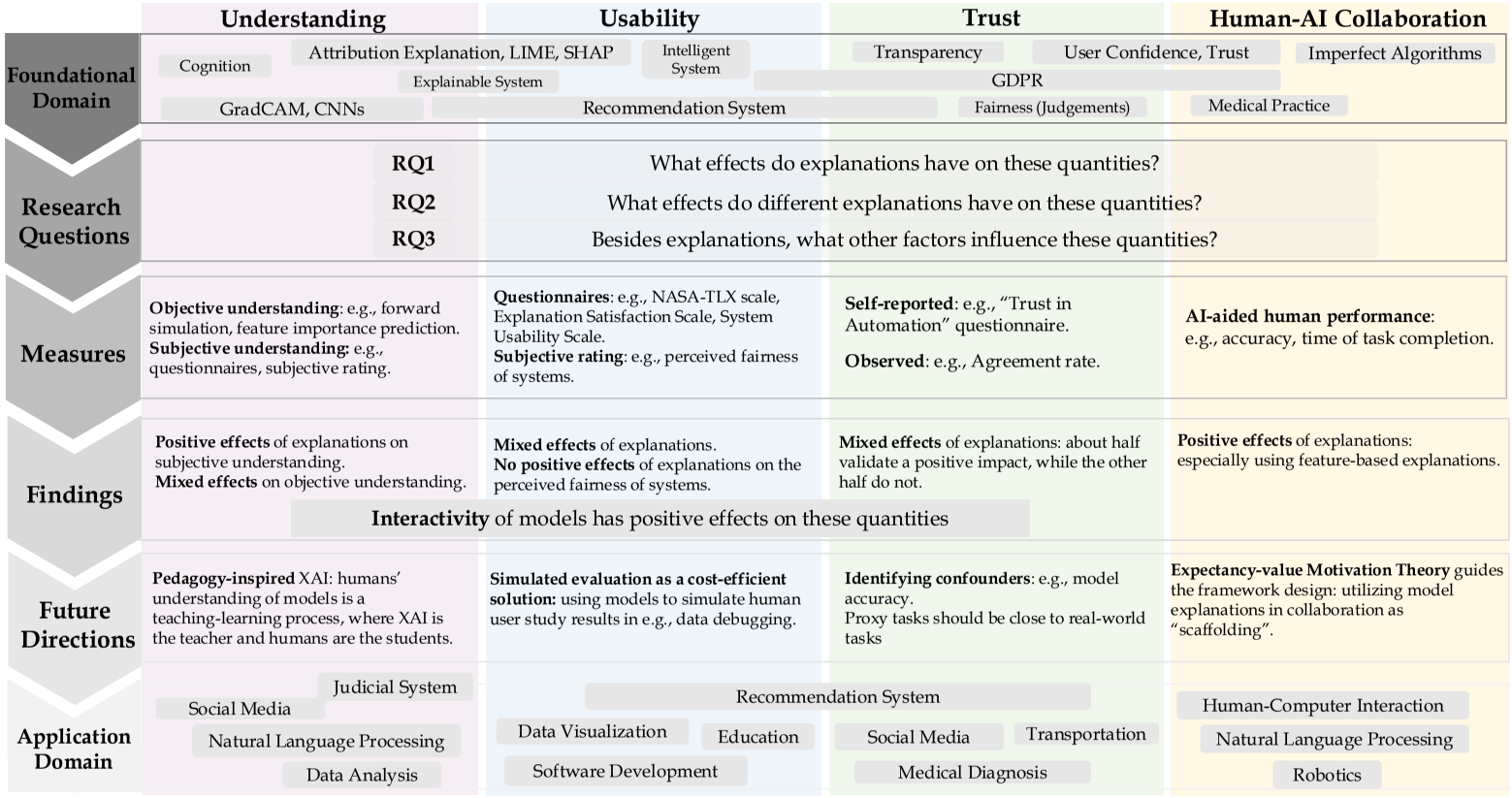


Figure 5.9: Roadmap of our literature analysis. We find out the foundational works of core papers and their application domains using a data-driven method introduced in Appendix E.1. Three main research questions in user studies are distilled from core papers. We distill important messages in this figure for each category: methods related to measures, findings of the research questions are summarized, and future directions based on the findings.

5.3 Guidelines for Human-grounded Evaluation

As highlighted in the previous section, automated evaluation yields inconsistent outcomes and may not fully reflect the human perspective on model explanations. Therefore, conducting user studies becomes essential in XAI to more comprehensively gauge the efficacy of explanations, particularly for applications in real-world scenarios [48]. Yet, only a minor fraction (approximately 20%) of XAI evaluation studies incorporate human subjects [44]. While there have been efforts to develop taxonomies and define the nuances and impacts of various human-centric evaluations [49, 50, 51], a systematic discussion on the recent advancements in user studies and their outcomes is still lacking. Furthermore, Yang et al. [52] observe that the field of XAI is evolving disparately across different communities, such as machine learning and human-computer interaction (HCI). This observation underscores the need for effective guidance in designing XAI user studies, which is essential for aligning the objectives of the XAI algorithm and application designers with the actual needs of users. Section 5.3 aims to address this gap in contemporary XAI user study methodology by offering practical guidelines derived from a thorough and structured review of the relevant literature.

Drawing upon the messages from prior research, we present useful guidelines for conducting XAI user studies, serving as a comprehensive checklist for those in the field. Specifically, we examine publications from the past *five* years in prominent conferences such as CHI, IUI, UIST, CSCW, FA(cc)T, ICML, ICRL, NeurIPS, and AAAI, focusing on the intersection of "explainable AI" and "user study". Our initial collection comprised over one hundred papers. After a detailed review, we narrow down our selection to **97** core papers that meet our specific requirements: (1) the implementation of explainable models or methods, and (2) the involvement of human participants in evaluations. Keywords used can be found in Table 5.7.

	Explainable AI	**User Study**
Keywords	XAI, explainable AI, explanation, explainable, explanatory, interpretable, intelligible, black-box, machine learning, explainability, interpretability, intelligibility, explain attribution, feature	user study, participant , human subject, empirical study, lab study, user evaluation, human evaluation

Table 5.7: Keywords for our paper search query. Paper must contain at least one keyword from each group.

5.3.1 Analysis

To conduct an analysis of the papers gathered on user studies in XAI, we first sort them into four distinct groups, each defined by its specific objective. From these papers, we extract three key research questions that focus on how the explanations provided by models impact these objectives. Our analysis includes a summary of the methods employed in these studies for measuring the objectives. We discuss notable conclusions drawn from these papers and suggest potential future research directions based on these insights. Moreover, we explore both previous works that these user studies reference (i.e., their foundational literature) and follow-up papers that cite these studies. This exploration provides insights into the key works and the evolving trends in human-centric XAI research. Figure 5.9 illustrates a comprehensive overview of our analytical process. This section outlines the criteria applied for their categorization. Then, we introduce the foundational and application in these documents, offering a broader view of XAI user studies.

Categorization of User-Study Objectives. In light of the comprehensive nature of core papers addressing various aspects of model explanations, we organize them into clusters to better study their commonalities and differences. [122] defines *interpretability* within machine learning systems as the ability to explain or present model predictions in understandable terms to a human. The authors contend that beyond aiding understanding, interpretability plays a crucial role in qualitatively determining the fulfillment of other essential criteria such as *usability* and *trust*. During a profound study of the relevant literature that was previously selected, we identified four sensible categories, that are derived from the considered dependent variables in user studies (desiderata of interpretability). These four categories are **trust, understanding, usability**, and **human-AI collaboration performance**. We observe from these papers that typically, each measure aligns with only one of these categories, making this method of classification both intuitive and practical.

These categories represent various functions (goals) of XAI. Since interpretability is described as "*the ability to explain or to present in understandable terms to a human*," the primary aim of XAI is to foster human comprehension. Specifically, comprehension in the realm of interacting with an ML model means a user's understanding or "mental model" of the model's workings. This understanding is enhanced through system interaction and straightforward explanations [40]. The term "Usability" is extensively explored in the field of HCI [209] and is considered a crucial requirement for XAI [122]. As defined by [210], usability refers to the degree to which a product enables users to perform their intended tasks successfully, efficiently, and satisfactorily. Therefore, this category includes user studies that leverage model explanations to aid users in completing specific tasks. In evaluating usability, various factors are assessed, such as the system's ease of use and the cognitive load it demands. The concept of "detection of undesired behavior" pertains to scenarios where explorations reveal discriminatory tendencies in a model, like the employment of features that are not desired. "Trust" in AI is summarized as the user's confidence in a model's accuracy, a personal comfort

level with understanding and using it, and the willingness to let the model make decisions [147]. It contains more requirements. Performance in human-AI collaboration involves situations where the AI system offers predictions, while humans maintain ultimate decision-making authority. In such contexts, the utilization of model explanations contributes to achieving performance that surpasses the capabilities of either the AI system or the human decision-maker when acting independently. The reviewed user studies contain various categories focusing on dependent variables, specifically concerning the operation of XAI methods. These operations are connected to the models' reasoning and knowledge representation. Exploring XAI from a broader viewpoint, particularly regarding generalization and robustness, continues to be a vital area for further investigation via user studies.

Foundations of User Studies. Our data-driven bibliometric analysis of references in key papers, presented in Figure 5.9, underscores prominent research themes in the "Foundational Domain". Model explanations and interpretability emerge as core elements. This includes studies introducing explanation techniques such as LIME [86], SHAP [211], and diverse attribution methods. These techniques are regularly explored in research focusing on understanding and usability. Moreover, convolutional networks, frequently used in experiments, leverage tools like GradCAM [212] and assorted saliency maps for crafting model explanations. Significantly, a substantial portion of research within recommender systems includes studies on Explainable Artificial Intelligence (XAI), particularly those focusing on recommendation solutions. The European Union's General Data Protection Regulation (GDPR) [213] often features in key discussions, especially related to the "right to explanation" debate [214], a topic that has greatly impacted the progression towards more explainable AI systems. Although these explanations are ultimately aimed at human users, there is a lack of emphasis on human comprehension. For example, references to studies in "Cognition" are relatively sparse, especially when compared to algorithmic subjects. Millecamp et al.[215] advocate for the incorporation of social sciences, like cognitive science and psychology, into XAI theory. However, references to psychology are minimal, indicating that few XAI user studies investigate the psychological aspects of XAI. We point out an emerging field in XAI frameworks that are grounded in theories of human cognition and behavior [40], which can provide valuable theoretical underpinnings and conceptual tools for a better assessment of XAI from user perspectives. Further information on these common references is available in Appendix E.2.

Impact of User Studies. Figure 5.9 illustrates various applications that rely on insights from core research papers. We observe a broad spectrum of applications in terms of user comprehension and trust. Trust, for instance, is a recurring theme in areas such as medical diagnosis and transportation, underscoring its importance in high-stakes situations. In subsequent studies, recommendation systems have emerged as a key area of focus. Research on user-friendliness significantly influences domains like data visualization, software engineering, and educational technology, where models often act as

facilitators for end users. Human-computer interaction enhancements are especially vital for advancing robotics and natural language processing. The notable presence of recommendation systems in both foundational research and their influential outcomes suggests that XAI is a fundamental element in modern recommendation systems. For a detailed survey of these essential studies and their application areas, refer to Figure E.1.

5.3.2 Guidelines

The guidelines, as demonstrated in Figure 5.11, offer advice to avoid common pitfalls that researchers could easily overlook. Our guidelines are structured sequentially to align with the phases of user studies: before, during, and after user study, each corresponding to study design, execution, and analysis, respectively.

5.3.2.1 Before User Study

In initiating a user study, the main task involves selecting objectives to be measured. One can choose between two types of measurements: a general measurement or one tailored to the specific application in question. The general measurement is adapted from existing, established research, such as adopting measures like "trust in automation" [216, 217, 218] or "general trust in technology" [219, 220]. In order to quantify "trust" effectively, it is necessary to review how it has been previously defined and measured in the realms of social sciences, XAI, and technical fields [221]. On the other hand, an application-based measurement is chosen based on the specific objectives of the application, like in a chess game [222], where the metric is the proportion of games won by humans aided by model explanations (Human-AI collaboration).

Examined literature suggests that past research often faces challenges in demonstrating the superiority of XAI, especially against a baseline group lacking explanations. In scenarios where only varying explanation methods are evaluated, one technique inevitably emerges as superior, yet the overall advantage of XAI remains undisclosed from the studied papers. Hence, to truly ascertain the efficacy of XAI, it is crucial to contrast it with a baseline that lacks any explanations. For studies aiming for a comparative approach, employing baselines like random explanations is advisable [223, 94, 224]).

When deploying a proxy task, its difficulty should be gauged and monitored carefully. In the past, the forward simulation task has been criticized as being unrealistically complex for domains such as computer vision [225]. Therefore, alternatives like feature importance queries [226] and manipulatability checks [227, 228] have been suggested. It is also important to select a proxy task that, while being more straightforward, still retains numerous aspects of the intended application [122]. Notably, the proxy task should be designed close to the final anticipated application, as even slight differences in the tasks may void the validity of the findings on the proxy tasks in the real world [229].

The reliance on measurements on how the measured entity is defined is a common theme. As an example, the study in [230] quantifies objective understanding through the metric of failure prediction, which assesses the accuracy of a user's prediction when the model's prediction is incorrect. When it comes to subjective metrics such as subjec-

tive understanding or trust, one-dimensional approaches (for example, asking a single question like *"Do you trust the model explanation?"*) are limited in their ability to capture the various aspects of the measured entities [231]. Additionally, there is often a weak correlation between subjective queries and behavioral metrics. This is evident when users express trust in a model, yet their actions do not align with the model's recommendations [232]. This disparity is also observed between objective and subjective understanding, as highlighted in studies like [233, 46, 234]. To address these limitations, it is advisable to employ both self-reported and observed methods concurrently.

Various psychological concepts can be deployed to assess the diverse aspects of human interaction with XAI. For example, within the expectancy-value framework, the *subjective task value* is commonly employed to examine the individual motivation for specific actions [235], an area yet to be extensively explored in XAI encounters. This subjective task value is composed of intrinsic value (pleasure), attainment value (personal significance), utility value (practicality), and cost (required effort or time) [235, 236]. An effective explanation interface is expected to have a positive correlation with the subjective task value, thereby enhancing an individual's interest and willingness to engage with model explanations. Concerning the cost aspect, particularly the cognitive load in utilizing model explanations, it is generally assessed in contemporary research using standard Likert scales [237, 238]. Scholars in cognitive load are examining the efficacy of various visual representations in assessment scales, beyond traditional numerical Likert scales, such as pictorial scales including emoticons (faces showing different emotions), or images depicting varying weights [239]. Their findings indicate that while numerical scales are better suited for complex tasks, pictorial scales are more effective for simpler tasks.

The adoption of online platforms for pre-registration, such as AsPredicted[3], has become popular [240]. This procedure involves researchers uploading a detailed plan of their intended study on the Internet prior to the commencement of data collection. This pre-registration document typically contains various aspects, including the variables to be measured, the hypotheses, the criteria for data exclusion, and the predetermined sample size. Thorough pre-registration is instrumental in countering concerns of selective reporting or p-hacking [241], thereby enhancing the research's trustworthiness. Furthermore, conducting expert interviews and preliminary studies using a think-aloud method [242], as exemplified in references [228, 243], are frequently cited as valuable techniques. These methods aid in refining the explanation system and the design of the study, and in acquiring initial qualitative insights or augmenting the qualitative analysis [229, 244].

In preparation for a user study, careful planning of distinct steps and having explicit plans for various scenarios are crucial. Informing participants beforehand about the meeting location with researchers, necessary items to bring, and ways to prepare for the study is beneficial. Should the study be in-person, it is advisable to remind participants a day in advance and provide contact details for assistance with locating the site or if cancellation is needed. Upon their arrival, researchers should be equipped with a comprehensive plan including all phases of the study. The protocol needs to include minute

[3] `https://aspredicted.org`

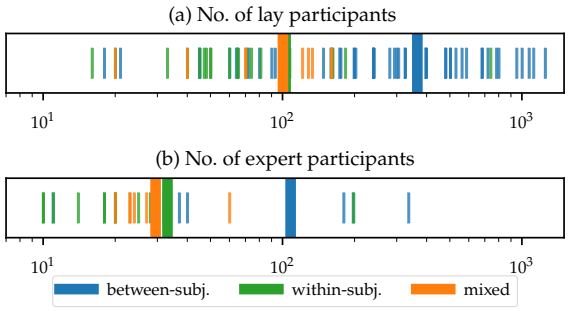


Figure 5.10: Distribution of participant numbers in the surveyed user studies by design and participant type (each bar represents one study). Per-design means are indicated in bold.

details, such as where to store backpacks, water bottles, and lunch boxes, and strategies for handling unforeseen events like uncooperative participants and multifunctional systems. A critical element of the study is ensuring participant consent, with special attention needed when involving vulnerable groups like children and pregnant women, necessitating potentially different consent methods. Preplanning the experiment script offers the advantage of refining language to eliminate unintentional cues. Researchers can inadvertently influence participants with their verbal and nonverbal actions, potentially biasing results towards their expectations [221]. To maintain the experiment's validity and safeguard data integrity, investing effort in creating a thorough experiment script is highly valuable.

5.3.2.2 During User Study

Ensuring an adequate number of participants is essential for robust user study analysis. For a general idea of typical sample sizes, see the participant data in Figure 5.10, which examines the number of subjects in various experimental setups. Typically, around 350 individuals with no special skills are recruited for between-subject experiments. However, it is important to emphasize that the number of participants needed varies greatly depending on the design of the study and should be assessed on a case-by-case basis, such as through a statistical power analysis [245]. IAdditionally, recruited participants should have the same knowledge background as the end users that applications are designed for. For instance, when evaluating an interface explaining loan approval decisions to bank customers, including only computer science students who might already understand model explanations, is inappropriate. Since AI applications are intended for diverse audiences throughout their development cycle, the nature of model explanations must adapt accordingly [246].

Maintaining the integrity of collected data requires implementing checks to ensure attention and prevent manipulation, particularly in extensive or web-based surveys in-

volving non-expert participants. Kung et al.[247] advocate for these measures, ensuring they do not affect the reliability of the scales used. In experiments where subjects are exposed to multiple conditions, randomizing the sequence of these conditions is crucial to eliminate the influence of sequence order, as discussed by Panigutti et al.[248]. This is important because participants might acquire knowledge from earlier conditions. To address this issue of learning bias, Tsai et al. [249] employ a Latin square design.

5.3.2.3 After User Study

Following the gathering of data, statistical evaluations are conducted to identify significant impacts. The selection of appropriate tests is influenced by the experimental setup and the characteristics and distribution of the collected data. Commonly, ANOVA tests and T-tests are employed for comparing distributions across various conditions. For mediation analysis, Structural Equation Models (SEM) or multi-level models are typically utilized. It is essential to perform checks for distributional assumptions. In cases where Likert-type data is gathered, as is often the case in questionnaires, non-parametric tests like the paired Wilcoxon signed-rank test or the Kruskal-Wallis H test for multiple groups are advisable to circumvent the need for normality assumptions.

When aggregating various measures into a single tool, evaluating the legitimacy of this combination using reliability assessments like tau-equivalent reliability (commonly referred to as Cronbach's α) is crucial. For instance, when integrating objective and subjective measures of a concept like understanding, it is vital to ensure adequate concordance. In situations where multiple elements (such as data points or visual representations) are evaluated by numerous assessors, statistical tools like Cohan's κ and Fleiß's κ for more than two evaluators [250] are useful for determining the level of agreement among these assessors that exceeds random chance, providing a gauge for the dependability of these evaluations.

In the concluding stage of manuscript preparation, it is important to include comprehensive details to enable the audience to gauge the study's explanatory strength. At the participant level, this entails providing the total count of participants, the distribution of participants across different treatment groups, details of their recruitment, the process of obtaining consent, the incentives offered, and the specific treatment conditions they underwent. Additionally, presenting some descriptive statistics of the gathered data aids in evaluating the suitability of the statistical methods employed. In terms of analysis, it is important to outline the process for verifying the assumptions underpinning the statistical tests applied and to specify the precise version of the test utilized (for instance, specifying the use of "a two-way ANOVA with independent variables X and Y" rather than a general reference to an ANOVA test).

5.3.3 Discussion

Automatic vs. human evaluations. Automatic evaluations align with functionally-grounded metrics as elaborated in [122, 44]. Such metrics are designed to rigorously assess key aspects like the "faithfulness" "fidelity," or "truthfulness" of model explana-

Before the User Study

Measured Quantity

- measurable quantity according to a general definition or application-based measure
- conceptualize measures in social sciences and in the context of XAI

Control Group

- include baseline without explanations
- include baseline with other (e.g., random) explanations in case of a comparative study

Proxy Task

- carefully balanced difficulty
- close to the anticipated real-world application
- participants might have some background knowledge of the task but not too much to bias their judgment or performance

Measurements

- measurements are aligned to definitions
- independent measurements vs. participant-interdependent measurements
- aggregation of the measured quantities across participants
- avoid one-dimensional measures
- pre-register study online to prevent p-hacking
- run a pre-study and expert interviews for initial feedback on the envisioned measures

During the User Study

Experiment Script

- prepare a script with explicit instructions and procedure of the experiment
- obtain consent
- avoid inadvertent cues

Participants

- recruit a representative number of participants for each study subgroup according to the experimental design and required expertise
- adhere to the real-world distribution of users according to the application at hand

Data Quality

- attention checks during a long survey
- preventing order and learning effect (within-subject)

After the User Study

Statistical Verification

- comparison between conditions: use ANOVA, t-test, etc.
- many dependent variables: use SEM, multi-level model, etc.
- distributional assumption checks for different statistical tests (e.g., normality tests)
- Compute measures like Cronbach's α for verifying scales and Fleiß's κ for detecting randomness in the answers of participants or consistent behavior, i.e., agreement beyond chance in their answers

Effect Verification

- correlation between the subjective and behavioral measures
- agreement or disagreement with expected (e.g., theoretical) results

Reporting

- (aggregated) measurements (across participants) and their statistical validity
- model and hyperparameters for reproducibility
- number of participants, characteristics of participants, recruitment
- exact conditions
- analysis methods (name and parameters of the test)

Figure 5.11: Summary cards of the guidelines extracted from past XAI user studies

tions, as detailed in sources [44, 42, 251]. The concept of faithfulness in explanations is articulated as the degree to which explanations signify genuinely significant features within the input, as delineated in [251]. Automatic evaluations are tailored to gauge general objectives, independent of specific downstream tasks, contrasting with human evaluations, which are contextualized to particular use cases. While automatic evaluations objectively scrutinize the accuracy of explanations in mirroring models, human evaluations focus on assessing the interpretability of models through explanations from a human perspective. It is worth noting that there are algorithms designed for automated evaluation that endeavor to align with human assessments, a topic we will explore later.

For human-subject experiments, it is imperative that all utilized explanations pass through rigorous automatic evaluations, ensuring they genuinely represent the model's workings. This preliminary verification step is crucial to affirm the empirical study's validity and to prevent misleading users with unfaithful explanations. However, it is observed that many current human-subject experiments lack thorough pre-verification of explanation's functional faithfulness. Relying on unfaithful explanations risks measuring merely the placebo effect rather than genuine understanding. Ideally, an effective explanation should not only be faithful to the model but also comprehensible to users.

Identifying and handling confounders. The current research highlights the susceptibility of model explanation studies to considerable confounding factors. Papenmeier et al.[252] have shown that user trust is more strongly impacted by the accuracy of the model than by the fidelity of its explanation. In a similar manner, Yin et al.[253] have established that both the accuracy score perceived by users and the one presented to them play a significant role in the formation of trust. [254] discovers another confounder, namely the ambiguity of input samples. If the class of input sample is intuitive for users to recognize, the users tend to trust model explanations, even the model explanations are in low quality (low faithfulness).

Effective explanations should also expose the model's limitations. When users encounter unexpected explanations, they might react with negative feedback, affecting the evaluations of the explanations. Thus, it is crucial for explanations to aid users in *calibrating* their trust [255, 256], that is, to trust the model's decisions when accurate but to be skeptical otherwise. Regarding this, there is a debate on how to approach such scenarios: In assessing model fairness, numerous studies [257, 258, 259, 216, 260] view an increase in perceived fairness positively. In contrast, Dodge et al.[261] consider a decrease in fairness as a positive outcome. Additionally, other variables play a role, such as the timing of model errors (Nourani et al.[262]), and the specific characteristics of the models (Ross et al.[228], Poursabzi et al.[263]).

Mitigating personal biases for XAI. Most existing XAI techniques and their associated user studies offer *one-size-fits-all* solutions, overlooking the individual biases ingrained in users' mental frameworks. These biases significantly shape how users perceive AI models, a factor that must be integrated into XAI's design, development, and evaluation processes. To address this, several studies aiming to elucidate reinforcement

learning policies have applied cognitive science theories. They formulate a human user model [53, 54, 55, 56], based on which tailored explanations are generated and assessed for their effectiveness in catering to individual user models. In XAI, references [57, 58] adopt a Bayesian Teaching framework to accurately gauge human perceptions of model explanations. It is important to note that in user studies, participant feedback can vary significantly due to differences in cultural and educational backgrounds [264]. This type of personal bias can be mitigated by involving a diverse and large sample of participants, representative of the intended audience. Therefore, we strongly advocate for the consideration of personal biases in the development of XAI to ensure more effective and personalized user interactions.

Simulated evaluation as a cost-efficient solution. Given the high costs associated with conducting experiments involving human subjects, Chen et al. [265] introduced a simulated evaluation framework (`SimEvals`) designed to pre-select potential explanations for user studies. This is achieved by evaluating the explanatory power of these explanations. Specifically, they explore three scenarios where model explanations are applied: forward simulation, counterfactual reasoning, and data debugging. The effectiveness of different explanations is assessed based on human performance across these tasks. A significant disparity in effectiveness between the two explanation types indicates a need for further investigation. In parallel, early experiments using large language models to simulate human-like textual responses in specific contexts have yielded unexpectedly human-like results [266]. However, as affirmed by Chen et al. [265], substituting human evaluation with this simulated approach is currently impractical due to factors such as cognitive biases impacting human decisions. Enhancing the simulation of human evaluations requires more focus on emulating human cognitive processes. Simultaneously, researchers in XAI should employ existing models that approximate human cognition to enable rapid prototyping and assessment of explanations.

5.4 Conclusion

This chapter discussed the use of both automatic and human-grounded evaluation metrics in XAI. It began with addressing a significant bias known as Class Information Leakage through masks, which is prevalent in automatic evaluation metrics. This was explored through both theoretical and experimental analyses. To counteract this bias, the chapter introduced a novel imputation technique termed Noisy Linear Imputation. Building on this, a new evaluation strategy, ROAD, was presented. ROAD not only effectively mitigated the identified bias but also demonstrated remarkable efficiency when compared to ROAR, achieving up to a 99% reduction in runtime. The accessibility of our method is a key advantage, due to its minimal resource demands, making it highly suitable for practical applications. Beyond automatic evaluation metrics, the chapter offered comprehensive guidelines for conducting human user studies in XAI, drawing upon an extensive review of 97 SOTA works in the field. Overall, this chapter contributed valuable perspectives for future developments in XAI evaluation methodology.

6 Tailoring Explanations to User Expertise

6.1 Introduction

As the significance of AI systems in our daily lives grows, it becomes a challenge for human users to understand the decisions these systems make. Experts from various fields, including AI research, legal frameworks, and design, have identified the importance of making AI systems transparent. This is crucial because many AI models function as a "black box," meaning their decision-making processes are not easily interpretable or comprehensible to humans. Therefore, developing solutions for AI transparency, such as techniques for XAI, is vital for ensuring users' safe use and proper understanding of these systems. In the various contexts of XAI, this study primarily addresses the explanation of tasks related to image classification [267]. Attribution explanations, exemplified by methods like GradCAM [212], SHAP [211], and LIME [86], are extensively employed in current XAI approaches for image classification. Although these methods contribute to the foundation of our research, they consistently overlook a critical aspect: the consideration of human factors, which may be attributed to the challenges in incorporating models of human users.

We advocate that modeling human aspects is pivotal in XAI research, given that the concept of explainability is fundamentally oriented towards human understanding [40]. Several studies in the realm of explaining reinforcement learning strategies employ theories from cognitive science for creating models of human users, which then serve as the basis for crafting explanations [53, 54, 55, 56]. Aligning more closely with our area of interest, the research conducted by [57] and [58] adopts the Bayesian Teaching approach to conceptualize human perception, subsequently leading to the generation of explanations that are centered around human cognition.

A disadvantage of previous studies is their uniform approach to all users, assuming a single explanation set is effective for *every* user. On the contrary, our method seeks to create customized explanations for individual users, focusing on their *specific task expertise*. This strategy is influenced by human annotator models in active and imitation learning, as seen in [59, 60]. Similar to these works, our user model aims to capture both the decisions and reasoning process (expertise in concepts used for image classification) of the human user in the context of a given classification task.

To address the void in research where *personalization* lacks in the explanation process, we introduce the framework that provides **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise (I-CEE). Drawing inspiration from prevalent XAI methods in image classification, our framework adopts the *explanation-by-examples* strategy, offering attribution explanations (local explanations) for selected training data instances. The novelty of I-CEE lies in its user-specific approach to choosing example explanations. When dealing

with an image classification task, I-CEE identifies a group of m task-relevant concepts. It then characterizes the user's task-specific expertise as an m-dimensional vector, each element ranging from $[0, 1]$, indicating their expertise level in each concept. Utilizing this user model, I-CEE subsequently opts for the local explanations most suitable to bridge the gaps in the user's knowledge.

I-CEE is designed to accelerate user understanding of the decision-making process of machine learning models by selecting the set of local explanations that can best increase the user's task-specific expertise. This approach contrasts most existing XAI work, which often relies on random or one-size-fits-all local explanations. Previous methods miss the chance to accelerate understanding of models by offering explanations customized to the user's needs.

The contributions of this study are outlined as follows:

- This research identifies the potential for creating tailored explanations for the decisions of image classification models and introduces a new framework named I-CEE. This framework marks a step forward in the direction of human-centered explanations.
- The effectiveness of I-CEE is evaluated by assessing the simulatability of the explanations it produces across four datasets. The findings show that our method surpasses state-of-the-art XAI frameworks in simulatability, enhancing users' ability to predict the model's labels.
- We conduct comprehensive human-subject studies ($N = 100$) to evaluate our framework. The experimental outcomes reveal that our framework is more successful in aiding users to comprehend the decision-making process of the ML model compared to the leading method, Bayesian Teaching [58]. Moreover, it is subjectively more preferred by the participants.

6.2 Problem Statement

Consider a machine learning classifier denoted as f or the *target model*, which is trained on a dataset $\mathcal{D}$ consisting of image-label pairs $(\mathbf{x}, y)$. This classifier, defined as $f : \mathbb{R}^d \to 1 : K$, assigns to each input image $\mathbf{x} \in \mathbb{R}^d$ a label $y \in 1 : K$, meaning $f(\mathbf{x}) = y$, with K representing the total number of classes. In some cases, the label y predicted by the classifier might not correspond to the actual label y^*. To provide clarity on how these target models function, various feature attribution techniques have been introduced, offering local explanations [86, 211]. These techniques allocate an importance value to each input pixel, represented as $\mathbf{e} \in \mathbb{R}^d$, often illustrated through a saliency map. Within the *explanation-by-example* approach, users are presented with a selection of training data images, along with their local explanations and predictions, denoted as $(\mathbf{x}, \mathbf{e}, y)$. Given the limited time users have to understand the model, choosing the most informative examples is important.

In learning through examples, we address the challenge of identifying the most enlightening set of example images (along with their respective explanations). To define this

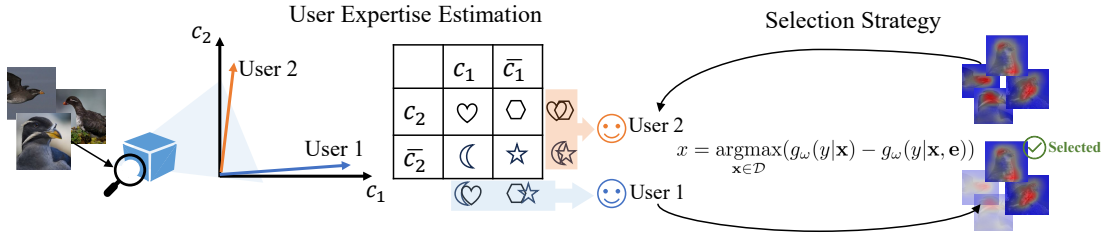


Figure 6.1: Overview of I-CEE. **Left:** The target model is first projected into a concept space, which is then used to estimate user expertise. Two users are illustrated. User 1 uses the concept c_1 in the reasoning process and can differentiate only two classes (highlighted in blue). Likewise, User 2 is able to distinguish two classes based on c_2 (in orange). **Right:** Based on user models, explanations with images $(\mathbf{x}, \mathbf{e})$ in the training set that maximize Hypercorrection Effect are selected and delivered to the users.

issue, we start with three fundamental elements: the target model denoted as f, a data collection symbolized by $\mathcal{D}$ (with the total number of data points being N, represented as $|\mathcal{D}| = N$), and a method for attributing features to produce local explanations. With these components in place, our goal is to select a smaller subset $S \subset \mathcal{D}$ from the training dataset, containing $M \ll N$ images, which are most effective in enhancing *simulatability*. This means they are instrumental in aiding users to anticipate the outcomes predicted by the machine learning model. As the problem objective hinges on a human-centered metric, its successful resolution warrants a human-centered approach.

6.3 I-CEE: Image Classification Explanations tailored to User Expertise

This section introduces our methodology I-CEE, containing two stages (Figure 6.1). First, the system constructs a model of the user, focusing on assessing their expertise related to the task at hand (refer to lines 3-4 in Algorithm 1). Subsequently, I-CEE employs this user model alongside a query strategy to choose example images and explanations that provide valuable insights (see lines 5-8).

6.3.1 User Expertise Estimation

The practice of predicting the labeling outcomes of an ML model by a user can be seen as a form of image annotation. In this scenario, the annotators may have varying levels of *expertise* or strengths, influencing their labeling decisions [59]. For example, while some users are better at recognizing textual patterns, others may find understanding shapes more natural. During this annotation activity, humans often engage in "concept-based thinking" for reasoning and decision-making. This involves identifying commonalities among different examples and categorizing them in an organized manner based on their similarities [268, 269, 270]. In acknowledging these elements of human cognition, and

Algorithm 1 I-CEE

1: **Input:** Target model $f(\cdot)$, data $\mathcal{D}$, user annotation y^u.
2: **Output:** A set of example images and explanations $\mathcal{S}$.
3: Discover concepts by solving Eq. 6.3.
4: Estimate user expertise by solving Eq. 6.4.
5: **for** $\mathbf{x} \in \mathcal{D}$ **do**
6: Calculate Hypercorrection Effect for $\mathbf{x}$ using Eq. 6.5.
7: **end for**
8: Return top-K image samples.

drawing inspiration from annotator models in active learning, we devise a model to estimate a user's proficiency in applying different relevant concepts for the task.

We first discover the underlying concepts in the feature space of the target model. Using the discovered concepts, we model a user with a vector representing their ability to utilize each concept when annotating images. Figure 6.2 provides an overview of the user model. To arrive at the model, I-CEE begins with applying the concept discovery algorithm on the target model [268] that aims to recover m concept $[\mathbf{c}_1, \cdots \mathbf{c}_m]$, such that

$$f(\mathbf{x}) = h(\Psi(\mathbf{x})) = h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}))), \tag{6.1}$$

where $\Psi(\mathbf{x}) \equiv [\psi(\mathbf{x}^1), \ldots, \psi(\mathbf{x}^T)]$ denotes T activation vectors. The function $h(\cdot)$ symbolizes the transformation process that converts the intermediate outputs of these activation vectors into image labels.[1]

The concept score $s_\mathbf{c}(\cdot)$

$$s_\mathbf{c}(\mathbf{x}) = \langle \psi(\mathbf{x}^i), \mathbf{c}_j \rangle|_{j=1}^m|_{i=1}^T \in \mathbb{R}^{m \cdot T} \tag{6.2}$$

is designed to measure how well each pair of concepts and activation vector aligns. Similarly, $\Xi_\theta : \mathbb{R}^{T \cdot m} \to \mathbb{R}^{T \cdot n}$ represents a trainable transformation that reinterprets concept scores within the activation space. All concept vectors and their corresponding scores are normalized to unit length.

To discover concepts (namely, determining $\mathbf{c}, \theta$), the objective is to minimize the ensuing cross-entropy loss:

$$\mathcal{L}_{(\mathbf{c},\theta)} = -\sum_{i=1}^{N} y_i \log(h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}_i)))), \tag{6.3}$$

in which y denotes the prediction derived from the target model $f(\cdot)$.

After completing concept discovery (a one-time process), the expertise estimation for each user takes place within the concept space. To achieve this, we freeze all model parameters ($\Psi(\cdot)$, $s_\mathbf{c}(\cdot)$, $\Xi_\theta(\cdot)$, and $h(\cdot)$), which are trained according to Eq. 6.3. This

[1] It is pertinent to note that Ψ and h can be interpreted as the intermediate and final layers of an image classification neural network. Since h and Ψ are not a part of the training process for the user model, we refrain from specifically indicating their parameters, such as weights and biases, in our notation.

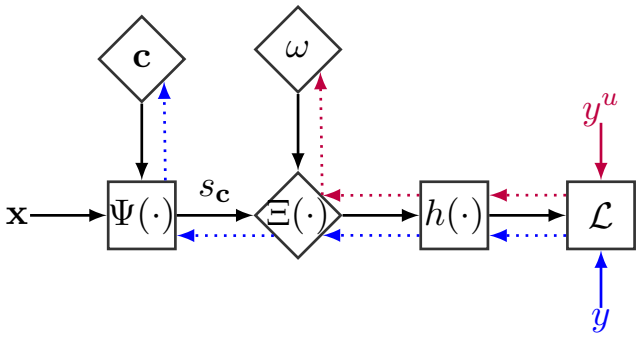


Figure 6.2: User Modeling: Square nodes are deterministic, while diamond nodes are trainable. Loss back-propagated for concept discovery (Eq. 6.3) is marked in blue, while that for expertise estimation (Eq. 6.4) is in red.

allows us to learn an expertise vector $\omega \in \mathbb{R}^m$ for every user. The individual expertise of users is reflected in the distinct values of ω, influenced by their unique domain knowledge and the manner in which they apply concepts to generate predictions. Specifically, users are tasked with annotating images, and their predicted annotations are replicated using ω. The expertise vector ω for each user is derived by minimizing the subsequent cross-entropy loss:

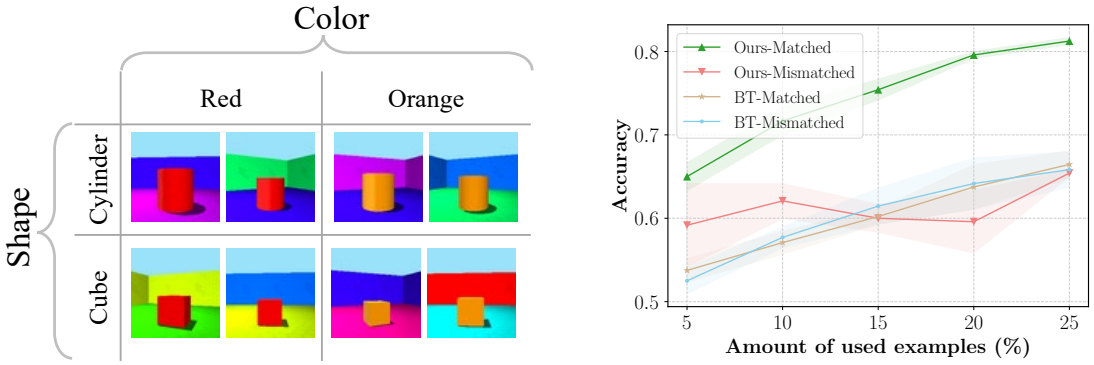
$$\mathcal{L}_\omega = -\sum_{i=1}^{N} y_i^u \log(h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\mathbf{x}_i))), \tag{6.4}$$

where y^u denotes annotated labels collected from the user.

After learning ω, a user model can be represented by $g_\omega(\cdot) = h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\cdot))$. When $\omega_1 \approx \omega_2$, it suggests that Users 1 and 2 share a closely related "reasoning process" due to the resemblance in their concept usage. Similarly, if $\omega \approx \mathbf{1}_m$, it indicates that the reasoning approach of this user closely mirrors that of the target model f.

6.3.2 Selection Strategy

Our objective is to choose a collection of instructive examples that can significantly enhance the user's ability to simulate scenarios. To assess the value of examples, we utilize the hypercorrection effect from educational psychology. As the human needs to learn how the model makes the decision, the model's prediction is viewed as the "correct" answer, whereas the human's disagreed initial belief is the "error". Providing feedback on the correct answer, especially when accompanied by explanations, is essential for effective learning, as indicated by previous studies [271]. The effectiveness of learning from an error example is enhanced when there's higher confidence in the error, meaning lower confidence in the correct answer [272, 273]. In implementing the hypercorrection effect in I-CEE, we select images where, after understanding the model's rationale, the user's confidence in the model's prediction diminishes. We propose that this selection method will foster greater learning achievements. In practical terms, I-CEE is designed to select a subset $\mathcal{S} \subseteq \mathcal{D}$, comprising instances that exhibit the highest Hypercorrection



(a) Synthetic dataset

(b) (Mis-)Matched expertise

Figure 6.3: (a): Overview of four classes in the synthetic dataset. **(b)**: User simulatability accuracy when trained with examples that match/mismatch with the user expertise.

Effect:

$$x = \underset{\mathbf{x}\in\mathcal{D}}{\operatorname{argmax}} \underbrace{(g_\omega(y|\mathbf{x}) - g_\omega(y|\mathbf{x},\mathbf{e}))}_{\text{Hypercorrection Effect of } \mathbf{e}}, \tag{6.5}$$

$g_\omega(\cdot)$ symbolizes the user model, while $\mathcal{D}$ signifies the training dataset. Additionally, $\mathbf{e}$ and y represent the local explanation and machine prediction respectively, corresponding to the image $\mathbf{x}$.

6.4 Experimental with Simulated Users

Prior to implementing a user study, we first assess our method by carrying out comprehensive experiments including simulated users for one synthetic and three authentic image classification tasks. Appendix F provides additional information on the experimental implementation.

6.4.1 Dataset

Synthetic Dataset. To assess the effectiveness of our proposed approach in a simulated environment, we develop a synthetic dataset[2]. This dataset comprises four categories, each characterized by two attributes: color and shape, as depicted in Figure 6.3(a). For example, a user proficient in discerning colors (more so than shapes) would group red cylinders and red cubes under the same category, distinguishing them from orange objects. Conversely, a user adept at recognizing shapes would differentiate between cylinders and cubes, irrespective of color. Other visual elements like angles or background hues are randomly assigned, being non-critical for this classification task. We produce 300 images for each category, allocating 80% for training and 20% for testing purposes.

[2]This dataset is inspired by 3d-shapes [274].

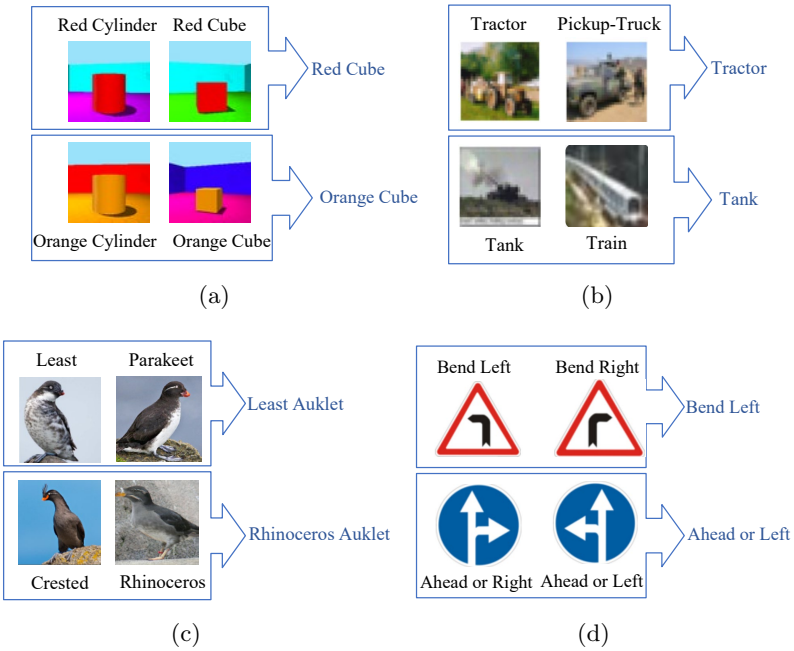


Figure 6.4: Illustration of annotation given by the simulated user on the (a) synthetic, (b) CIFAR-100, (c) CUB-200-2011 and (d) GTSRB dataset. Original label is in black, and the label given by the simulated user is in blue.

Our classification model employs a ResNet-18 [169], and for the generation of explanations, we utilize GradCAM [212]. In terms of annotation behavior, we simulate this using Eqs. 6.3-6.4, following the same modeling methodology as I-CEE.

Realistic Datasets. Our evaluation of I-CEE extends to three real-world datasets: CIFAR-100 [203], CUB-200-2011 [1], and the German Traffic Sign Recognition Benchmark (GTSRB)[275]. We simulate a user with predefined annotations for each dataset, whose behavior deviates from the target model. Specifically, this simulated user in each dataset can only differentiate between two out of four closely related classes. This user's perspective is used as the basis for all method assessments. For example, in the CUB-200-2011 dataset, the simulated user categorizes both Crested and Least Auklet as Crested Auklet, and Parakeet and Rhinoceros Auklet as Parakeet Auklet. We adhere to the original training-test divisions of these datasets and employ the same approach as in the synthetic dataset, using ResNet-50[169] for classifier training and GradCAM to generate explanations.

6.4.2 Implementation of I-CEE

This section provides an overview of the implementation details of our proposed method. This encompasses the training procedure for our simulated user model, as well as an explanation of the selection strategy we employ.

We choose four visually similar classes in each dataset for training simulated user models. This restriction to a limited number of classes is intentional, as our goal is to employ these classes in examining the comprehension of actual human users through the chosen examples. To simulate a user, we first learn a concept within the latent space through the application of Equation (6.2) and Equation (6.3). In the case of the synthetic dataset, we employ a dimensionality of $m = 8$, where this setting yields a test accuracy close to 100%, eliminating the necessity for a larger dimension. For realistic datasets, users can determine the number of concepts, denoted as m. This decision pertains to the dimension of the expertise vector, $\omega \in \mathbb{R}^m$, as outlined in Equation (6.4). To ascertain an appropriate dimension for ω, we experiment with various concept spaces, each characterized by a different value of m, across realistic datasets. Extensive information on this process is provided in the subsequent section. Ultimately, for each realistic dataset, we establish the value of m as 64.

The user model, denoted as $g_\omega(\cdot)$, undergoes training with user-provided annotations as per Equation (4). During this process, all parameters within the network remain static, except for ω. To create simulated users with varying levels of expertise, we generate simulated user-annotated labels by classifying two similar but different classes into one class. This setup necessitates that the user's expertise differs from that of the target model, as it is unable to differentiate between all four classes. The four chosen classes, along with the user annotations for each class in the realistic datasets, are depicted in Figure 6.4. The training of $g_\omega(\cdot)$ is carried out using the Adam Optimizer, set at a learning rate of $1e^{-2}$, and spans over 40 epochs. Note that this training configuration is also applied in the experiments showcased in Figure 6.5.

We employ Equation (6.5) to select images that enhance users' understanding and insight into the model's reasoning processes through specific examples. Utilizing the trained user model, we compute the likelihood that the input image belongs to class y, as determined by the target model, denoted as $g_\omega(y|\mathbf{x})$. When the input is given as $(\mathbf{x}, \mathbf{e})$, we apply the explanation $\mathbf{e}$ as a weighted overlay (the saliency map), which is derived by normalizing the original saliency map, onto the input image. This method is frequently adopted to assess the impact of explanatory techniques [126, 42].

6.4.3 Baseline Methods

We evaluate I-CEE in comparison to a recent approach in human-centric XAI, specifically Bayesian Teaching (BT) [58]. In BT, the emulation of a user's prediction behavior for an image class is accomplished using a ResNet-50-PLDA (probabilistic linear discriminate analysis [276]) framework. This method operates on the premise that users engage in Bayesian reasoning, choosing images and explanations that more closely align the user's

understanding with that of the desired model. There are distinct differences in user modeling and example selection between I-CEE and BT.

In evaluating the selection of examples independently, we conduct comparisons with query approaches originating from active learning (AL). In our unique implementation of AL query methodologies for Explainable Artificial Intelligence (XAI), the learner is represented by a simulated user, while the target model assumes the role of the annotator. Our benchmarks include the use of Expected Gradient Length (EGL) [277], Density-Weighted Method (DWM) [278], and a random sampling approach as foundational comparisons. In this study, EGL is utilized to select sample pairs (x, e) which, upon acquiring their annotated labels, induce the most significant alteration in the current model. This alteration is quantified by observing the gradient of the objective function relative to the model's parameters. Nevertheless, EGL's selection may include atypical samples that cause pronounced gradient shifts. Addressing this concern, [278] suggests combining a density-weighting approach with EGL's querying methodology. Here, each sample's weight is determined based on its mean resemblance to other samples in the dataset. Our contribution in this research involves augmenting EGL with an adjustment for the belief change in EGL calculations when e is part of the input, termed as EGL-Shift. Precisely, we calculate the variation between the EGL values of (x, e) and just x. Through EGL-Shift, our goal is to mitigate the influence of the image itself on the training gradient, instead accentuating the effect of explanations.

6.4.4 Evaluation Metric

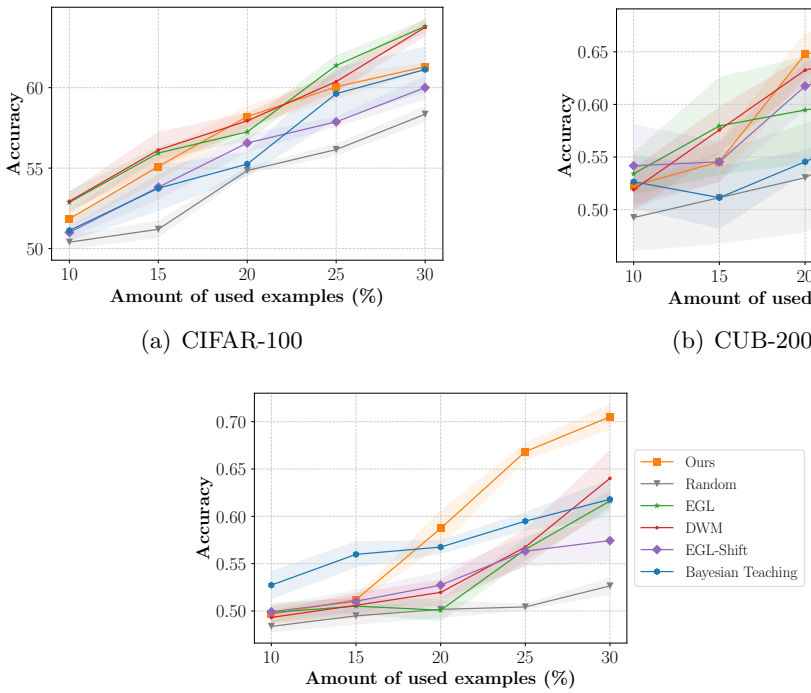
In assessing our approach, we employ the concept of simulatability, a frequently utilized surrogate for evaluating how well a user grasps the decision-making process of the model, as referenced in [46, 279]. Simulatability is quantified by the degree to which a user is able to accurately foresee a model's prediction. This measure is applicable in both simulated experiments and studies involving human participants.

Our approach adheres to the experimental framework suggested in [91, 92] to examine the impact of chosen examples. In particular, each technique yields a prioritized collection of sample images $\mathcal{S}$, with the order determined by the *informativeness* as specified in each method. We express the proportion of the number of sample images $|\mathcal{S}|$ to the volume of training data $\mathcal{D}$ as $p = |\mathcal{S}|/|\mathcal{D}|$. The hypothetical user is then retrained utilizing these sample images $\mathbf{x}$ along with their respective labels $y = f(\mathbf{x})$, where it is important to remember that f represents the target model.

Using the updated user model g'_ω, we calculate the accuracy with which the user predicts the model's forecasts on the test set, i.e., the user's simulatability:

$$\text{Acc} = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{1}(y_i = g'_\omega(\mathbf{x_i})), \tag{6.6}$$

where N_t is the number of samples in the test set.



(a) CIFAR-100

(b) CUB-200-2011

(c) GTSRB

Figure 6.5: Comparison with baseline algorithms using simulated users across three datasets. On the x-axis, the percentage of utilized examples (denoted as p) is depicted, while the y-axis represents the accuracy of simulatability. (Averaged results from 5 runs.)

6.4.5 Experimental Results

Ablation Study. To assess the effectiveness of our model for $g(\cdot)$, we examine (1) the accuracy of ω in representing user expertise, and (2) the benefits of customized explanations adapted to user expertise levels. Our evaluation involves simulating two users on a synthetic dataset: User 1, who relies solely on color for classification, and User 2, who depends exclusively on shape. We infer annotations for each user according to the attributes associated with each class (Figure 6.3(a)).

Upon evaluating each user, we examine their proficiency vectors: ω_1 and ω_2 (where $\omega_i \in \mathbb{R}^8$). Each element within ω_i signifies the user's proficiency in a distinct concept. The four highest values in both ω_1 and ω_2 are inversely related, reflecting that each user possesses expertise in contrasting areas (meaning, each user relies on different concepts during decision-making processes). To assess the effectiveness of our user model in terms of expertise, we conduct a study in which User 1 is trained using a selection of examples tailored to the User 1 model ("Matched"), as opposed to a selection of examples suited for User 2 ("Mismatched"). In Figure 6.3(b), it is evident that the simulated user exhibits high simulatability accuracy when presented with examples that align with their expertise (denoted as "Ours Matched"). In contrast, when examples

are chosen that do not optimally suit the Hypercorrection Effectspecific to the user (referred to as "Ours Mismatched"), the simulatability accuracy significantly drops. This suggests that these examples are inadequate in offering meaningful insights about the target model. Furthermore, our user simulation model is juxtaposed with the Bayesian Teaching approach. Notably, there are negligible differences in simulatability accuracy in both matched and mismatched scenarios under the Bayesian Teaching framework. This implies that Bayesian Teaching may not effectively replicate the diverse behaviors of different users and, as a result, is less proficient in providing examples that enhance user simulatability (showing less improvement in performance compared to our method).

Comparison. We conduct a comparison of I-CEE against various baselines using three distinct real-world datasets, as detailed in Figure 6.5. The evaluation focuses on user prediction accuracy, measured at $p = [10, 15, 20, 25, 30]\%$. On CIFAR-100, our approach consistently surpasses BT and EGL-Shift, although it falls short when compared to EGL and DWM. This outcome may stem from the ambiguous nature of CIFAR-100's explanations, a consequence of the images' low resolution. Consequently, Hypercorrection Effect struggles to be effectively captured, as the explanations are marred by noise. Conversely, on the datasets CUB-200-2011 and GTSRB, our method generally exceeds the performance of other baselines at the majority of these percentages. For example, on the CUB dataset, our method begins to lead in performance beyond the 20% mark. It is worth noting that 20% of the training data equates to 24 images, a count feasibly manageable for human examination, which will be elaborated on in the subsequent section. In the GTSRB case, there's a noticeable performance disparity between our method and the close competitor BT. This could be attributed to differences in user model architecture. Our model mimics user behavior through learning ω in concept space, maintaining the final classifier's effectiveness. In contrast, BT employs a PLDA layer for image classification, potentially leading to inferior results when dealing with images with highly similar latent features, like traffic signs. This is suboptimal since humans excel at identifying key concepts and disregarding visually similar but irrelevant features. Through more accurate user modeling, our method provides valuable learning samples in most scenarios during these simulation experiments.

6.5 Experiments with Human Users

We conducted a study involving human participants with the CUB-200-2011 and GTSRB datasets, adhering to the same parameters used in our simulation tests. These particular datasets are selected due to their increased difficulty and the higher resolution of the images. As a benchmark, we employ Bayesian Teaching [58], which stands as the most advanced and relevant to our area of interest.

In this user study, we aim to study the following research questions:

- **R1**: Our framework selects informative samples that can increase human understanding of the model.

- **R2**: Human understanding of the model is affected by task domains.

6.5.1 User Study Details

We present the procedure and some essential details of our human user study in this section.

Procedure. First, participants were instructed to analyze two categories (out of four) and record the characteristics that help differentiate these categories. The purpose of this exercise was to encourage participants to adopt the mindset of a hypothetical, pre-defined user for whom the model explanations are customized. Subsequently, participants were presented with 20 model explanations, either generated through our approach (experimental group) or via Bayesian Teaching (control group). They were asked to note the features they consider when deducing the model's predictions.

In the evaluation phase, the participants first complete a quiz comprising 15 questions, where they predict the labels assigned by the model (the images for this quiz are evenly drawn from all four categories within the test set). This part is termed "objective understanding". Following this, they assess their perceived comprehension via seven questions, rated on a 7-point Likert scale, a process we describe as "subjective understanding".

The concrete procedure of our user study is as follows:

1. Participants complete a demographic survey, such as their experience with AI models.
2. Participants complete the warmup task. By doing this, participants adapt their reasoning to the simulated model, for which the examples on the following page are selected.
3. Participants complete the experimental task. They are asked to the model's classification for 15 images.
4. Participants complete a questionnaire to rate their subjective understanding of model explanations.
5. Repeat Steps 2-4 on another dataset.

Prior to commencing Step 3, which is the experimental phase, participants are prompted to select their preferred task. The options presented are: "I will select the label I believe is accurate for the image" and "I will select the label I presume the model will predict". This single-choice question also serves as an attention check. By doing this, we can control whether all participants fully understand the task. All participants in our user study made the correct choice, i.e., "choose the label that the model would predict".

Objective Understanding Questions. Figure 6.6 presents a sample question from our user study, aimed at assessing objective comprehension (simulatability). The study comprises 15 questions, broadly distributed across four distinct categories. On the left, the

Figure 6.6: Question on objective understanding: participants are asked to predict the model's prediction given selected model explanations.

test image is displayed, while the right side features model explanations, specifically the top 20 as ranked by the chosen selection strategy. For both the control and experimental groups, different algorithms select the examples on the right, yet the test images on the left remain identical for both groups.

Subjective Understanding Questions The question utilized for gauging subjective comprehension is adapted from references [280, 281]. Responses are recorded using a 7-point Likert scale, where 1 signifies "Strongly Disagree" and 7 indicates "Strongly Agree2.

- I understood the explanations within the context of this study.
- The explanations provided enough information for me to understand how the Machine Learning model arrived at its label. (Alternative: I would need more information to understand the explanations.)

- I think that most people would learn to understand the explanations very quickly.
- I would like to have more examples to understand the machine's reasoning and how the machine arrived at its labeling.
- The explanations were useful and helped me understand the machine's reasoning.
- I believe that I could provide an explanation similar to the machine's explanation for a new image.

Participants. We recruit $N = 100$ individuals (mean age 28.8 ± 8.6, comprising 49 females, 50 males, and 1 person of undefined gender) via the research platform Prolific[3], and systematically distributed them across two experimental conditions (each with 50 participants). Among these participants, 51 had previous AI exposure through usage of Alexa, Siri, ChatGPT, or ML-related academic courses. All participants successfully completed an attention check during the study. The research protocol received approval from the Technical University of Munich Institutional Review Board (IRB). At the start of the experimental session, informed consent was obtained from each participant through Prolific. For their involvement in this half-hour study, participants received a compensation of £4.50.

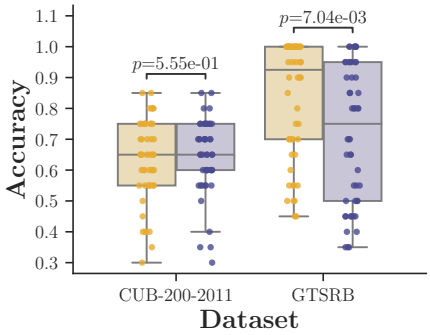
6.5.2 Results

Analysis on R1. The outcomes of the accuracy of simulatability under each condition for each dataset are presented in Figure 6.7(a). Regarding GTSRB, a notable enhancement in user simulatability accuracy of 11.5% ($p = 0.007$) is noted when applying our framework. In contrast, on the CUB dataset, the user prediction accuracy in two conditions is comparable, showing no substantial impact.

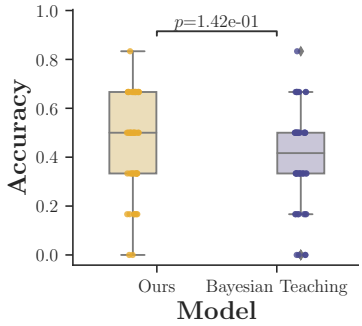
Upon examining the test samples where the target model's predictions are incorrect (with 6 out of 15 images being misclassified), our approach shows enhanced effectiveness compared to BT. In the experimental condition, users attained an accuracy of 46.3%, contrasting with 40.3% in the control condition, as depicted in Figure 6.7(b). This suggests that users are more adept at predicting the target model's errors using our method, which represents a more complex task. Further proof of our model's efficacy is presented in Figure 6.8. We analyzed the number of words users use to identify the model's distinguishing features across four classes. With our method, there is a consensus among users about specific features (bird body parts) for each class. For instance, approximately 68% of users chose "Head" for differentiating Rhinoceros, while about 20% preferred "Belly" for Least Auklet. In contrast, agreement among users in Bayesian Teaching is less common; for instance, only about 10% agreed on "Body" for Least Auklet, with others providing varied descriptions. These findings underscore our method's effectiveness in enhancing user comprehension of the target model.

As illustrated in Figure 6.7(c), the enhancement in subjective comprehension (as measured by rating scores) shows no substantial difference on CUB, where the average rating

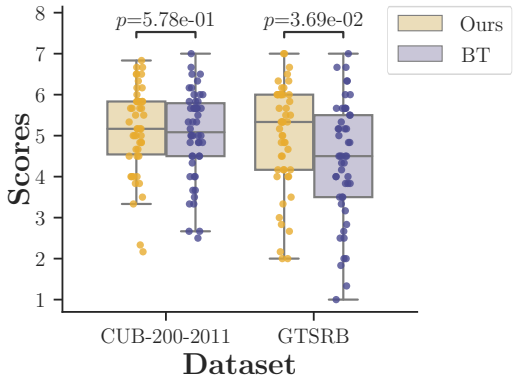
[3]https://www.prolific.co/



(a) User Accuracy (all predictions)



(b) User Accuracy (inaccurate pred., CUB)



(c) User evaluation score.

Figure 6.7: Results of experiments with human users ($N = 100$) comparing I-CEE with the baseline Bayesian Teaching (BT). (a) Simulatability accuracy on all predictions, (b) Simulatability accuracy on images where the target model made inaccurate predictions in the CUB-200-2011 dataset, (c) User's subjective perception of model explanations.

score is 5.14 for our approach and 5.02 for BT. However, on GTSRB, our method outperforms BT with a significant $p = 0.037$. This considerable advancement in GTSRB can be attributed to our method's selection of explanations that impart new insights for differentiating four classes. Conversely, Bayesian Teaching opts for examples that only underscore key features for two classes, thereby limiting users' ability to grasp the model's decision-making process for the remaining classes.

Analysis on R2. The data demonstrates that the task domain (dataset) influences the users' objective comprehension. However, the impact on subjective understanding varies less across different tasks, such as no notable distinction between two datasets when applying our method, as depicted in Figure 6.7(c). Following the user study, participants provided feedback on the comparative usefulness of model explanations

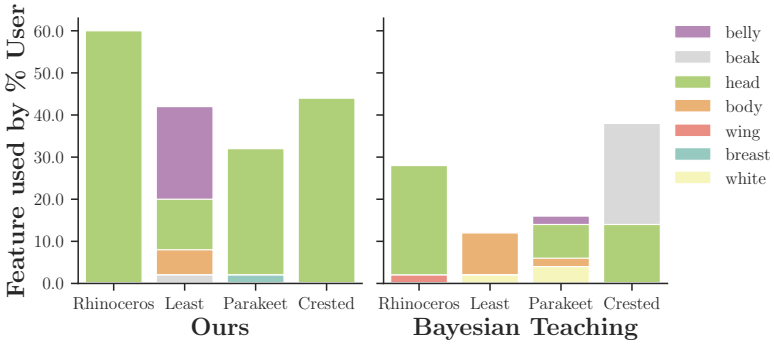


Figure 6.8: Illustration of features used by human users for distinguishing each class on CUB-200-2011.

across two datasets. Most users in both groups found the explanations beneficial, but in the experimental group, seven users, and in the control group, fourteen users, reported that explanations about bird species were more helpful than those about road signs. The ambiguity in the road sign images might be due to the salient area always being a circle covering the road sign, which appears to be "the only one distinguishing feature" among various classes.

6.6 Discussion

Limitations and Future Work. Future research on our framework may explore these directions. First, it is important to delve into more sophisticated models for assessing expertise. Our current methodology simulates user expertise by using a concept-based reasoning strategy for categorizing images, as outlined in [268]. An alternative method could be the application of Large Language Models to mimic the input of multiple individuals in a textual format, as discussed in [282, 283]. Moreover, our existing framework does not address the sample complexity related to estimating user expertise. Therefore, forthcoming studies should focus on techniques that can gauge user expertise efficiently using a limited amount of annotations from actual users. Future research should assess datasets that incorporate a broader range of instances, as recommended by certain participants.

Implications for XAI Systems. We contend that incorporating user modeling is crucial for delivering explanations that specifically address user-specific misunderstandings or confusion. Future XAI systems should harness and address the unique preferences and sources of confusion of individual users. This necessitates the creation of human-in-the-loop systems, enabling users to actively engage in the explanation generation process.

6.7 Conclusion

This chapter introduced a human-centric XAI framework, I-CEE, designed to provide tailored explanations for image classification ML models based on user expertise. Our framework identified task-relevant concepts in image classification and utilized these concepts to create user models that consider human expertise. Then, it selected examples and explanations to assist users in grasping the missing concepts necessary for accurate predictions of model decisions. We assessed our approach through simulated experiments on four datasets and presented findings from a comprehensive human-subject study ($N = 100$). In these experiments, it was observed that I-CEE surpassed previous methods, highlighting the potential of human-centered XAI and suggesting avenues for future research in XAI system design.

7 Conclusions and Future Work

In summary, this cumulative dissertation addresses current challenges existing in HAI for the development of advanced AI models that aim to better assist humans. More specifically, it proposes novel methods for the incorporation of three essential human factors—perception, response, and reasoning—into AI models. These factors are linked to various stages of model decision-making processes. Specifically, the dissertation explores the integration of **human attention** into models to enhance perception. This approach improves the model's ability to process inputs effectively, which is shown for an example application on the context of medical diagnosis. Additionally, it introduces a model designed to predict **human intentions**, focusing on the realm of advanced autonomous driving. Lastly, the dissertation highlights the importance of improving **human comprehension** of these opaque models by providing post-decision explanations. This part first presents guidelines for the design of user studies aimed at assessing the effectiveness of XAI in facilitating end users' comprehension and trust in AI models. Beyond including human participants into the evaluation of XAI, a novel framework is introduced, offering personalized model explanations. Through this framework, individuals can gain improved insights into black-box models, as the explanations are tailored to their specific reasoning needs.

This dissertation further introduces innovative methodologies for designing models, providing thus novel insights into the realization of HAI. It marks a substantial progression towards the principles of HAI models: the development of AI systems that not only technically reflect human intelligence, but also emphasize the impact of AI on humans, and enhance human abilities rather than replacing them.

7.1 Conclusions

This dissertation highlights the efficacy of the proposed methods in integrating human factors, specifically, human attention, intention, and comprehension, into the design of AI models.

Part I addresses the power of human visual perception, i.e., human attention, in the context of challenging classification tasks. This part also explores how to integrate human attention into model design. Since gaze data can serve as a proxy for human attention, **eye-tracking** tools were employed to explore the intricacies of human attention. An image comparison task was used to guide participants to focus on discriminative features while comparing two visually similar images from distinct fine-grained classes. Remarkably, by training a model using only 5% of an image—the areas where human attention was focused—and masking out the remainder, the model's performance matched

that achieved based on the full image. This highlights the effectiveness of human attention in classification tasks. To harness human attention, two strategies were introduced in this dissertation: 1) data augmentation using these attention regions, and 2) integration of human-derived saliency features into a vanilla network using a dual-branch network. The proposed knowledge fusion techniques were evaluated on two tasks: a fine-grained classification of bird species and a *disease diagnosis* task using chest X-ray images. The findings show that integrating human gaze-based attention efficiently improves the model performance.

To harvest an efficient human-AI collaboration, the model understanding humans' intentions is essential. Part II targets at proposing advanced AI models for human intention predictions. This part opts for the utilization of advanced autonomous driving as a use case, where there is a rising demand for an AI model to function as a "co-pilot". More specifically, this part first introduced a novel framework for driver maneuver intention prediction. The novelty of this model is that it combines the features from driving scenes, i.e., videos of road traffic, with the features of human drivers' actions. Both features were extracted from monitoring videos. Motion features from outside videos were obtained from optical flow images using ConvLSTM layers, while human action features were extracted by a 3D CNN. The proposed method surpassed the SOTA model significantly in driver maneuver prediction accuracy. Another framework was designed to predict the objects a driver focuses on while driving, which can indicate driver intentions. It offered an understanding of human intentions and the capability to make decisions based on these identified objects. The efficacy of this model was demonstrated through experiments on two public datasets featuring driver gaze. Results show that our model not only achieved state-of-the-art performance in attention prediction, surpassing previous works, but also did so with significantly reduced computational resources. More importantly, this model was integrated with YOLOv5, enabling it to provide detailed object information. This integration was beneficial in complex and crowded traffic scenarios, enhancing hence the model's ability to inform further decision-making processes.

Addressing the third human factor, namely reasoning, within AI models is explored in Part III. In this section, Explainable Artificial Intelligence (XAI) serves as a valuable tool to elucidate the reasoning processes of opaque models. Specifically, Part III initiates the discussion by highlighting the existing gap between XAI algorithms and their utility for end users in XAI applications. To enhance human comprehension of these models, this part focuses on two research challenges elaborated upon in Chapter 5 and Chapter 6, respectively: (1) XAI application design should consider user experiences, and (2) XAI algorithm design should consider user backgrounds.

Chapter 5 examined the limitations in automatic metrics like fidelity. For example, these metrics yielded inconsistent evaluation outcomes as demonstrated in various instances due to the bias term. This chapter also demonstrated the source of the bias term in automatic evaluation metrics, which was defined as "Class Information Leakage". To mitigate the bias, a novel evaluation strategy "Noisy Linear Imputation" was introduced. Experimental results on various datasets show that the proposed method was able to effectively mitigate the influence of the bias caused by conventional imputation methods. This chapter also addressed the importance of human-grounded evaluation in measuring

user perception of XAI methods, which cannot be evaluated through automatic metrics. However, human-grounded evaluation posed other challenges, such as selected examples in user studies containing biases that can easily trick users into trusting models with non-meaningful explanations, as studied in [254]. Therefore, a detailed user-centered evaluation guideline is proposed to help design human-grounded assessments.

Human reasoning should be integrated into XAI algorithms to provide more effective model explanations. This integration aids users in better comprehending the workings of AI models. Chapter 6 proposed a novel XAI framework coined as I-CEE, aiming at estimating user expertise in reasoning image classification tasks ("concept-based thinking") and providing informative explanation examples. Informativeness was calculated based on the Hypercorrection Effect, inspired by educational psychology. This enabled I-CEE to select examples that could better inform users of the reasoning mechanism of the model. Results on simulated users and real human users indicated that I-CEE could estimate individual users' reasoning, and select explanations tailored to their expertise, effectively improving user understanding of the model. This work highlighted the importance of considering human factors in generating model explanations.

7.2 Future Work

While the results in promoting HAI through the methods proposed in this dissertation are promising, there remains room for further development in AI to better support humans. This section first discusses limitations discovered from the current works as follows. (1) Gathering human expertise and knowledge can be very costly, such as human gaze data from radiologists during the examination of X-ray images. To enable future integration with human expert knowledge, an algorithm that is efficient in learning is therefore essential. (2) In the current work, XAI frameworks provide static local explanations and do not support interactions. In cases of large datasets or models with a vast number of features, creating comprehensive and understandable static explanations over all different cases becomes impossible and potentially overwhelming for the user. (3) HAI also aims to emphasize its societal impacts, a topic that goes beyond the scope of this dissertation. In the upcoming section, three prospective research directions aimed at overcoming the existing limitations are explained.

Human knowledge integration via efficient learning algorithms. In future work, it is important to research how learning human knowledge could be enabled from small data samples, especially when considering the costs of collecting human expert data. In contrast to AI models, in fact, humans are good at learning new knowledge with very few samples. For instance, from a very young age, children engage in a form of contrastive learning when they differentiate between categories. The power of human contrastive learning can be observed in Chapter 3, where human subjects are asked to compare two visually similar images from distinct fine-grained bird species. Although none of the participants were ornithologists, they could identify distinct visual features for a given bird species after a few seconds of comparison, whereas models require a large corpus of

data samples for training within the learning scheme. To address this research challenge, in future work, I plan to introduce an active learning scheme [151] in contrastive learning for an efficient learning process, thus harnessing human intelligence in contrasting. Given a pretrained model, we can then use the active contrastive learning framework to further improve the capability of the model for a specific task, which boosts efficient fine-tuning. The model first identifies similar pairs and then asks human annotators for precise labels. Besides labels, comparison cues such as distinct areas in images or important words in sentences should also be considered. The human annotation is further used to retrain the model. For a traditional deep learning model, a contrastive loss or a supervised loss can be used depending on the human-annotated cues.

Human comprehension of complex models via interactions. XAI methods considered in this work provide local explanations. As AI models become deeper and more complex, explaining black-box models based on parameters and several explanation examples using traditional XAI becomes less efficient. To better explain the decision-making processes of these complex models, my future research will focus on designing interactive explanation frameworks that allow dynamic explanations based on user feedback, moving hence beyond conventional post-hoc explanatory approaches.To enable interactivity, interpretable models, for instance the concept bottleneck architecture, can be deployed, which provides explanations of important features for the decision in the network forward pass. Using these frameworks, users can view explanations and modify features that are used in making the final decision. In this approach, users gain insights into the model's functionality through personalized queries. In addition to interactive explanation design, assessing the efficacy of these explanations extends beyond solely human understanding. It is important to evaluate user experience with a specific emphasis on cognitive load. Future work should consider physiological measures such as heart rate along with eye-tracking metrics like pupil dilation.

Enhancing impact on humans and society. Emphasizing the social impact of AI models and ensuring their alignment with human values in addressing critical societal issues is another key objective of HAI. Going beyond individual use cases, it is worth considering applications that can benefit a large public, such as the prediction of urban planning challenges, public health crises, or educational disparities, all of which have the potential to positively impact society at large. To address the social impact of AI, future work considers formulating complex problems in social-impact situations. It involves working closely with experts in the field and a thoughtful assessment of various decision-making options. To solve these problems, novel multi-modal fusion techniques should be studied in developing AI-based solutions. For instance, combining different types of data, such as text and images from social media, with specialized data, like public health, educational, or environmental statistics, can benefit the prediction of AI models. These approaches must take into account the broader societal impact. Furthermore, it is important to focus on the ethical impact, ensuring that these technologies are accessible to diverse communities and do not cause unfairness in the decisions, which involves continuously

monitoring and adjusting AI algorithms to prevent biases that could negatively affect certain groups.

Bibliography

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[2] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 8(1):92, 2021.

[3] A. Jain, S. Soh, B. Raghavan, A. Singh, H. S. Koppula, and A. Saxena. Brain4cars: Sensory-fusion recurrent neural models for driver activity anticipation.

[4] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney. Predicting driver attention in critical situations. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 658–674. Springer, 2019.

[5] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon, Y. Defretin, et al. ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. *Zenodo*, 2021.

[6] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[7] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney. Predicting driver attention in critical situations. In *ACCV*, 2018.

[8] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara. Predicting the driver's focus of attention: the dr(eye)ve project. *TPAMI*, 2018.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *ICPR*, 2016.

[10] N. Liu, J. Han, and M.-H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.

[11] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[12] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, ICML*, 2017.

[13] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[15] A. Pal, S. Mondal, and H. I. Christensen. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *CVPR*, 2020.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[18] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[19] W. Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.

[20] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[21] S. HAI. About hai, 2018. Accessed: 2023-11-12. URL: `https://hai.stanford.edu/about`.

[22] I. E. Association. What is ergonomics (hfe), 2023. Accessed: 2023-11-30. URL: `https://iea.cc/about/what-is-ergonomics/`.

[23] M. Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[24] R. D. Rimey and C. M. Brown. Control of selective perception using bayes nets and decision theory. *IJCV*, 12(2-3):173–207, 1994.

[25] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[26] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*, 2007.

[27] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI*, volume 2020, 2020.

[28] Y. Rong, W. Xu, Z. Akata, and E. Kasneci. Human attention in fine-grained classification. In *British Machine Vision Conference (BMVC 2021)*, 2021.

[29] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, pages 3794–3803, 2021.

[30] K. Saab, S. M. Hooper, N. S. Sohoni, J. Parmar, B. Pogatchnik, S. Wu, J. A. Dunnmon, H. R. Zhang, D. Rubin, and C. Ré. Observational supervision for medical image classification using gaze data. In *MICCAI*, pages 603–614. Springer, 2021.

[31] K. Shanmuga Vadivel, T. Ngo, M. Eckstein, and B. Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*, 2015.

[32] A. B. Vasudevan, D. Dai, and L. Van Gool. Object referring in videos with language and human gaze. In *CVPR*, pages 4129–4138, 2018.

[33] W. Wang, J. Shen, X. Dong, and A. Borji. Salient object detection driven by fixation prediction. In *CVPR*, pages 1711–1720, 2018.

[34] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, 2006.

[35] M. Kümmerer, T. S. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *Journal of Vision*, 2016.

[36] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, volume 28, 2015.

[37] M. Barz and D. Sonntag. Automatic visual attention detection for mobile eye tracking using pre-trained computer vision models and human gaze. *Sensors*, 21(12):4143, 2021.

[38] N. Kumari, V. Ruf, S. Mukhametov, A. Schmidt, J. Kuhn, and S. Küchemann. Mobile eye-tracking data analysis using object detection via yolo v4. *Sensors*, 21(22):7668, 2021.

[39] K. Panetta, Q. Wan, A. Kaszowska, H. A. Taylor, and S. Agaian. Software architecture for automating cognitive science eye-tracking data analysis and object annotation. *IEEE Transactions on Human-Machine Systems*, 49(3):268–277, 2019.

[40] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[41] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–20, 2023.

[42] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6021–6029, 2020.

[43] J. Haug, S. Zürn, P. El-Jiz, and G. Kasneci. On baselines for local feature attributions. *AAAI Workshop on Explainable Agency in AI Workshop*, 2021.

[44] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 2023.

[45] Y. Rong, N.-R. Kassautzki, W. Fuhl, and E. Kasneci. Where and what: Driver attention-based object detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(ETRA):1–22, 2022.

[46] P. Hase and M. Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, 2020.

[47] D. Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, 2018.

[48] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 408–416, 2019.

[49] G. Hoffman. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 2019.

[50] M. Chromik and M. Schuessler. A taxonomy for human subject evaluation of black-box explanations in xai. *Exss-atec@ iui*, 94, 2020.

[51] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[52] Q. Yang, N. Banovic, and J. Zimmerman. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–11, 2018.

[53] C. Baker, R. Saxe, and J. Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33. International Organization for Standardization, 2011.

[54] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2), February 2019.

[55] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir. Exploring computational user models for agent policy summarization. In *IJCAI: proceedings of the conference*, volume 28, page 1401. NIH Public Access, 2019.

[56] P. Qian and V. Unhelkar. Evaluating the role of interactivity on improving transparency in autonomous agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, page 1083–1091. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[57] S. C.-H. Yang, N. E. T. Folke, and P. Shafto. A psychological theory of explainability. In *International conference on machine learning*, pages 25007–25021. PMLR, 2022.

[58] S. C.-H. Yang, W. K. Vong, R. B. Sojitra, T. Folke, and P. Shafto. Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, 11(1):9863, 2021.

[59] P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

[60] M. Beliaev, A. Shih, S. Ermon, D. Sadigh, and R. Pedarsani. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pages 1732–1748. PMLR, 2022.

[61] L. Qiuxia, S. Khan, Y. Nie, S. Hanqiu, J. Shen, and L. Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 2020.

[62] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *TPAMI*, 2014.

[63] P. Majaranta and A. Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 2014.

[64] Y. Xia, J. Kim, J. Canny, K. Zipser, T. Canas-Bajo, and D. Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *WACV*, 2020.

[65] C. Braunagel, W. Rosenstiel, and E. Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *ITSM*, 2017.

[66] D. Weber, T. Santini, A. Zell, and E. Kasneci. Distilling location proposals of unknown objects through gaze information for human-robot interaction. In *IROS*, 2020.

[67] A. Shafti, P. Orlov, and A. A. Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *ICRA*, 2019.

[68] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni. Eye-hand behavior in human-robot shared manipulation. In *HRI*, 2018.

[69] N. Castner, T. C. Kuebler, K. Scheiter, J. Richter, T. Eder, F. Hüttig, C. Keutel, and E. Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ETRA*, 2020.

[70] N. Karessli, Z. Akata, B. Schiele, and A. Bulling. Gaze embeddings for zero-shot image classification. In *CVPR*, July 2017.

[71] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017.

[72] M. Sun, Y. Yuan, F. Zhou, and E. Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018.

[73] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[74] F. Zhang, M. Li, G. Zhai, and Y. Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MMM*. Springer, 2021.

[75] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*, 2017.

[76] Z. Li, Y. Yang, X. Liu, F. Zhou, S. Wen, and W. Xu. Dynamic computational time for visual attention. In *ICCV*, 2017.

[77] P. Sermanet, A. Frome, and E. Real. Attention for fine-grained categorization. In *ICLRW*, 2015.

[78] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017.

[79] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, 2019.

[80] P. Zhuang, Y. Wang, and Y. Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, 2020.

[81] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *CVPR*, 2020.

[82] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NeuIPs*, 2014.

[83] Q. Lai, S. Khan, Y. Nie, H. Sun, J. Shen, and L. Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020.

[84] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata. Attribute prototype network for zero-shot learning. *arXiv preprint arXiv:2008.08290*, 2020.

[85] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[86] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *KDD*, 2016.

[87] S. Wang, T. Zhou, and J. Bilmes. Bias also matters: Bias attribution for deep neural network explanation. In *ICML*, 2019.

[88] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[89] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[90] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

[91] C.-K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar. Representer point selection for explaining deep neural networks. In *NeuIPs*, 2018.

[92] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

[93] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. In *NeuIPs*, 2019.

[94] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. In *NeurIPs*, 2019.

[95] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[96] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[97] C. Sen, T. Hartvigsen, B. Yin, X. Kong, and E. Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *ACL*, 2020.

[98] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical co-attention for visual question answering, 2016.

[99] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[100] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *ECCV*, 2016.

[101] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3182–3190, 2015.

[102] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[103] D. Zhou, H. Ma, and Y. Dong. Driving maneuvers prediction based on cognition-driven and data-driven method. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2018.

[104] P. Gebert, A. Roitberg, M. Haurilet, and R. Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent vehicles symposium (IV)*, pages 969–974. IEEE, 2019.

[105] M. Tonutti, E. Ruffaldi, A. Cattaneo, and C. A. Avizzano. Robust and subject-independent driving manoeuvre anticipation through domain-adversarial recurrent neural networks. *Robotics and Autonomous Systems*, 115:162–173, 2019.

[106] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1652–1657. IEEE, 2015.

[107] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci. Online recognition of driver-activity based on visual scanpath classification. *IEEE Intelligent Transportation Systems Magazine*, 9(4):23–36, 2017.

[108] J. Wolf, S. Hess, D. Bachmann, Q. Lohmeyer, and M. Meboldt. Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research*, 11(6), 2018.

[109] P. Kumar, M. Perrollaz, S. Lefevre, and C. Laugier. Learning-based approach for online lane change intention prediction. In *IV*, pages 797–802. IEEE, 2013.

[110] E. M. S. Machado, I. Carrillo, M. Collado, and L. Chen. Visual attention-based object detection in cluttered environments. In *SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, pages 133–139. IEEE, 2019.

[111] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[112] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[113] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCVW*, volume 1, pages 1–2. Prague, 2004.

[114] M. Barz, S. Kapp, J. Kuhn, and D. Sonntag. Automatic recognition and augmentation of attended objects in real-time using eye tracking and a head-mounted display. In *ACM ETRA*, pages 1–4, 2021.

[115] C. Liu, Y. Chen, L. Tai, H. Ye, M. Liu, and B. E. Shi. A gaze model improves autonomous driving. In *ACM ETRA*, pages 1–5, 2019.

[116] A. Makrigiorgos, A. Shafti, A. Harston, J. Gerard, and A. A. Faisal. Human visual attention prediction boosts learning & performance of autonomous driving agents. *arXiv preprint arXiv:1909.05003*, 2019.

[117] l. Kai, H. Sheng, Z. Xiong, W. Li, and L. Zheng. Improving driver gaze prediction with reinforced attention. *IEEE Transactions on Multimedia*, 2020.

[118] M. Shirpour, S. S. Beauchemin, and M. A. Bauer. Driver's eye fixation prediction by deep neural network. In *VISIGRAPP*, 2021.

[119] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012.

[120] E. Aksoy, A. Yazıcı, and M. Kasap. See, attend and brake: An attention-based saliency map prediction model for end-to-end driving. *arXiv preprint arXiv:2002.11020*, 2020.

[121] T. Deng, H. Yan, L. Qin, T. Ngo, and B. Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *T-ITS*, 2019.

[122] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[123] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[124] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32:10967–10978, 2019.

[125] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[126] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[127] P. Sturmfels, S. Lundberg, and S.-I. Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[128] S. H. Park and K. Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, 2018.

[129] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19(1):1–18, 2019.

[130] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem. Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339, 2020.

[131] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel. Artificial neural networks in medical diagnosis, 2013.

[132] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505, 2017.

[133] X. Dastile, T. Celik, and M. Potsane. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, 2020.

[134] M. Ala'raj, M. F. Abbod, M. Majdalawieh, and L. Jum'a. A deep learning model for behavioural credit scoring in banks. *Neural Computing and Applications*, 34(8):5839–5866, 2022.

[135] P. M. Addo, D. Guegan, and B. Hassani. Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38, 2018.

[136] Y. Xia, C. Liu, Y. Li, and N. Liu. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*, 78:225–241, 2017.

[137] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93:106384, 2020.

[138] N. Van Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, and V. Kostakos. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

[139] T. Sourdin. Judge v robot?: Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal, The*, 41(4):1114–1133, 2018.

[140] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[141] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[142] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[143] P. Tambe, P. Cappelli, and V. Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42, 2019.

[144] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December*, 7, 2018.

[145] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu. Employee turnover prediction with machine learning: A reliable approach. In *IntelliSys*, pages 737–758. Springer, 2018.

[146] D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[147] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[148] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1369–1385. Association for Computing Machinery, 2023.

[149] J. J. Ferreira and M. S. Monteiro. What are people doing about xai user experience? a survey on ai explainability research and practice. In *International Conference on Human-Computer Interaction*, pages 56–73. Springer, 2020.

[150] I. Lage and F. Doshi-Velez. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*, 2020.

[151] B. Settles. Active learning literature survey. 2009.

[152] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[153] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

[154] S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

[155] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[156] M. I. Posner and S. E. Petersen. The attention system of the human brain. *Annual review of neuroscience*, 1990.

[157] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[158] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.

[159] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *NeuIPs*, 2019.

[160] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[161] A. Dubey, O. Gupta, R. Raskar, and N. Naik. Maximum-entropy fine grained classification. In *NeuIPs*, 2018.

[162] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 2015.

[163] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017.

[164] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[165] A. Kanehira and T. Harada. Learning to explain with complemental examples. In *CVPR*, 2019.

[166] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In *ECCV*, 2018.

[167] A. Olsen. The tobii i-vt fixation filter. *Tobii Technology*, 2012.

[168] J. N. Vickers. *Perception, cognition, and decision training: The quiet eye in action.* Human Kinetics, 2007.

[169] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[170] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2018.

[171] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[172] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[173] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*. PMLR, 2019.

[174] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[175] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[176] S. Huang, X. Wang, and D. Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *AAAI*, 2021.

[177] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019.

[178] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, and S.-N. Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019.

[179] Y. Rong, Z. Akata, and E. Kasneci. Driver intention anticipation based on in-cabin and driving scene monitoring. In *ITSC*, pages 1–8. IEEE, 2020.

[180] Y. Rong, C. Han, C. Hellert, A. Loyal, and E. Kasneci. Artificial intelligence methods in in-cabin use cases: a survey. *IEEE Intelligent Transportation Systems Magazine*, 14(3):132–145, 2021.

[181] L. Pomarjanschi, M. Dorr, and E. Barth. Gaze guidance reduces the number of collisions with pedestrians in a driving simulator. *ACM TiiS*, 1(2):1–14, 2012.

[182] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual explanations for self-driving vehicles. In *ECCV*, pages 563–578, 2018.

[183] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[184] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015.

[185] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[186] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[187] G. Jocher, A. Stoken, J. Borovec, NanoCode012, A. Chaurasia, TaoXie, L. Changyu, A. V, Laughing, tkianai, yxNONG, A. Hogan, lorenzomammana, AlexWang1900, J. Hajek, L. Diaconu, Marc, Y. Kwon, oleg, wanghaoyang0106, Y. Defretin, A. Lohia, ml5ah, B. Milanko, B. Fineran, D. Khromov, D. Yiwei, Doug, Durgesh, and F. Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, 2021. URL: `https://doi.org/10.5281/zenodo.4679653`, `doi:10.5281/zenodo.4679653`.

[188] J. Choi, D. Chun, H. Kim, and H.-J. Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *ICCV*, pages 502–511, 2019.

[189] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking objects as points. In *ECCV*, 2020.

[190] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *CVPRW*, pages 390–391, 2020.

[191] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018.

[192] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara. Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops*, pages 54–60, 2016.

[193] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[194] A. L. Yarbus. Eye movements during perception of complex objects. In *Eye Movements and Vision*, pages 171–211. Springer, 1967.

[195] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[196] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[197] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *ICIP*, volume 1, pages I–253. IEEE, 2003.

[198] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022.

[199] Y. Rong, P. Qian, V. Unhelkar, and E. Kasneci. I-cee: Tailoring explanations of image classifications models to user expertise. *arXiv preprint arXiv:2312.12102*, 2023.

[200] M. O. Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[201] U. Ehsan and M. O. Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[202] J. R. Vergara and P. A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 2014.

[203] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[204] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

[205] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

[206] L. Bossard, M. Guillaumin, and L. V. Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[207] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[208] M. Kachuee, K. Karkkainen, O. Goldstein, S. Darabi, and M. Sarrafzadeh. Generative imputation and stochastic prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[209] N. Bevan. International standards for hci and usability. *International journal of human-computer studies*, 55(4):533–552, 2001.

[210] I. S. . E. of Human-System Interaction (Subcommittee). *ISO 9241-11:1998: Guidance on Usability*. 1998.

[211] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[212] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[213] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[214] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[215] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 397–407, 2019.

[216] A. I. Anik and A. Bunt. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[217] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

[218] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[219] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces*, pages 537–548, 2022.

[220] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[221] G. Hoffman and X. Zhao. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–31, 2020.

[222] D. Das and S. Chernova. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 510–518, 2020.

[223] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, 2019.

[224] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, and N. T. Vu. Human interpretation of saliency-based explanation over text. *arXiv preprint arXiv:2201.11569*, 2022.

[225] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

[226] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf. Do users benefit from interpretable vision? a user study, baseline, and dataset. In *International Conference on Learning Representations*, 2022.

[227] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5277–5285, 2022.

[228] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings*

of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2021.

[229] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[230] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh. Do explanations make vqa models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036—-1042, 2018.

[231] J. Ooge, S. Kato, and K. Verbert. Explaining recommendations in e-learning: Effects on adolescents' trust. In *27th International Conference on Intelligent User Interfaces*, pages 93–105, 2022.

[232] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer. I can do better than your ai: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251, 2019.

[233] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th International Conference on Intelligent User Interfaces*, pages 307–317, 2021.

[234] X. Wang and M. Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[235] J. Eccles. Expectancies, values and academic behaviors. *Achievement and achievement motives*, 1983.

[236] C. S. Hulleman, J. J. Kosovich, K. E. Barron, and D. B. Daniel. Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, 109(3):387, 2017.

[237] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[238] F. G. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology*, 84(4):429, 1992.

[239] K. Ouwehand, A. v. d. Kroef, J. Wong, and F. Paas. Measuring cognitive load: Are there more valid alternatives to likert rating scales? In *Frontiers in Education*, volume 6, page 702616. Frontiers Media SA, 2021.

[240] J. P. Simmons, L. D. Nelson, and U. Simonsohn. Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1):151–162, 2021.

[241] U. Simonsohn, L. D. Nelson, and J. P. Simmons. P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, 143(2):534, 2014.

[242] K. A. Ericsson and H. A. Simon. *Protocol analysis: Verbal reports as data.* MIT Press, 1984.

[243] W. Zhang and B. Y. Lim. Towards relatable explainable ai with the perceptual process. In *CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.

[244] Y. Wang, P. Venkatesh, and B. Y. Lim. Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation. In *CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[245] J. Cohen. *Statistical power analysis for the behavioral sciences.* Routledge, 2013.

[246] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li. Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle. In *Designing Interactive Systems Conference 2021*, pages 1591–1602, 2021.

[247] F. Y. Kung, N. Kwok, and D. J. Brown. Are attention check questions a threat to scale validity? *Applied Psychology*, 67(2):264–283, 2018.

[248] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi. Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[249] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[250] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[251] D. Alvarez Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[252] A. Papenmeier, G. Englebienne, and C. Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

[253] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[254] T. Leemann, Y. Rong, T.-T. Nguyen, E. Kasneci, and G. Kasneci. Caution to the exemplars: On the intriguing effects of example choice on human trust in xai. In *XAI in Action: Past, Present, and Future Applications*, 2023.

[255] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[256] Y. Rong, N. Castner, E. Bozkir, and E. Kasneci. User trust on an explainable ai-based medical diagnosis support system. *arXiv preprint arXiv:2204.12230*, 2022.

[257] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 392–402, 2020.

[258] J. Schoeffer, N. Kuehl, and Y. Machowski. "there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. *arXiv preprint arXiv:2205.05758*, 2022.

[259] E. Rader, K. Cotter, and J. Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[260] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912, 2018.

[261] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.

[262] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.

[263] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[264] A. Springer and S. Whittaker. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *IUI*, pages 107–120, 2019.

[265] V. Chen, N. Johnson, N. Topin, G. Plumb, and A. Talwalkar. Use-case-grounded simulations for explanation evaluation. *arXiv preprint arXiv:2206.02256*, 2022.

[266] G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022.

[267] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[268] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

[269] S. L. Armstrong, L. R. Gleitman, and H. Gleitman. What some concepts might not be. *Cognition*, 13(3):263–308, 1983.

[270] J. B. Tenenbaum. *A Bayesian framework for concept learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

[271] J. Metcalfe. Learning from errors. *Annual Review of Psychology*, 68(1):465–489, 2017.

[272] B. Butterfield and J. Metcalfe. Errors committed with high confidence are hypercorrected. *Journal of experimental psychology. Learning, memory, and cognition*, 27:1491–4, 12 2001. doi:10.1037/0278-7393.27.6.1491.

[273] J. Metcalfe and B. Finn. People's hypercorrection of high-confidence errors: Did they know it all along? *Journal of experimental psychology. Learning, memory, and cognition*, 37:437–48, 03 2011.

[274] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[275] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

[276] S. Ioffe. Probabilistic linear discriminant analysis. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, pages 531–542. Springer, 2006.

[277] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

[278] B. Settles, M. Craven, and L. Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:, 2008.

[279] P. Hase, S. Zhang, H. Xie, and M. Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*, 2020.

[280] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023.

[281] Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483*, 2021.

[282] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[283] G. V. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[284] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[285] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015.

[286] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

[287] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989.

[288] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[289] S. A. Cholewiak, P. Ipeirotis, V. Silva, and A. Kannawadi. SCHOLARLY: Simple access to Google Scholar authors and citation using Python, 2021. URL: `https://github.com/scholarly-python-package/scholarly`, `doi:10.5281/zenodo.5764801`.

[290] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[291] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[292] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[293] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[294] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[295] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

[296] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10, 2012.

[297] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

[298] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[299] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[300] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[301] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[302] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[303] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

[304] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[305] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[306] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[307] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[308] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.

[309] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5686–5697, 2016.

[310] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[311] V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

[312] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[313] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

[314] B. Y. Lim, A. K. Dey, and D. Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2119–2128, 2009.

[315] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[316] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[317] C. J. Cai, J. Jongejan, and J. Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262, 2019.

[318] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[319] R. F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2390–2395, 2016.

[320] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18(5):455–496, 2008.

[321] M. Owens and K. Tanner. Teaching as brain changing: Exploring connections between neuroscience and innovative teaching. *Cell Biology Education*, 16:fe2, 07 2017. doi:10.1187/cbe.17-01-0005.

[322] S. A. Ambrose, M. DiPietro, M. W. Bridges, M. K. Norman, and M. C. Lovett. *How Does Students' Prior Knowledge Affect Their Learning*, chapter 1, pages 10–39. John Wiley & Sons, 2010.

A Human Attention in Fine-grained Classification

This chapter provides additional results for the work discussed in Chapter 3. The content is adapted from the work that was published in BMVC 2021 [28].

A.1 Gaze Data Analysis

This section demonstrates the distinctiveness of feature attributes identified through our analysis of human gaze data for detailed species classification. The CUB dataset provides inherent attributes for each picture as 312-dimensional binary vectors. These vectors are utilized to pinpoint the most distinctive attributes for each bird category within the dataset. Our data collection experiments encompass 100 comparison pairings. In these experiments, every image from one class is compared against every image from another class. For example, if there are M images in the first class and N images in the second, the total number of possible pairings between the two classes is $M \cdot N$. For each pairing, we generate a comparison attribute vector. In this vector, a value of 1 indicates identical attributes in both images, while 0 denotes differing attributes. Thus, like the ground-truth vectors, these comparison vectors are also 312-dimensional binary arrays. We aggregate these $M \cdot N$ vectors into a singular 312-dimensional vector that encapsulates the distinctive attributes of the two compared classes. For instance, if the entry for the attribute `has-wing-color::brown` in this aggregated vector is 354, it suggests that in 354 of the image pairs, the `has-wing-color::brown` attribute varies. Finally, we categorize these attributes into seven anatomical sections: head, beak, breast, belly, back, wing, and leg. By summing the values of attributes corresponding to each body part in the comparison vector, we can assess the variation in that part between the two classes. The body part with the highest total is identified as the most distinctive feature differentiating the two classes.

In our study, when participants viewed an image, their attention was consistently drawn to specific parts of a bird's body that are distinctive. The area where a viewer's gaze lands typically exhibits a high variety of differing characteristics when comparing two distinct bird species. By utilizing the coordinates of each body part's center in an image, we are able to link each participant's eye fixation (gathered from five individuals for that image) to the closest body part. This linkage is based on the proximity between the fixation point and the body part's central coordinate. As depicted in Figure A.1, the histogram outlines the frequency of particular bird body parts capturing attention across the entire CUB-GHA dataset. Our observations reveal that in 3855 images, humans generally focus on three specific body parts. A significant majority of the images (92.5%)

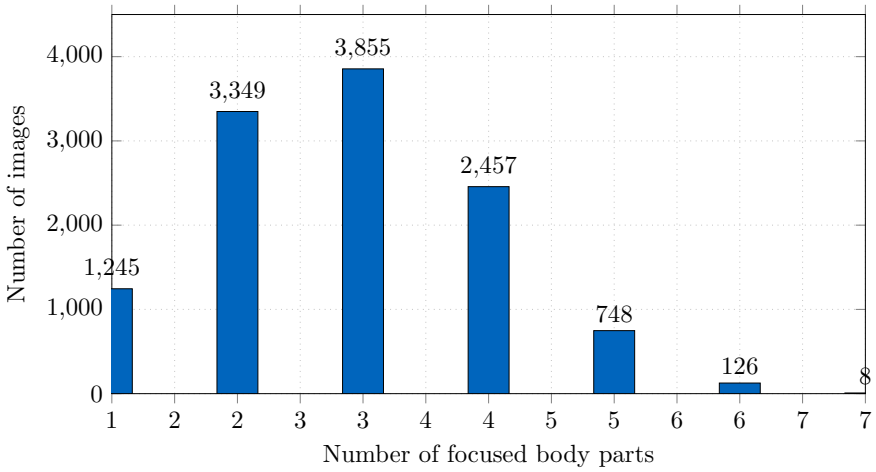


Figure A.1: Histogram of the number of focused bird body parts in CUB-GHA. **Y-axis** refers to the amount of images with the certain number of parts (**X-axis**).

show attention to fewer than five parts. Only a small number of images display attention to all seven bird body parts. For each image, we aggregate the total duration of fixations on each body part, which serves as a measure of the amount of attention it receives from the participants. A higher cumulative duration signifies greater attention. We then order the seven body parts in each image based on these duration totals and assess how frequently the top-k most focused body parts correspond to the most distinctive one, as determined by pre-established ground-truth attributes. This correspondence rate is detailed in Table A.1.

Our findings indicate that in 84.4% of the images, participants accurately identified the most distinctive body part. Moreover, in instances where participants deemed up to four parts as essential for classification, the ground-truth distinct body part was identified in 98.3% of the images. These results suggest that human gaze data in the CUB-GHA dataset effectively points to the discriminative body parts or attributes important for classifying bird species.

Top-k	1	2	3	4
Hit rate (%)	84.40	93.60	97.18	98.31

Table A.1: Hit rate of the most discriminative body part. Top-k refers to the k longest focused body parts by humans in CUB-GHA.

A.2 Additional Comparison between ME and HA

We conduct a quantitative analysis to compare the similarity between HA and MEs. This evaluation uses various metrics: Kullback-Leibler divergence (KL-D), correlation coefficient (CC), and similarity (SIM), commonly applied in image similarity assessments [170]; rank-correlation (Rank-Co) as presented in [96]; the shuffled AUC metric (sAUC) for evaluating each pixel in saliency maps as part of a classification task; and information gain (IG), which measures performance relative to a baseline [170, 61]. CAM and Grad-CAM exhibit close similarities, for instance, Grad-CAM scores 0.565 on CC and 1.242 on KL-D, whereas CAM shows 0.563 and 1.248 respectively. Furthermore, IG and IxG display comparable results on these metrics, with IG scoring 0.699 versus 0.694 for IxG on CC, and 1.318 for IG against 1.310 for IxG on KL-D. These parallels are also evident in the qualitative data. Across all metrics, Grad-CAM appears most similar to HA, achieving the highest ratings in all six metrics. This aligns with findings from the KAR, indicating that among all MEs, Grad-CAM has superior performance.

	KL-D ↓	CC ↑	SIM ↑	Rank-Co ↑	sAUC ↑	IG ↑
CAM	1.248	0.563	0.399	**0.761**	0.460	0.938
Grad-CAM	**1.242**	**0.565**	**0.415**	**0.761**	**0.508**	**1.376**
IG	1.318	0.546	0.361	0.699	0.436	0.921
IxG	1.310	0.543	0.375	0.694	0.461	1.001

Table A.2: Similarity comparison between MEs and HA saliency map. (↓: the lower the better; ↑: the higher the better.)

B Driver Intention Anticipation

This chapter provides additional related work and results for Section 4.2. The content is adapted from the work that was published in ITSC 2020 [28].

B.1 Related Work

The intention to maneuver can be identified through drivers' actions, like glancing at external mirrors or peering through windows. Consequently, methodologies from human action recognition have been effectively utilized in this context. An action embodies both spatial and temporal elements. It is a common understanding that deep CNNs excel in capturing spatial domain features, whereas RNN frameworks and LSTM units are renowned for their proficiency in decoding temporal series patterns.

LSTM and RNN techniques are therefore often combined with 2D CNNs in video processing applications to deal with both spatial and temporal information, for example as in [36]. The formulation from [36] is shown in Eq. B.1 with a minor modification since it contains no bias component.

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \cdot c_{t-1}) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \cdot c_{t-1}) \\
g_t &= tanh(W_{xc} * x_t + W_{hc} * h_{t-1}) \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \cdot c_t) \\
h_t &= o_t \cdot \tanh(c_t)
\end{aligned}
\tag{B.1}
$$

In Equation B.1, the subscript t denotes the time step in the sequence. The term x_t represents the input at time t. The symbols i_t, g_t, f_t, and o_t signify the gates within the LSTM cell. The variable c_t stands for the cell state, and h_t indicates the hidden state at time t. The various W symbols correspond to the weight matrices involved in convolution operations. The symbol $*$ is used to denote convolution, whereas $\cdot$ signifies element-wise multiplication. The functions σ and tanh represent the sigmoid and hyperbolic tangent functions, respectively, and are applied on an element-wise basis. The ConvLSTM, as detailed by the authors in [36], is adept at learning features useful for either regression or classification tasks. An application includes the development of an encoding-forecasting architecture, utilizing ConvLSTM cells to predict subsequent frames in a sequence.

One essential element of video analysis is motion features. Motion describes changes in both temporal and spatial spaces and is often estimated on an image plane based on the optical flow, a method with several decades of research history. This method has been

studied since Horn and Schunck's work in 1981 [284]. Optical flow involves calculating the movement of individual pixels between successive frames, aiding in understanding object motion. Its applications are broad, notably in automotive technology, as highlighted by Menze and Geiger [285], due to its additional feature provision. Historically, optical flow extraction was seen as an optimization challenge, tackled by various methodologies like the energy-based approach [286] and region-based matching [287]. However, the advent of deep learning has revolutionized this field. CNNs have delivered remarkable outcomes. FlowNet [288] and FlowNet 2.0 [6] are two exemplary end-to-end network models in this domain. These networks process a pair of consecutive frames to directly compute the optical flow.

B.2 Additional Experimental Results

We implement a testing procedure using a threshold policy analogous to the one described in [101, 3], applied to our two-stream video model. The model predicts *"go straight"* if the calculated probability does not exceed the threshold. Figure B.1 illustrates that the effectiveness diminishes when the threshold exceeds 0.4 across various video input lengths. This is attributed to the model being developed with a balanced loss function, enabling it to discern motion characteristics associated with all five maneuvers. Typically, the model issues predictions with probabilities above 0.4, indicating a high level of confidence. Consequently, our model operates efficiently without the need for a threshold policy.

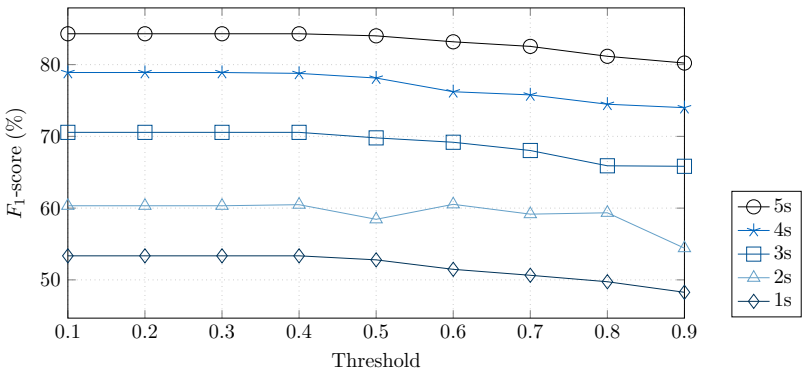


Figure B.1: Effect of using thresholds. Two-stream input with different video lengths (from 1 to 5 seconds).

Additionally, Fig. B.2 shows the confusion matrix of three models using different data sources. Prediction is made based on time period [-5,0]. From this, an improvement in all classes can be observed when using two video streams.

Straight	.80	.07	.02	.07	.04
L Lane	.12	.71	.04	.08	.05
L Turn	.02	.09	.77	.02	.10
R Lane	.07	.07		.80	.06
R Turn	.04	.13	.04	.07	.72
	Straight	L Lane	L Turn	R Lane	R Turn

(a) Inside videos

Straight	.64	.19		.17	
L Lane	.31	.55		.12	.02
L Turn	.02	.05	.88	.03	.02
R Lane	.35	.23	.02	.37	.03
R Turn		.04	.02	.05	.89
	Straight	L Lane	L Turn	R Lane	R Turn

(b) Outside videos

Straight	.87	.05	.01	.05	.02
L Lane	.12	.75	.03	.07	.03
L Turn		.05	.88	.04	.03
R Lane	.08	.06	.01	.80	.05
R Turn	.02	.02		.02	.94
	Straight	L Lane	L Turn	R Lane	R Turn

(c) In and outside videos

Figure B.2: The confusion matrix of using different video streams. The prediction is made at the last second before the occurrence of a maneuver.

C Driver Attention-based Object Detection

This chapter provides additional results for Section 4.3. The content is adapted from the work that was published in PACMHCI 2022 [45].

C.1 Results of Our YOLOv3- and CenterTrack-based Models

For a fair comparison, we evaluated object-level metrics using the objects detected by YOLOv5 across all models as detailed in Section 4.3. Additionally, in table C.1, the object-level outcomes for our models based on YOLOv3 and CenterTrack, which utilize 16×16 grids, are presented, highlighting their performance based on the objects they detected.

	AUC	**Prec (%)**	**Recall (%)**	**F_1 (%)**	**Acc (%)**
CenterTrack	0.83	69.80	74.62	72.13	75.33
YOLOv3	0.84	70.23	73.42	71.79	76.22

Table C.1: Comparison of different models on BDD-A dataset with own detected objects (Th = 0.5). For all metrics a higher value indicates better performance.

C.2 Results of Different Input Sequence Lengths of LSTM

In Table C.2, the results for varying input sequence lengths are presented. This is in the context of incorporating an LSTM layer with a hidden size of 256 before the dense layer in our YOLOv5-based model with 16×16 grids. The results across all sequence lengths are similar.

C.3 More Qualitative Results

C.3.1 BDD-A Dataset

In Figure C.1, additional examples of our YOLOv5-based model applied to the BDD-A dataset are presented. The first row demonstrates the model's accurate detection of a vehicle in the two straight-ahead lanes, while disregarding parked vehicles two lanes over and a car in a turning lane. The second row highlights the model's identification of a central traffic light and two parked vehicles, which are pivotal if the driver were to continue straight. However, as the driver is making a left turn, the ground truth annotations focus on items on the road being turned onto.

	Object-level					Pixel-level	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	*KL*	*CC*
2	0.85	72.40	72.68	72.54	78.00	1.16	0.60
4	0.85	72.58	73.02	72.80	78.18	1.16	0.60
6	0.85	72.52	73.04	72.78	78.16	1.18	0.60
8	0.85	73.13	70.44	71.76	77.83	1.17	0.60
16	0.85	71.84	73.39	72.61	77.86	1.18	0.60

Table C.2: Comparison of different input sequence lengths when using one LSTM layer. Our model uses the 16×16 grids. For all metrics except D_{KL}, a higher value indicates the better performance. ($Th = 0.5$)

Figure C.1: Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.5) and not using attention-based object detection on BDD-A test set. (The Second row is a failed case.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

C.3.2 DR(eye)VE Dataset

Figure C.2 and Figure C.3 present additional instances of object detection using our YOLOv5-based model on the DR(eye)VE dataset. In Figure C.2, the model accurately identifies vehicles on the road, while disregarding cars parked further away in two lanes. Conversely, Figure C.3 showcases the model's detection of a cyclist adjacent to a vehicle and a car positioned to the right. This differs from the ground-truth, which highlights objects that the driver will encounter later. This discrepancy might be attributed to the driver perceiving nearby objects through peripheral vision.

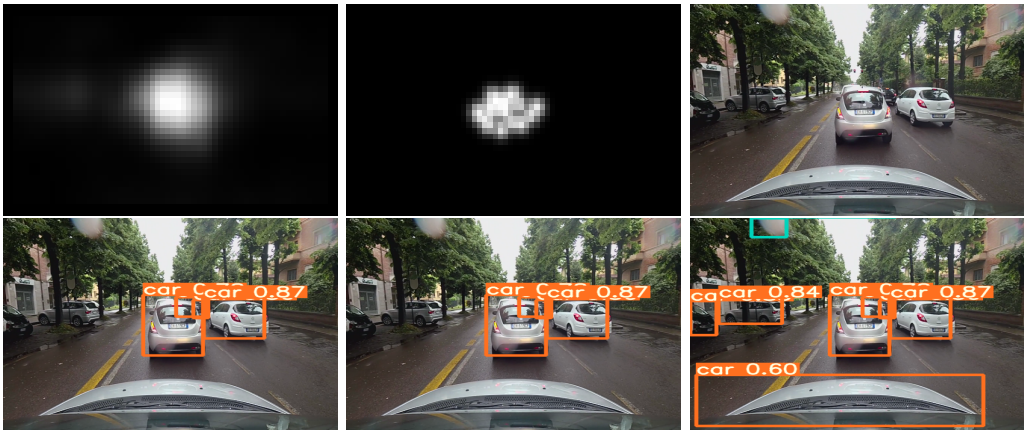


Figure C.2: Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.4) and not using attention-based object detection on DR(eye)VE test set. The first row contains the predicted attention map (Left), ground-truth attention map (Middle) and original frame (Right). The second row contains our object detection (Left), ground-truth (Middle), and object detection without driver attention (Right). Better view in colors.

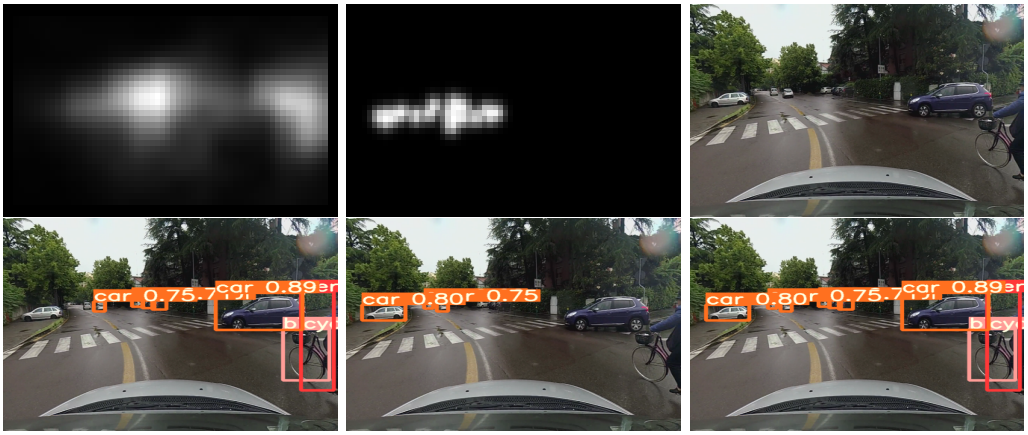


Figure C.3: Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.4) and not using attention-based object detection on DR(eye)VE test set. (Failed case.) The first row contains the predicted attention map (Left), ground-truth attention map (Middle) and original frame (Right). The second row contains our object detection (Left), ground-truth (Middle), and object detection without driver attention (Right). Better view in colors.

D A Consistent and Efficient Evaluation Strategy for Attribution Methods

This chapter provides additional results for the work discussed in Chapter 5. The content is adapted from the work that was published in ICML 2022 [45].

D.1 Additional Experiments on Food-101

D.1.1 Implementation Details

A vanilla ResNet-50 model [169] was trained on the Food-101 dataset [206]. Specifically, the training was conducted using the SGD optimizer with an initial learning rate of 0.01. This rate was subsequently decreased by a factor of 0.1 every 10 epochs. The training spanned 40 epochs, utilizing a batch size of 32, and resulted in the model attaining an accuracy of 81.67% on the test set. For the implementation of the GAN imputation operator, a GAIN model was first trained on the Food-101 dataset as outlined in section 5.2.5. This training employed hyper-parameters of $\alpha = 100$ and $hr = 0.1$, with the model being trained for 100 epochs at a batch size of 32. Eight explanations were computed, and both ROAD and ROAR evaluations were conducted, using the same parameters specified in experiments for CIFAR-10.

D.1.2 Consistency Analysis

In Table D.1, a comprehensive analysis of the Spearman Correlation across rankings from eight evaluation strategies ("Retrain"/"No-Retrain", MoRF/LeRF, along with fixed/Noisy Linear/GAN imputation) for the Food-101 dataset is presented. Bold results in the table highlight the uniformity achieved by employing three different imputation mechanisms. High consistency is noted between the corresponding Retrain and No-Retrain methods, underscoring that the efficiency enhancements discussed in the main text are achievable in larger datasets. The consistency between MoRF and LeRF shows improvement with the retraining process as compared to fixed imputation, although a minor decline is observed with the No-Retain approach. Due to the proximity of the curves on this dataset, particularly in the No-Retrain scenario, even marginal differences might lead to ranking shifts, making the results generally more variable than those observed on CIFAR-10. In summary, while similar trends are noted, the consistency boost between MoRF and LeRF in the No-Retrain case is less significant. It is also worth noting that perfect alignment between MoRF and LeRF may not always be preferable.

		Retrain MoRF			No-Retrain MoRF			Retrain LeRF			No-Retrain LeRF		
		fixed†	lin	gan	fixed	lin*	gan	fixed	lin	gan	fixed	lin	gan
Retrain MoRF	fixed†	1.00 ±0.00											
	lin	0.48 ±0.03	1.00 ±0.00										
	gan	0.50 ±0.04	0.79 ±0.03	1.00 ±0.00									
No-Retrain MoRF	fixed	**0.12** ±0.01	0.57 ±0.02	0.50 ±0.01	1.00 ±0.00								
	lin*	0.61 ±0.01	**0.81** ±0.02	0.67 ±0.04	0.31 ±0.01	1.00 ±0.00							
	gan	0.74 ±0.01	0.79 ±0.02	**0.67** ±0.04	0.35 ±0.01	0.86 ±0.00	1.00 ±0.00						
Retrain LeRF	fixed	**-0.26** ±0.02	0.41 ±0.02	0.30 ±0.02	0.53 ±0.01	0.10 ±0.01	0.11 ±0.01	1.00 ±0.00					
	lin	-0.40 ±0.02	**0.26** ±0.04	0.19 ±0.04	0.30 ±0.03	-0.05 ±0.01	0.09 ±0.01	0.83 ±0.01	1.00 ±0.00				
	gan	-0.18 ±0.01	0.46 ±0.04	**0.32** ±0.04	0.50 ±0.03	0.13 ±0.02	0.14 ±0.03	0.89 ±0.02	0.83 ±0.01	1.00 ±0.00			
No-Retrain LeRF	fixed	0.79 ±0.02	0.79 ±0.03	0.63 ±0.05	**0.32** ±0.01	0.85 ±0.00	0.89 ±0.00	**0.02** ±0.01	-0.15 ±0.02	0.10 ±0.03	1.00 ±0.00		
	lin	-0.28 ±0.02	0.35 ±0.02	0.28 ±0.04	0.46 ±0.00	**-0.03** ±0.00	-0.06 ±0.00	0.89 ±0.01	**0.81** ±0.02	0.87 ±0.01	-0.11 ±0.00	1.00 ±0.00	
	gan	-0.45 ±0.02	-0.08 ±0.03	-0.04 ±0.04	0.23 ±0.00	-0.37 ±0.00	**-0.44** ±0.00	0.58 ±0.01	0.61 ±0.01	**0.54** ±0.00	-0.41 ±0.00	0.70 ±0.00	1.00 ±0.00

Table D.1: Food-10: Rank Correlations between all evaluation strategies used with standard deviations computed by considering the rankings obtained through five consecutive runs as independent. The ROAR benchmark is marked by † and our ROAD by *. Bold results highlight the consistency between Retrain and No-Retrain (still very high) as well as MoRF and LeRF evaluation strategies using different imputation operators (fair increase when using Noisy Linear and GAN imputations instead of fixed imputation in "Retrain", decrease in "No-Retrain").

D.2 Additional Results on CIFAR-10

D.2.1 Extended Figures

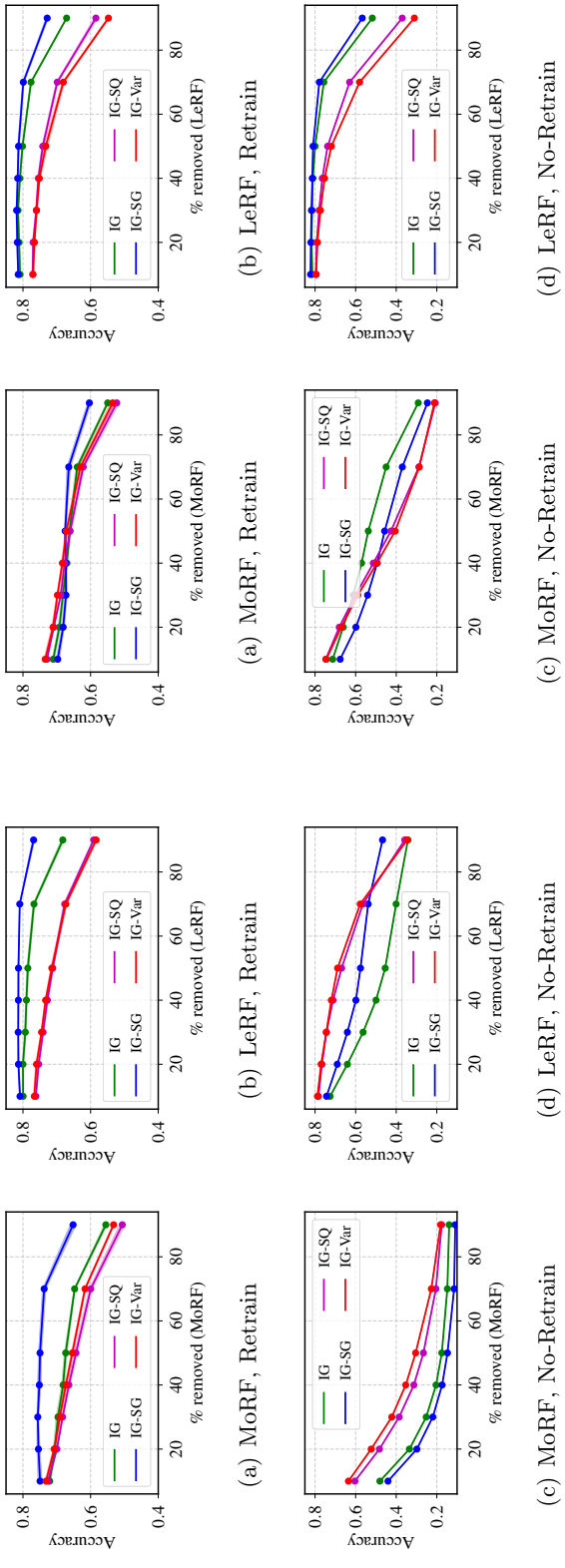
This section presents comprehensive qualitative findings from the application of four different evaluation strategies (namely "Retrain"/"No-Retrain", and MoRF/LeRF) across three distinct imputation methods (fixed value, Noisy Linear, and GAN imputation). The complete set of results for IG-family attribution methods using fixed value imputation can be found in Figure D.1. Similarly, Figure D.4 provides a detailed view of the GB-based attribution methods. The outcomes of employing our Noisy Linear Imputation for both IG- and GB-family attribution methods are depicted in Figure D.2 and Figure D.5, respectively. Notably, the adoption of our Noisy Linear Imputation enhances the alignment between evaluation rankings in both MoRF and LeRF frameworks, with or without retraining. This improvement is particularly evident when comparing the results in Figure D.2 against those in Figure D.1.

D.2.2 Consistency Analysis

In Table D.2, we present a comprehensive overview of the Spearman Correlation across all twelve evaluation approaches employed in our study, encompassing variations like "Retrain"/"No-Retrain", MoRF/LeRF, and methods involving fixed value, Noisy Linear, and GAN imputation. Our primary emphasis in this research was on evaluating the consistency between the Retrain and No-Retrain methods, as well as between MoRF and LeRF. The findings that are central to our main paper are highlighted in bold.

		Retrain MoRF			No-Retrain MoRF			Retrain LeRF			No-Retrain LeRF		
		fixed†	lin	gan	fixed	lin*	gan	fixed	lin	gan	fixed	lin	gan
Retrain MoRF	fixed†	1.00 ±0.00											
	lin	0.68 ±0.02	1.00 ±0.00										
	gan	0.76 ±0.01	0.82 ±0.01	1.00 ±0.00									
No-Retrain MoRF	fixed	**0.15** ±0.01	0.38 ±0.02	0.23 ±0.01	1.00 ±0.00								
	lin*	0.66 ±0.01	**0.84** ±0.01	0.86 ±0.01	0.43 ±0.01	1.00 ±0.00							
	gan	0.65 ±0.01	0.62 ±0.01	0.84 ±0.01	0.14 ±0.01	0.78 ±0.01	1.00 ±0.00						
Retrain LeRF	fixed	**-0.01** ±0.01	0.48 ±0.02	0.28 ±0.02	0.66 ±0.00	0.47 ±0.02	0.13 ±0.01	1.00 ±0.00					
	lin	0.16 ±0.01	**0.61** ±0.01	0.34 ±0.01	0.78 ±0.01	0.50 ±0.01	0.10 ±0.01	0.87 ±0.01	1.00 ±0.01				
	gan	0.15 ±0.01	0.59 ±0.01	0.32 ±0.01	0.74 ±0.00	0.50 ±0.01	0.10 ±0.01	0.90 ±0.01	0.96 ±0.01	1.00 ±0.00			
No-Retrain LeRF	fixed	0.49 ±0.01	0.44 ±0.01	0.69 ±0.01	**0.01** ±0.00	0.60 ±0.00	0.77 ±0.00	**0.09** ±0.01	0.03 ±0.01	-0.03 ±0.00	1.00 ±0.00		
	lin	0.21 ±0.01	0.60 ±0.01	0.38 ±0.01	0.81 ±0.00	**0.58** ±0.01	0.22 ±0.01	0.85 ±0.00	**0.94** ±0.01	0.91 ±0.00	0.10 ±0.00	1.00 ±0.00	
	gan	0.05 ±0.01	0.47 ±0.01	0.17 ±0.01	0.69 ±0.00	0.36 ±0.00	-0.07 ±0.01	0.85 ±0.00	0.86 ±0.01	0.90 ±0.01	-0.14 ±0.00	0.79 ±0.00	1.00 ±0.00

Table D.2: CIFAR-10: Rank Correlations between all evaluation strategies used with standard deviations computed by considering the rankings obtained through five consecutive runs as independent. Results indicated in bold correspond to those reported in Section 5.2. The ROAR benchmark is marked by † and our ROAD by *.



(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

Figure D.1: Consistency comparison using **Fixed Value** imputation on **IG**-based methods on CIFAR-10

(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

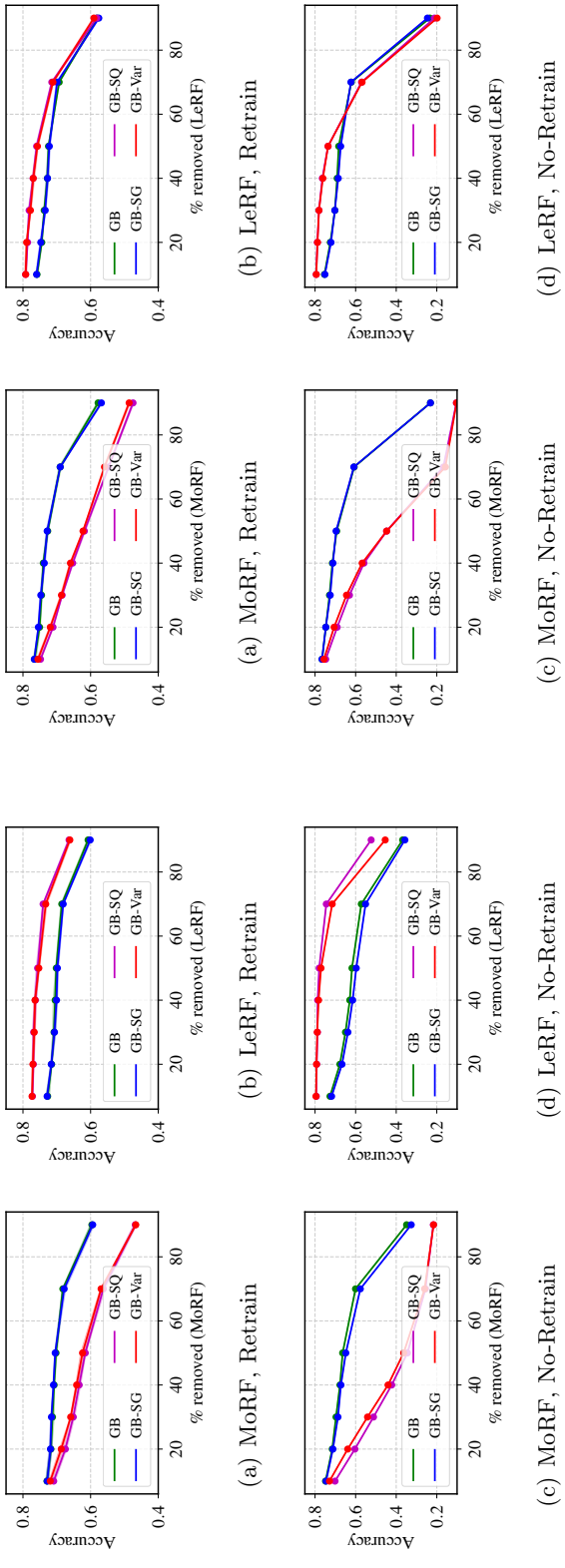
Figure D.2: Consistency comparison using **Noisy Linear** imputation on **IG**-based methods on CIFAR-10

(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

Figure D.3: Consistency comparison using **GAN** imputation on **IG**-based methods on CIFAR-10

(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

Figure D.4: Consistency comparison using **Fixed Value** imputation on **GB**-based methods on CIFAR-10



(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

Figure D.5: Consistency comparison using **Noisy Linear** imputation on **GB**-based methods on CIFAR-10

(a) MoRF, Retrain (b) LeRF, Retrain (c) MoRF, No-Retrain (d) LeRF, No-Retrain

Figure D.6: Consistency comparison using **GAN** imputation on **GB**-based methods on CIFAR-10

E Towards Human-centered XAI

This chapter provides additional analysis details for the guidelines discussed in Chapter 5. The content is adapted from the work that was published in TPAMI 2023 [41].

E.1 Data-driven Bibliometric Analysis

To conduct a bibliometric analysis driven by data, focusing on the references and citations of all key papers[1], we initially gathered prevalent references from each category. Facing a substantial volume of papers, each was tagged with a keyword indicative of its research theme, facilitating categorization based on content. Specifically, references were obtained directly from the papers in pdf format. Works citing these core papers were identified using the Google Scholar platform, utilizing the Python API, "Scholarly" [289]. This API also served to procure abstracts from Google Scholar for all references and citations. In our study, we employed GPT-4 [290] to assign keywords to the papers based on their titles and abstracts. This process was followed by a manual examination to ensure the relevance of these keywords. For visualization, we mapped the papers onto a two-dimensional semantic space, using the keyword embeddings and t-SNE [291] for this purpose.

The key research areas fundamental to user studies in XAI are depicted in Figure E.1 (**Left**). For the sake of visual clarity, this illustration includes only those works referenced in at least five of the main papers. In a similar vein, understanding the beneficiaries of XAI user study findings is crucial. Figure E.1 (**Right**) showcases the "consumers" of these human-centered XAI core papers, meaning the research domains that are influenced by these papers. Here, each dot symbolizes a distinct research topic, with the dot size reflecting the frequency of citations from this topic in our collection of core papers.

By examining the foundational elements and impact of XAI user studies, we gain comprehensive insights into significant topics within this research domain. This approach enables us to identify emerging and significant areas for forthcoming research, like cognition-based analytical tools in XAI. The raw data and code for these analyses are available at `https://github.com/yaorong0921/hxai-survey`.

E.2 Foundation of XAI User Studies

In our analysis of core papers, we have identified a multitude of essential literary sources for XAI researchers, which are instrumental in guiding their project development. We

[1] Here, "references" means sources listed in the references of a core paper, while "citations" are followed-up works citing a core paper

Topic	Fundamental works
Surveys of XAI	[122], [147], [297], [298], [299], [49], [300]
Theories for XAI	[292]: social sciences, [293]: theory for XAI design, [281]: a question bank for XAI design
XAI Methods	[301]: a survey, [86]: LIME, [302]: Anchors, [211]: SHAP, [303]: TCAV, [304]: explaining recommendation systems, [305]: intelligible models, [92]: influence function, [306]: counterfactual explanations, [11]: Integrated Gradient (IG), [307]: saliency maps for images, [212]: GradCAM
Principles of Explanations	[294, 295]: completeness and soundness, [296]: helping users build mental models
User studies for ML	[308]: image retrieval algorithm for medical uses, [309]: interactive model
User studies for XAI	[310]: justice perceptions, [261]: fairness [311]: human-AI team, [312]: usability, [259, 313, 263, 314, 315]: understanding, [316, 218, 229, 317]: trust and understanding
Trust	[255]: trust (calibration), [318]: trust in automation, [253]: impact of model accuracy on trust, [319, 320]: impact of system transparency on trust,

Table E.1: Fundamental works of the core papers (categorized according to topics).

meticulously examined over 3000 references across these pivotal papers, focusing particularly on those referenced by at least ten of them (approximately 50 papers). These significant papers are organized by topic in Table E.1.

The first category of papers contains comprehensive surveys on XAI. Miller et al. [292] suggest grounding XAI in social sciences, such as cognitive science and psychology. Meanwhile, Wang et al. [293] and Liao et al. [281] offer theoretical frameworks for crafting XAI systems. Another critical category is XAI methods, with the most widely utilized ones detailed in "XAI Methods". According to [294, 295], XAI explanations should be both sound and complete to positively influence users. Additionally, XAI aims to aid users in forming accurate mental representations of AI systems, as indicated by in [296]. We also reference previous user studies on ML systems and explainable interfaces, which are used for comparison or as blueprints for user study design. Lastly, we include various general works on user trust that extend beyond the conventional boundaries of XAI.

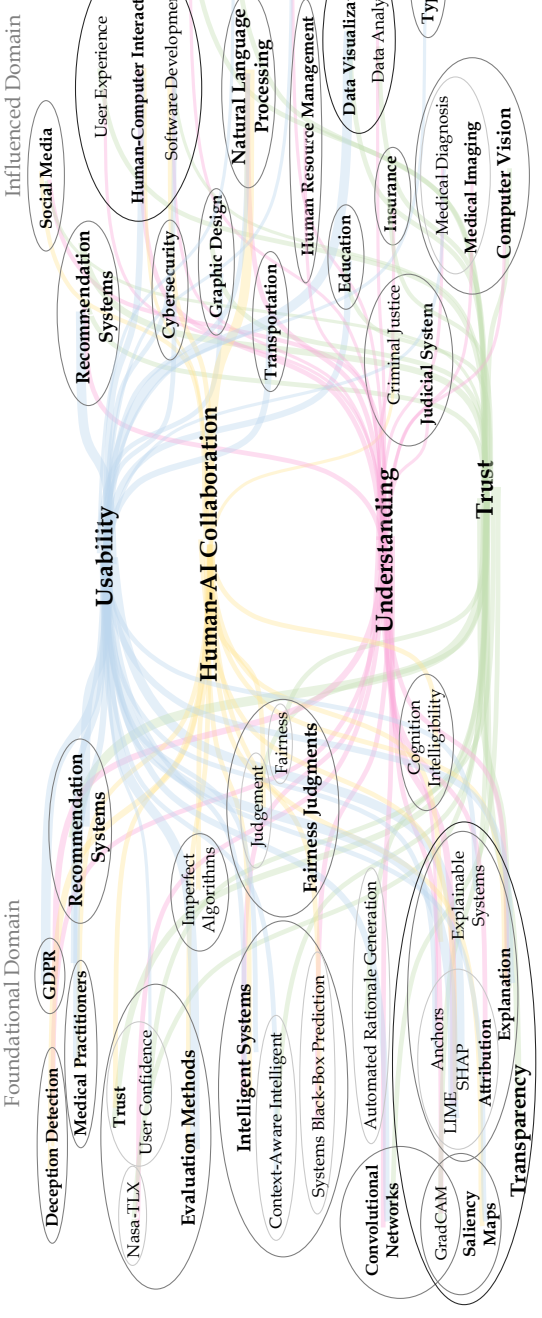


Figure E.1: Illustration of the **foundational** research domains (**Left**): Each dot represents a referenced paper, whose size reflects the number of studied core papers referring to it. Illustration of **influenced** research domains (**Right**): Each dot represents a research topic, whose size refers to the number of papers on the same topic. For a clear depiction, only several important research domains are labeled with text. Lines are used to depict reference links, with thicker lines representing a greater number of links. Core paper categories are in blue (**Middle**). Circles are used to indicate a hierarchical structure of keywords.

F I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise

This chapter provides additional details for the work discussed in Chapter 6. The content is adapted from the work that will be published in AAAI 24.

F.1 Addtional Related Work

Pedagogical Theories on Learning from Errors. XAI has been viewed as an educational interaction, with the XAI method acting as the instructor and the user as the learner [56]. Effective teaching, as supported by pedagogical studies, requires a teacher to evaluate a learner's existing knowledge and tailor their teaching methods accordingly [321, 322]. A key sign of misunderstanding is mistakes, typically arising from wrong associations or comprehension. Addressing these mistakes effectively demands providing correct answers along with clarifying explanations, which have been proven to be significantly beneficial [271]. These insights from educational research form the basis of our XAI framework, particularly influencing our approach to selecting examples. Specifically, I-CEE focuses on explaining those images where it predicts the user will err. Moreover, the greater the certainty of an error, the more potent the learning from it [272, 273]. This phenomenon is known as the hypercorrection effect. In line with the hypercorrection effect, our framework targets images where the user is less confident about the correct label (i.e., more confident about an incorrect label). We argue that utilizing such examples will lead to more effective learning outcomes.

F.2 Target Models and Explanations

F.2.1 Datasets

Our evaluation of the proposed method involves four distinct datasets. The synthetic dataset consists of 960 images for training purposes and 240 images for testing, categorized into four classes: Red-Cylinder, Orange-Cylinder, Red-Cube, and Orange-Cube. The CIFAR-100 dataset [203] includes a total of 60,000 images, where 50,000 are used for training and 10,000 for testing. This dataset covers 100 varied classes, each with 600 images. The CUB-200-2011 dataset, which is specifically designed for fine-grained analysis of different bird species, contains 11,788 images in total. Of these, 5,994 are allocated for training and 5,794 for testing, containing 200 bird species. On average, each species has about 30 images in the training set and another 30 in the testing set.

We apply fine-tuning to the ResNet-18 model, originally pre-trained on ImageNet, for our desired datasets [169]. For the synthetic dataset, these models undergo training using the Stochastic Gradient Descent (SGD) Optimizer. The learning rate is established at $1e^{-4}$, and the training spans 10 epochs. In the case of realistic datasets, we adjust the learning rate to $1e^{-3}$ and extend the training duration to 50 epochs. We resize the input images to 224×224 for all datasets, except for the CUB-200-2011 dataset, where the resizing dimension is 448×448. The training process includes a technique of random horizontal flipping as a form of data augmentation. The initial row of Table F.1 presents the test accuracy figures for the target model across each dataset, encompassing all test classes.

	Synthetic	CIFAR-100	CUB-200-2011	GTSRB
Test (all)	1.00	0.73	0.78	0.99
Test (subset)	1.00	0.82	0.81	0.99

Table F.1: Accuracy of target models. The first row indicates the accuracy of all test classes. The second row contains the accuracy for classes selected for training simulated user models.

Our approach utilizes GradCAM, as described in [212], applying it post the final convolutional layer in the designated model to produce explanations. We favor GradCAM for its alignment with human gaze-based attention in pinpointing key visual features, as noted in [28], enhancing interpretability over other explanatory techniques.

The saliency maps are adjusted to match the size of the original input, denoted as $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{e} \in \mathbb{R}^d$. Figure F.1 presents examples of these explanations across various datasets. Notably, in the CIFAR-100 dataset, the model emphasizes the train's locomotive, a critical characteristic. Similarly, in the GTSRB dataset, the model distinctly highlights the "left turn" symbol on traffic signs.

F.3 Hyper-parameter Settings

This section illustrates the impact of selecting m, the number of concepts, across various realistic datasets. Table F.2 presents the accuracy of the simulated user trained (evaluated with user annotations) alongside the count of trainable parameters in the concept vector $\mathbf{c}$ and the mapping function $\Xi(\cdot)$. The optimal value of m is highlighted in bold, balancing accuracy against the parameter count. In scenarios where user models must be deployed in resource-constrained real-world environments, minimizing the number of trainable parameters is preferable. The data indicates that an m value of 64 on three datasets already results in high test accuracy. Adding more concepts, although it might seem beneficial, doesn't significantly enhance performance and instead increases computational demands.

(a) Synthetic dataset (b) CIFAR-100 (c) CUB-200-2011 (d) GTSRB

Figure F.1: Illustration of model explanations on each dataset. The saliency map highlights the important area (feature) that is important for the model decision.

F.4 Details of Baselines

F.4.1 Bayesian Teaching

Our approach adopts the Bayesian Teaching framework, as outlined in [58]. However, we modify the image selection process due to the absence of two specific classes in our query set. Specifically, [58] employs a binary choice task, utilizing Bayesian Teaching probabilities to select two samples from both the predicted class of the target model and a pre-defined alternative class. These images are intended to guide the explainee model $f_L(\cdot)$ to classify a target image with the label assigned by the target model. In our adaptation, we omit the selection from the alternative class. More precisely, we focus on the probability that a given image $\mathbf{x}$ is part of class y, from which we sample another image τ^y. This probability is expressed as $f(\mathbf{x}|\tau^y)$, a concept borrowed from [58]). Under the PLDA model [276], this probability can be expressed in the form of the normal distribution as follows:

$$f(\mathbf{x} \mid \tau^y) = \mathcal{N}(u \mid \frac{\Psi}{2\Psi + \mathrm{I}}\, u^y, \frac{\Psi}{2\Psi + \mathrm{I}} + \mathrm{I}), \tag{F.1}$$

where u is the image $\mathbf{x}$ transformed by the shift vector $\mathbf{m}$ and rotation and scaling matrix A in the PLDA layer. Likewise, the image τ^y is transformed to u^y. Ψ is another parameter in the learned PLDA layer. To integrate the PLDA layer into the user model (also known as the explainee model), we train a ResNet-18 network, replacing its final layer with the PLDA layer. This training uses user annotations as labels. With the trained user model, we are able to calculate $f(\mathbf{x}|\tau^y)$, which facilitates the selection of images by ranking according to this term.

(a) CIFAR-100

m	16	32	**64**	128
Acc	85.00 ± 0.50	89.25 ± 0.34	93.50 ± 0.70	96.5± 0.23
# Param. (M)	0.94	0.97	1.05	1.19

(b) CUB-200-2011

m	16	32	**64**	128
Acc	25.75 ± 0.78	27.27 ± 0.89	63.35 ± 0.45	65.15 ± 0.60
# Param. (M)	3.67	3.73	3.85	4.08

(c) GTSRB

m	8	16	32	**64**
Acc	89.17± 0.34	85.83 ± 0.35	98.33 ± 0.23	100.0 ± 0.10
# Param. (M)	0.92	0.94	0.97	1.05

Table F.2: Effect of m on the user model performance.

F.4.2 Active Learning Baselines

Our study integrates foundational benchmarks from the field of active learning. These benchmarks offer a variety of selection methodologies, underscoring the efficiency of our introduced Hypercorrection Effect. The formula for Expected Gradient Length [277] (EGL) is presented as follows:

$$x_{\text{EGL}} = \underset{x}{\operatorname{argmax}} \sum_{i}^{K} f_\theta(y_i \mid \mathbf{x}, \mathbf{e}) \; \|\nabla \; l_\theta(\mathcal{L} \cup \langle \mathbf{x}, \mathbf{e}, y_i \rangle)\|, \tag{F.2}$$

where $f_\theta(\cdot)$ denotes the trained user model in our case with parameters θ. To integrate $\mathbf{e}$ into the input, we utilize the explanation $\mathbf{e}$ as a weighted mask, following the method outlined in the "Selection Strategy" section. The model's training is guided by the objective function $\mathcal{L}$, represented by the cross-entropy loss. Consider $\nabla \; l_\theta(\mathcal{L})$ as the gradient of this objective function in relation to θ. Given that the model has reached convergence in the final training phase, the Euclidean norm of the objective function, denoted as $\|\nabla \; l_\theta(\mathcal{L})\|$, is expected to approach zero, as discussed in [277]. Consequently, x_{EGL} can be expressed in a simplified form:

$$x_{\text{EGL}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i}^{K} f_\theta(y_i \mid \mathbf{x}, \mathbf{e}) \; \|\nabla \; l_\theta(\langle \mathbf{x}, \mathbf{e}, y_i \rangle)\|. \tag{F.3}$$

We expand Expected Gradient Length (EGL) by incorporating the concept of EGL-Shift, which focuses exclusively on the variable $\mathbf{x}$ in the input. The goal of EGL-Shift is to diminish the effect of the image on the training gradient while accentuating the role

of explanations. Specifically, the calculation of EGL-Shift is executed as follows:

$$x_{\text{EGL-Shift}} = \underset{\mathbf{x}}{\text{argmax}} \ (\sum_{i}^{K} f_\theta(y_i \mid \mathbf{x}, \mathbf{e}) \ \|\nabla \ l_\theta(\langle \mathbf{x}, \mathbf{e}, y_i \rangle)\| - \sum_{i}^{K} f_\theta(y_i \mid \mathbf{x}) \ \|\nabla \ l_\theta(\langle \mathbf{x}, y_i \rangle)\|). \tag{F.4}$$

The Density-Weighted Method (DWM) [153] is effectively integrated with a fundamental selection approach like EGL. This method focuses on selecting data points that not only exhibit uncertainty but are also indicative of the overall distribution present in the input data. It estimates the distribution for a given data point by assessing the degree of similarity it shares with other points in the dataset. The process of implementing DWM is detailed as follows:

$$x_{\text{DWM}} = \underset{\mathbf{x}}{\text{argmax}} \ \phi_A(\mathbf{x}) \cdot (\frac{1}{U} \sum_{u=1}^{U} \text{sim}(\mathbf{x}, \mathbf{x}^{(u)}))^\beta, \tag{F.5}$$

where $\phi_A(\mathbf{x})$ represents the computation of EGL for $\mathbf{x}$. The entire input dataset is denoted by U. As per the approach outlined in [153], we assign the value of 1 to β; The degree of resemblance between two images is determined by measuring the cosine similarity of their feature vectors within the latent space.

F.5 Computational Infrastructure

All experiments in the project "I-CEE" were conducted on the device as listed below:

Device Attribute	Value
Computing infrastructure	GPU
GPU model	NVIDIA GeForce RTX 2080 Ti
GPU number	1
CUDA version	11.3

Table F.3: Computational infrastructure details.

Human Attention in Fine-grained Classification

Yao Rong[1]
yao.rong@uni-tuebingen.de
Wenjia Xu[2]
xuwenjia16@mails.ucas.ac.cn
Zeynep Akata[1,3]
zeynep.akata@uni-tuebingen.de
Enkelejda Kasneci[1]
enkelejda.kasneci@uni-tuebingen.de

[1] University of Tübingen, Germany
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

The way humans attend to, process and classify a given image has the potential to vastly benefit the performance of deep learning models. Exploiting where humans are focusing can rectify models when they are deviating from essential features for correct decisions. To validate that human attention contains valuable information for decision-making processes such as fine-grained classification, we compare human attention and model explanations in discovering important features. Towards this goal, we collect human gaze data for the fine-grained classification dataset CUB and build a dataset named CUB-GHA (Gaze-based Human Attention). Furthermore, we propose the Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN) to integrate human gaze knowledge into classification models. We implement our proposals in CUB-GHA and the recently released medical dataset CXR-Eye of chest X-ray images, which includes gaze data collected from a radiologist. Our result reveals that integrating human attention knowledge benefits classification effectively, e.g. improving the baseline by 4.38% on CXR. Hence, our work provides not only valuable insights into understanding human attention in fine-grained classification, but also contributes to future research in integrating human gaze with computer vision tasks. CUB-GHA and code are available at https://github.com/yaorong0921/CUB-GHA.

1 Introduction

Through a lifelong learning process, humans have developed a selective attentional mechanism, which has received attention in many areas of artificial intelligence [54]. As human attention can be revealed from gaze data, it bears the potential to explain our behavior and decisions [31]. Many computer vision applications embrace human gaze information to detect salient objects for solving tasks [19, 34, 39]. To visually illustrate human attention in these tasks, it is common to add a Gaussian filter on fixation points to form a feature map [15], which is also called *saliency* map [21] (see Figure 1). Similar to how gaze explains

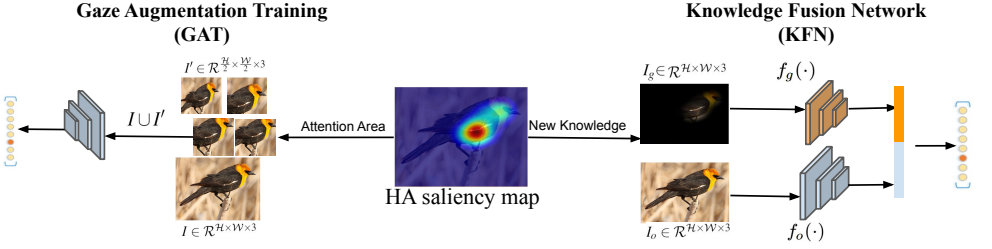


Figure 1: Overview of our proposed methodology. HA salicency map is used to obtain attention area which is used to enhance the training dataset in Gaze Augmentation Training (Left), while it is used as extra knowledge and fused together with the image knowledge in the Knowledge Fusion Network (Right).

human decisions, the post-hoc attention of a network, i.e. model explanation, tries to reveal important regions for neural network decision-making [12, 30, 35, 40, 42, 57]. Both can be visualized by means of saliency maps, thus allowing the study of similarities and differences between them. In this context, several previous works show that humans and models are looking at different regions when performing the same task [7, 36]. However, it is not clear whether a feature discovered by a human is more efficient for solving a given task or not. Our work addresses this research gap and the hypotheses that (1) human attention focuses on essential features for solving the task (e.g. fine-grained classification); (2) using human attention also allows improving model performance in accomplishing the task. To validate the first hypothesis, we first capture and present human attention in the style of a saliency map. We compare the regions that human attention covers with the ones that are discovered by the model (model explanation), and show that human attention hints on the regions that are more discriminative in the classification. We propose two modules which make use of the essential features revealed by human gaze to validate the second hypothesis: we use Gaze Augmentation Training (GAT) to train a better classifier and a Knowledge Fusion Network (KFN) to integrate the human attention knowledge into models.

Our contributions are as follows: (1) We collect human gaze data for the fine-grained data set CUB, enhance it by incorporating human attention and coin this new dataset as **CUB-GHA** (Gazed-based Human Attention). For this novel dataset, we also validate the efficiency of human gaze data in discovering discriminative features. (2) We propose two novel modules to incorporate human attention knowledge in classification tasks: Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN). (3) To showcase the relevance of our work for highly relevant applications, we evaluate our methods not only on our novel CUB-GHA dataset, but also on chest radiograph images from a recently released dataset CXR-Eye (which contains also gaze data). Our work shows that human attention knowledge can be successfully integrated in classification models and help improve the model performance with regard to the state-of-the-art in different classification tasks.

2 Related Work

Human Gaze in Machine Learning. Recent developments in hardware devices allow for the precise recording of eye movements in different activities, ranging from human-computer interaction [26, 27] to complex and dynamic real-world tasks, such as driving [4, 47] and robotics [3, 38, 45]. Furthermore, the way that visual information is processed can reveal

information about a person's strategy or level of expertise [5]. In the medical domain, researchers have validated that gaze data reveals patterns which can benefit AI models, as for disease (Pneumonia and Congestive Heart Failure) classification [18]. In computer vision, gaze data has proven its usefulness in various applications [19, 32, 34, 39]. E.g., [19] collects gaze (coordinates, duration, etc.) vectors for 60 bird classes in dataset [46] to form embeddings for zero-shot learning. [32] compares the attention map generated by an attention module (two convolutional layers) with human attention maps generated by the data from [19] and shows that human attention surpasses the attention module. [34] proposes a photograph cropping system using the collected fixation data to identify important content and compute the best crop. Eye tracking data is also used to extract dominant objects in videos [39]. Different from previous works which use gaze for specific tasks [19, 34, 39], our proposal GAT leverages human attention to train a better backbone which can be used in many different tasks and frameworks. Moreover, we evaluate GAT and KFN for two different classification tasks and thus show the general validity of our methods.

Attention Module in Fine-grained Classification. Many previous works [10, 14, 22, 23, 24, 37, 41, 51, 55, 56, 58] integrate attention modules in networks to localize the parts which are important for fine-grained classifications and make use of the information of the discriminative parts to improve the models' performance. [10, 22, 23, 24, 37] adopt the Recurrent Attention Model (RAM) [28], where an attention agent is deployed to predict locations of the discriminative regions, and train the classifier based on these cropped regions. The attention agent is trained with a reinforcement learning algorithm to address the non-differentiability due to the cropping operation. However, the architecture of this attention model is cumbersome with high computational cost. [14, 41, 51, 55, 56, 58], on the other hand, design attention modules using the output from intermediate layers in networks and enforce it to capture discriminative features. Compared to previous works, we do not use the intermediate outputs from networks to generate model attention but use human attention maps. Our method augments the training set with regions cropped according to human attention and thus accomplishes training a better classifier. We compare our method with previous works and demonstrate the profit of exploiting human attention in Section 5.

3 CUB-GHA Dataset

In this section, we first provide the details of our gaze data collection paradigm and then analyze the effect of machine explanation and human attention to the fine-grained classification model. To collect gaze data, we employ the CUB-200-2011 (CUB) [44] dataset with 11,788 images from 200 bird classes incorporating various annotations: image-level attributes, body part locations, and text descriptions of the bird. Our annotation leads to a human-gaze enhanced version, i.e. CUB-GHA.

We choose the fine-grained CUB dataset for two reasons: 1) The difference between two similar classes lies in local and compositional attributes, which can be precisely captured by human gaze. For instance, it is challenging to achieve a measure for unified human attention when comparing a bear and a horse as there are many differences between them. In contrast, distinguishing between two similar birds with different throat colors presents a more unified problem (as shown in Figure 2). 2) The CUB dataset is widely used for various computer vision tasks, such as fine-grained classification [9, 10, 56], zero-shot learning [1, 48, 49, 53], explainable artificial intelligence [2, 6, 16], etc. Thus, our CUB-GHA may serve as a valuable foundation for exploring the effect of human attention on those tasks.

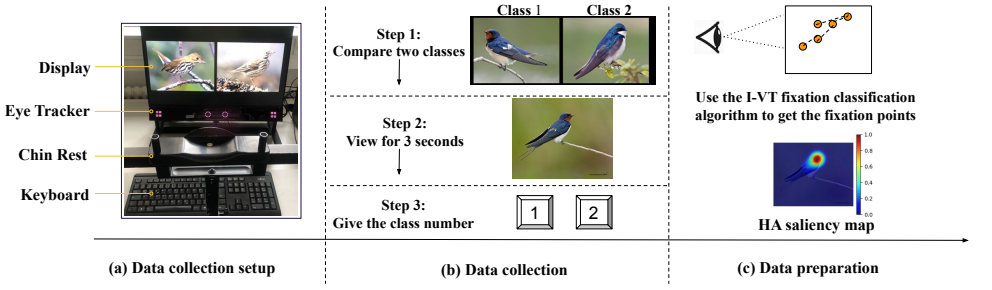


Figure 2: **(a)** Eye tracker set-up: We use a Tobii Spectrum eye-tracker to capture gaze information at a high frequency of 1200 Hz. **(b)** Data collection: Step 1 represents a schematic overview of the image comparison task where two images of different species are freely viewed. In Step 2, a randomly selected example of one of the species is shown to the user for which gaze data is then collected. To gamify this setting, the user is asked to choose the correct class in Step 3. **(c)** Preparing human attention data: we visualize human attention in Gaussian-based saliency maps.

3.1 Gaze Data Collection

Collection Framework. As illustrated in [19], humans fixate on class-discriminative features when they observe two very similar classes. In this paper, we adopt an image comparison game [19], where we encourage participants to look at the discriminative features when comparing two similar images from different categories. The comparison task is designed to be challenging to provide more powerful insights, i.e. two classes in one comparison pair are chosen to be very similar.

A schematic overview of our data collection is presented in Figure 2. Figure 2 (a) shows the experimental setup including a picture of the eye-tracker (Tobii Spectrum Eye Tracker, sampling at 1200 Hz) and the chin rest as well as the display (1920 $\times$ 1080 resolution). The chin rest is used to ensure precise recordings of the eye movements. Each image is re-scaled to fit to the screen and placed at the center. The average distance between the participant's nose and the screen is approximately 60 *cm*. The comparison task consists of three steps shown in Figure 2 (b). In step 1, we present two representative images at the same time, each from one bird class of the CUB dataset, e.g. representative images of Barn Swallow and Tree Swallow. We choose the comparison pairs under the same sub-classes, and then different persons manually check the visual similarity to make sure that the comparison is not too simple. The participants are allowed to observe the images for as long as they want. When the participant is ready for the classification task, in step 2, an image from one of the two classes of the CUB dataset is shown. The participant has to choose which category the image belongs to by viewing the image. Note that the image shown for classification is displayed for only 3 seconds to avoid explorative gaze behavior unrelated to the task. One collection session includes one image from each class, meaning that there are 200 images reviewed per session. Every image in CUB is reviewed by five different participants. 25 subjects (19 males and 6 females with mean age 27.64 $\pm$ 4.15) participate in the experiment. Although the participants do not take part in the same number of sessions and instances, we make sure that every participant views all classes in every session. It is worth noting that all participants are domain novices with no specific knowledge about birds.

Gaze Data Preparation. The raw gaze data is preprocessed to extract fixation locations using the Velocity-Threshold Identification (I-VT) algorithm [29]. The resulting fixation points offered in the dataset include coordinates and duration information. Based on this

information, we generate saliency maps for human gaze as shown in Figure 2 (c). Every fixation location is modelled as a Gaussian distribution $G(\mu, \sigma^2)$, where σ is 75 pixels (in the displace resolution), according to the ratio of the distance to the screen and the approximate foveal area of 2°. The duration of the fixation is then used as a weight for its Gaussian distribution. Finally, the saliency map is presented in grayscale image form. From here on, we note the human attention saliency map generated from gaze data as HA.

3.2 Gaze Data Analysis

In this section, we validate the hypothesis that *HA covers discriminative regions for the fine-grained classification*. Given the same image and the same (visual) task, HA and model explanation (ME) reveal regions which are important in making decisions for humans and models, respectively. Thus, we compare HA with four MEs provided by a trained classifier (vanilla ResNet-50 [11]) with a classification score of 85.58% on CUB , and validate that HA is able to discover features that better differentiate the bird from other bird classes. The four ME used are Class Activations Maps (CAM) [57], Gradient-based CAM (Grad-CAM) [35], InputXGradient (IxG) [40], and IntegratedGradients (IG) [42].

For quantitative comparison, we compare HA and ME using the keep and retrain (KAR) procedure (proposed in the appendix to [12]) to validate if the important regions highlighted by HA and ME help the model to make decisions. Concretely, we gradually insert important pixels to a blank image according to their values in HA or ME saliency maps. The modified percentage of pixels is [5,10,15,20,25,30,50,70,90]. After a certain amount of pixels are inserted, we retrain a new model using the modified train images and report the accuracy on modified test sets. Modified images at 5%, 20% and 70% of pixels inserted using Grad-CAM are shown in Figure. 3 (middle). The intuition behind this is that the class-discriminative information should be included in the pixels that are evaluated as very important; with more pixels inserted which are relatively less important, the model performance will not improve much. If a saliency map selects the informative features as being the important ones for classification, the increase of accuracy at the beginning of insertion is rapid, i.e. the resulting higher Area Under the Curve (AUC) indicates a better feature importance estimate.

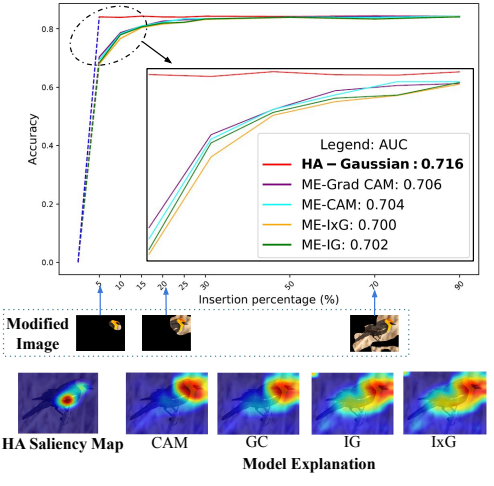


Figure 3: Comparison of HA and ME in discriminative feature discovery. **Top:** Test accuracy on modified datasets using different saliency maps. The x-axis is the insertion percentage and the y-axis is the accuracy on test set. The AUC of each curve is reported in zoom-in image. **Middle:** modified images (using Grad-CAM as an example). **Bottom:** Illustration of HA and four MEs.

The keep and retrain curves and the AUC scores for each method are shown in Figure. 3 (top), and the qualitative saliency maps for HA and four MEs for one image are shown in the bottom. We see that HA and MEs do not focus on the same image regions: humans consider the white feathers on the black wing as a more important feature, while the model uses the

yellow head as the most important feature (see the original image in Figure 1). HA discovers more informative and important features for the fine-grained classification model than the MEs do, e.g. HA obtains an AUC score of 0.716 compared to Grad-CAM (0.706) and IG (0.702). With only 5% important pixels revealed, the model trained with HA modified images can reach an accuracy of 81% while the model trained with ME modified images only reaches an accuracy of around 70%. More details of the analyses can be found in the supplementary material.

4 Methodology

In this section, we introduce how we incorporate the gaze information to improve the classification performance, i.e. using gaze to augment training data (GAT) or as an extra information source (KFN). The illustration of the architecture is shown in Figure 1.

4.1 Gaze Augmentation Training

Motivated by the assumption that the model should pay attention to the discriminative image regions (highlighted by HA), we enhance our model's reaction to those regions by adding them as augmentation in training as illustrated in Figure 1 (left).

To get the k augmentation images for the input image $I \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 3}$ (where $\mathcal{H}$ and $\mathcal{W}$ represent the width and height of the input image), we implement a sliding window algorithm to find areas which contain human attention. A window with the size of (w,h) slides on the HA map $A \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$ from the upper left to the right bottom corner (with stride size s in both dimensions). We rank all the window areas according to the averaged pixel values inside windows and get k cropped images according to top-k highest scores. We resize the cropped images to the half of the width and height of the I, i.e. $I' \in \mathcal{R}^{\frac{\mathcal{H}}{2}\times\frac{\mathcal{W}}{2}\times 3}$, as suggested in [10, 37, 55] where the attended regions are resized into smaller sizes. I' has the same label y as I does. To get various regions, we use various window sizes and the non-maximum suppression. The training set is extended to $I \cup I'$. We train the model on the enlarged dataset with cross-entropy loss. Note that GAT just needs human gaze information in training and the model takes only original images as inputs in the test phase.

4.2 Knowledge Fusion Network

As shown in Fig. 1 (right), our KFN is a two-branch network that fuses the knowledge from HA and the original image features together. The first branch is the image knowledge branch. This branch takes the original images $I_o \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 3}$ as the input, where $\mathcal{H}$ and $\mathcal{W}$ represent the width and height of the input image, respectively. We use a CNN backbone $f_o(\cdot)$ to extract image feature $f_o(I_o) \in \mathcal{R}^{D_o}$ from I_o, where D_o denotes the dimension of the feature channel. Another branch, the HA knowledge branch, incorporates the gaze features of this image. We multiply the gaze information (HA) with the input image by $I_g = I_o \odot A$, where $A \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}\times 1}$ is the HA saliency map. Through this operation, pixels in the image get different weights from the gaze: the area where humans pay attention to is brighter than the rest. I_g contains visual features which are important for the classification. Another CNN backbone $f_g(\cdot)$ is utilized to extract the gaze feature as $f_g(I_g) \in \mathcal{R}^{D_g}$. Then the gaze feature and original image feature are concatenated together to form the fused feature $f(I_o, I_g) \in \mathcal{R}^{(D_o+D_g)}$. It this way, we integrate HA into a multiclass classification task to study the

potential of HA to improve the performance of the image classifier. The whole network is trained with cross-entropy loss.

5 Experiment

In this section, we first introduce datasets and implementation details. Then we show the results of our proposed GAT and KFN. To show the general validity of our methods, We test on two datasets: CUB-GHA and Eye Gaze Data for Chest X-rays (CXR-Eye) [17].

5.1 Datasets and implementation details

CUB-GHA includes 11788 images in total, with 5994 images for training and 5794 for validation [44]. Each image contains eye gaze data from 5 participants. CXR-Eye includes 1083 chest X-ray images with gaze data from a radiologist while performing routine radiology readings [17]. The goal of this dataset is to make a prediction based on the chest X-ray image, whether the subject has one of two clinically prevalent diseases (pneumonia or congestive heart failure (CHF)), or the subject is healthy (normal). The human gaze data is also visualized in the saliency map style. Each image is annotated with one label out of three classes. We choose this dataset because it is a unique human gaze dataset in the medical domain. For such safety-critical applications (e.g. computer-aided diagnosis), we believe the integration of human attention can increase the acceptance and trust of these applications among users.

In our experiments on the CUB dataset, the input images are resized to 448×448 (the images are cropped to this size with the smaller edge first resized to 448) and then randomly flipped horizontally in training. We use the SGD optimizer [33] with an initial learning rate of 0.001. In the experiments on the CXR dataset, the input images are resized to 224×224 and a random horizontal flip is used in training. We use the Adam optimizer [20] with an initial learning rate of 0.0005. Since the CXR-Eye dataset is relatively small, we run 5-fold cross validation and report the average accuracy of the five validation sets as the final score. All experiments are run for totally 100 epochs training on a single NVIDIA GeForce RTX 3090 and the learning rate decreases after every 50 epochs by a factor of 0.1.

For GAT and KFN, we use ResNet-50 [11] and EffiecientNet-b5 [43] pretrained on ImageNet as backbones on CUB and CXR, respectively. In GAT, we crop the original image using three sets (large, medium and small) of window sizes (more details can be found in the supplementary material). Inside each set of window sizes, we run a sliding window algorithm and get k augmentation images for each image in the training set. Concretely, k is set to 2 for large, 3 for medium and 4 for small scale, which results in 9 augmentation images in total. When combining GAT and KFN, we use the GAT trained classifier as backbone in our KFN and fine-tune the KFN for only 20 epochs.

5.2 Evaluation on CUB-GHA

Ablation study. To measure the influence of GAT and KFN on the fine-grained classification, we design an ablation study on the CUB dataset where we train a ResNet-50 with cross-entropy loss as the baseline, and several variants by adding GAT and KFN training modules to the baseline. From the results shown in Table 1, we observe that both GAT and KFN can improve the fine-grained classification accuracy by a large margin. GAT (with

HA) improves the baseline model by 2.42% to 88%, which indicates that human gaze falls on areas containing discriminative features for classification. When using HA in KFN, the accuracy score is increased from 85.58% to 86.99%, which demonstrates that KFN integrates the knowledge of human attention successfully. To show the effectiveness and uniqueness of HA knowledge, we use two machine explanation methods Grad-CAM [35] and IG [42] as the saliency maps, replacing HA in GAT and KFN. HA surpasses both methods in the GAT and KFN modules, e.g. KFN (HA) gains 86.99% while KFN (IG) gains 85.66%. It indicates that human gaze contains unique knowledge that can not be acquired by the model itself. From the result of GAT+KFN, we observe that the combination of both exceeds using any of them alone.

Method		Acc.
ResNet-50 [11]		85.58
GAT	Grad-CAM [35]	87.68
	IG [42]	87.73
	HA	88.00
KFN	Grad-CAM [35]	85.04
	IG [42]	85.66
	HA	86.99
GAT+KFN	HA	**88.66**

Table 1: Ablations study of GAT and KFN on CUB. "Acc." denotes the accuracy in %.

Method	Acc.
MixUp [52]	86.23
CutMix [50]	86.15
SnapMix [13]	87.75
Ours (GAT)	**88.00**
OSME+MAMC[41]	86.30
TASN [56]	87.90
API [58]	87.70
ACNet [14]	88.10
Ours (KFN+GAT)	**88.66**

Table 2: Comparison with the state-of-the-art methods on CUB. **Top:** Comparison of GAT with data augmentation methods. **Bottom:** Comparison of GAT+KFN with attention-based models.

Comparison with state-of-the-art. We compare our proposed modules with several state-of-the-art methods. Note that for a fair comparison, we compare with the results of using ResNet-50 as the backbone and the input resolution of 448×448. First, we compare our GAT with other data augmentation methods, i.e., MixUp [52], CutMix [50] and SnapMix [13] in Table 2 (top). The difference between our GAT and other data augmentation methods is that we do not generate synthetic images. MixUp combines two images and their labels linearly, while the rest replace one part of the image with one part from other images. Our GAT simply extends the dataset with the cropped images, which introduces very low computation cost to train the classifier. Among all these works, training a ResNet-50 with GAT outperforms with other state-of-the-art augmentation methods and achieves an accuracy of 88%. Moreover, this better trained backbone can be combined easily with other framework to further improve the performance, for instance we combine it with our KFN and thus get better results.

We compare our full network with the attention-based methods on CUB in Table 2 (bottom). We choose these methods (OSME+MAMC [41], TASN [56], API [58] and ACNet [14]) due to their high performance and relevance in simulating human attention by attention modules. They apply attention modules to capture discriminative features from the intermediate output in the network, while we use and integrate the HA directly. For instance, [14, 41] applies several layers on the top of the output of the residual block to obtain the region features; API [58] simulates the comparison behavior of humans as our participants do in the data collection in order to learn discriminative representations. Our full network out-

Method	S3N [8]	S3N + GAT (Ours)	CrossX [25]	CrossX + GAT (Ours)	MMAL [51]	MMAL + GAT (Ours)
Accuracy	87.95%	88.91%	87.70%	88.51%	89.25%	**89.53%**

Table 3: Combining our GAT model with the state-of-the-art methods on CUB.

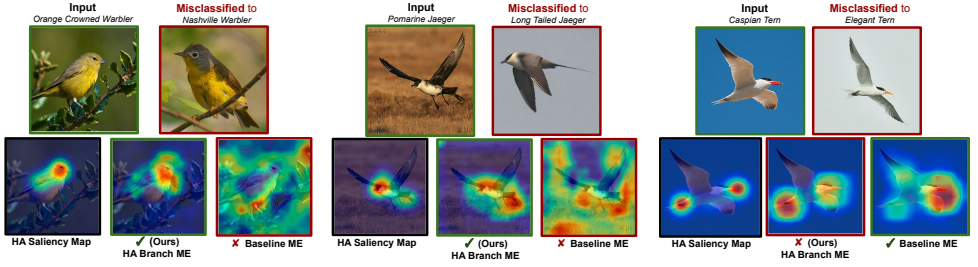


Figure 4: Illustration of model explanations using HA. Two improved examples and one failure example of our model are shown. For each example, we show the input and misclassification classes; HA saliency map, model explanation of our model, and the baseline model.

performs all state-of-the-art models, achieving 88.66% compared to the attention networks API (87.70%) and ACNet (88.10%). The high performance of our KFN and GAT validates that human gaze can benefit a model's performance in the task.

We combine our module with other state-of-the-art models flexibly and thus improve the performance. In Table 3, we show our re-implementations with official code and our improvement by combining our GAT in S3N [8], CrossX [25] and MMAL [51] models. Please note that no HA information is needed in the inference phase. Our combination of MMAL and GAT improves MMAL from 89.25% to 89.53%. We improve CrossX from 87.70% to 88.51% and S3N from 87.95% to 88.91%, which also surpass the best results given in [8, 25].

Qualitative results. We show two examples from two classes whose accuracy is improved the most compared to the baseline model (vanilla ResNet-50), and one example of a class where our model fails to classify correctly in Figure 4. In the first example, the baseline model looks at the belly of an Orange Crowned Warbler and misclassifies it as a Nashville Warbler who also has a yellow fluffy belly. Our model instead focuses on the throat, which is discriminative between the two classes: an Orange Crowned Warbler has a yellow throat, while a Nashville Warbler has a clear mixture of gray and yellow colors on its throat. In the second example, the discriminative feature is the tail. The baseline model mistakes the background as the tail, while our model localizes the tail successfully. Moreover, our model explanation is also more compact and similar to the human saliency map. In the third example, we show a failure of our model: Our model attends to the feet instead of beak which causes the misclassification of a Caspian Tern as an Elegant Tern. Although our model aligns with the human attention, it puts more weight on the feet of birds, since the color of feet is an important feature for distinguishing between a Caspian Tern and a Common Tern (or an Artic Tern).

5.3 Evaluation on CXR-Eye

Comparison with state-of-the-art. The state-of-the-art work on CXR-Eye [18] uses the Efficient-b5 [43] as the classifier, however, it deploys random splits to create training, validation and test sets. For a fair comparison, we re-run its network using our 5-fold cross validation setting and report the average of five validation accuracies as the score for this method. The result of this baseline is 70.97%. When implementing GAT, the result is improved to 71.86%; when implementing KFN, the accuracy is improved by 3.45% to 74.42%.

The full model (GAT+KFN) achieves 75.35% exceeding Efficient-b5 [18] by 4.38%. When comparing the performance boost from GAT and KFN, the KFN improves the model on CUB more than GAT. The reason for the difference is how the gaze data is collected.

In CXR-Eye, the gaze data of the radiologist is collected in an interpretation routine. From the examples shown in Figure 5 (sec. column), we see that fixations spread over many locations (light blue area). These locations may play an important role in diagnoses, but GAT localizes the area that the radiologist fixates for relatively longer time. KFN can integrate the knowledge of all potential locations therefore improves the performance by a larger margin.

Qualitative results. To study the influence of integrating HA into the network, we compare the model explanation (Grad-CAM [35]) of each branch in KFN and the qualitative results are shown in Figure 5. From the figure, we see that the HA branch follows more the human attention while the image branch is focusing different areas.

In the first example (top), human attention focuses more on the left side than the right and the HA branch also does, while the image branch looks more on the right side. The image branch in the second example concentrates on a wrong area, but the HA branch corrects the attentive area to the right. Therefore, KFN improves the performance compared to a model only using images. Most importantly, incorporating gaze knowledge helps to increase the trust and acceptance of the model-based decision in applications such as medical diagnostics, since the model aligns with human behavior.

Figure 5: Illustration of the influence of using HA in model explanation. **Left to Right:** the original Chest X-ray image; HA saliency map; Model explanation of the Image Branch (w/o HA knowledge) and Model explanation of the HA Branch.

6 Conclusion

In this work, we investigate human attention in classification tasks on the CUB and CXR datasets. In particular, we collect a new gaze dataset, CUB-GHA, and show that human attention focuses on the discriminative regions for a fine-grained classification task. To study the hypothesis that human attention helps a model in the decision-making, we propose the Gaze Augmentation Training and Knowledge Fusion Network which integrate human attention knowledge into the network. Our proposed method improves the accuracy in classification by a large margin on both datasets, showing the general validity of our methods. Thus, our work indicates that human attention provides hints on distinct features in different classification tasks.

The aim of our work is to demonstrate the potential benefit of human gaze data in classification. As a by-product of this work, we provide the research community with a gaze-enriched dataset CUB-GHA, which can be incorporated with other existing comprehensive annotations (textual explanations, attributes and bounding boxes, etc.). Researchers can therefore validate multiple applications, where human gaze is required in the interaction with a machine.

7 Acknowledgement

This work has been partially funded by the ERC (853489 - DEXIM) and by the DFG (2064/1 – Project number 390727645). The authors thank all participants who contributed to the CUB-GHA dataset.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2015.

[2] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *ECCV*, 2018.

[3] Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *HRI*, 2018.

[4] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *ITSM*, 2017.

[5] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ETRA*, 2020.

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *NeuIPs*, 2019.

[7] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 2017.

[8] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019.

[9] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeuIPs*, 2018.

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[12] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *NeuIPs*, 2019.

[13] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *AAAI*, 2021.

[14] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *CVPR*, 2020.

[15] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[16] Atsushi Kanehira and Tatsuya Harada. Learning to explain with complemental examples. In *CVPR*, 2019.

[17] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy Wu, Matthew Tong, Arjun Sharma, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth Krupinski, and Mehdi Moradi. Eye Gaze Data for Chest X-rays (version 1.0.0), 2020. URL https://physionet.org/content/egd-cxr/1.0.0/.

[18] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 2021.

[19] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *CVPR*, July 2017.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[21] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *Journal of Vision*, 2016.

[22] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *ICCV*, 2017.

[23] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[24] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*, 2017.

[25] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019.

[26] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 2014.

[27] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *TPAMI*, 2014.

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NeuIPs*, 2014.

[29] Anneli Olsen. The tobii i-vt fixation filter. *Tobii Technology*, 2012.

[30] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

[31] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual review of neuroscience*, 1990.

[32] LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 2020.

[33] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[34] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, 2006.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[36] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *ACL*, 2020.

[37] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. In *ICLRW*, 2015.

[38] Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *ICRA*, 2019.

[39] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*, 2015.

[40] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*. PMLR, 2017.

[41] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018.

[42] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*. PMLR, 2019.

[44] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[45] Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling location proposals of unknown objects through gaze information for human-robot interaction. In *IROS*, 2020.

[46] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[47] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *WACV*, 2020.

[48] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[49] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeuIPs*. Curran Associates, Inc., 2020.

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[51] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MMM*. Springer, 2021.

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[53] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017.

[54] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI*, volume 2020, 2020.

[55] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017.

[56] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, 2019.

[57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[58] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, 2020.

Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring

Yao Rong
Human-Computer Interaction
University of Tübingen
yao.rong@uni-tuebingen.de

Zeynep Akata
Explainable Machine Learning
University of Tübingen
zeynep.akata@uni-tuebingen.de

Enkelejda Kasneci
Human-Computer Interaction
University of Tübingen
enkelejda.kasneci@uni-tuebingen.de

***Abstract*— Numerous car accidents are caused by improper driving maneuvers. Serious injuries are however avoidable, if such driving maneuvers are detected beforehand and the driver is assisted accordingly. In fact, various recent research has focused on the automated prediction of driving maneuver based on hand-crafted features extracted mainly from in-cabin driver videos. Since the outside view from the traffic scene may also contain informative features for driving maneuver prediction, we present a framework for the detection of the drivers' intention based on both in-cabin and traffic scene videos. More specifically, we (1) propose a Convolutional-LSTM (ConvLSTM)-based auto-encoder to extract motion features from the out-cabin traffic, (2) train a classifier which considers motions from both in- and outside of the cabin jointly for maneuver intention anticipation, (3) experimentally prove that the in- and outside image features have complementary information. Our evaluation based on the publicly available dataset Brain4cars shows that our framework achieves a prediction with the accuracy of 83.98% and F_1-score of 84.3%.**

I. INTRODUCTION

According to the World Health Organization [2], about 1.35 million people die in car accidents every year worldwide. These statistics, however, do not include non-fatal injuries from traffic accidents. Most of these accidents are caused by improper driver behavior: Based on the statistics from the Department for Transport (DfT) in Great Britain, a survey [6] revealed that there were 15,560 accidents reported due to poor turn or maneuver, which ranked top 5 in causes of road accidents in 2017. As automated vehicle technology emerges, it promised to be safer than human driving [3], [4], [5]. However, there is still much research to be conducted in order to reach to the fully automated level working at any possible traffic situation and weather conditions. On the half way to autonomous driving vehicles, it is therefore necessary to provide already existing Advanced Driver Assistance Systems (ADAS) the functionality for collaboration with the human driver in the most efficient way, for example to alert the driver in case of a dangerous maneuver.

Recently, many researchers focused on detecting maneuver intention of the driver before execution. For example, Brain4cars [1] and Honda Research Institute Driving Dataset (HDD) [7] are two datasets specifically designed for learning driver behaviors. HDD for example [7] uses three high-resolution video cameras, GPS, signals from LiDAR sensor and vehicle CAN-Bus to record the traffic scenes. Brain4cars [1] provides videos from inside and outside of the car. GPS and vehicle dynamics are also recorded with the videos. These videos show different behavior patterns of maneuvers from driver side and road traffic. Images convey massive information, and much of the literature shows the possibility to predict driver intention according to the drivers' videos, since the drivers turn their heads to glance in the side mirrors. Previous work based on the Brain4cars dataset, such as [1], [9], [10], [11], [12], have all achieved maneuver prediction. Although the reported results are quite impressive, there are still some issues that deserve scrutiny.

More specifically, most of the previous works in the driver maneuver prediction domain mainly use videos from driver observation. Various research has shown that driver behavior, and especially eye movements of the driver, can not only be used for activity recognition [27], [28] but also to ensure safe take-over behavior in conditionally autonomous driving [29]. Additionally, video frames of driver observations are used to extract features e.g. head postures [1], [9], [10], [12]. However, in these works, the traffic information is manually encoded into a vector with four elements, where the first two Boolean values indicate whether a lane exists on the right or left side of the vehicle, the third bit (also Boolean) implies if an intersection or turn exists in 15 meters, and the last value represents the current speed of the car. Therefore, video information of the outside view is not further processed. In addition, manual encoding as employed so far is not applicable to practical use-cases. (2) [11] proposes using two 3D ResNet-101 models for two streams separately. However, it shows that using only driver videos works better than using both video streams. The reason behind this poor performance of outside videos is that there is no large dataset for on-road traffic training, which makes training with the Brain4cars dataset from scratch very difficult. In contrast, for driver observation videos, there is large human activity dataset available such as Kinetics [14].

Intuitively, the outside video, i.e., the scene perspective, should be very informative and provide information that the inside video does not convey. Therefore, our work aims (1) extracting the vehicle motion information from the traffic videos effectively and improving the results which only used one video stream; (2) proposing an end-to-end method without using manual encoding information, and (3) keeping the model as light-weighted (less parameters) as possible to offer applicability to resource-limited mobile platforms.

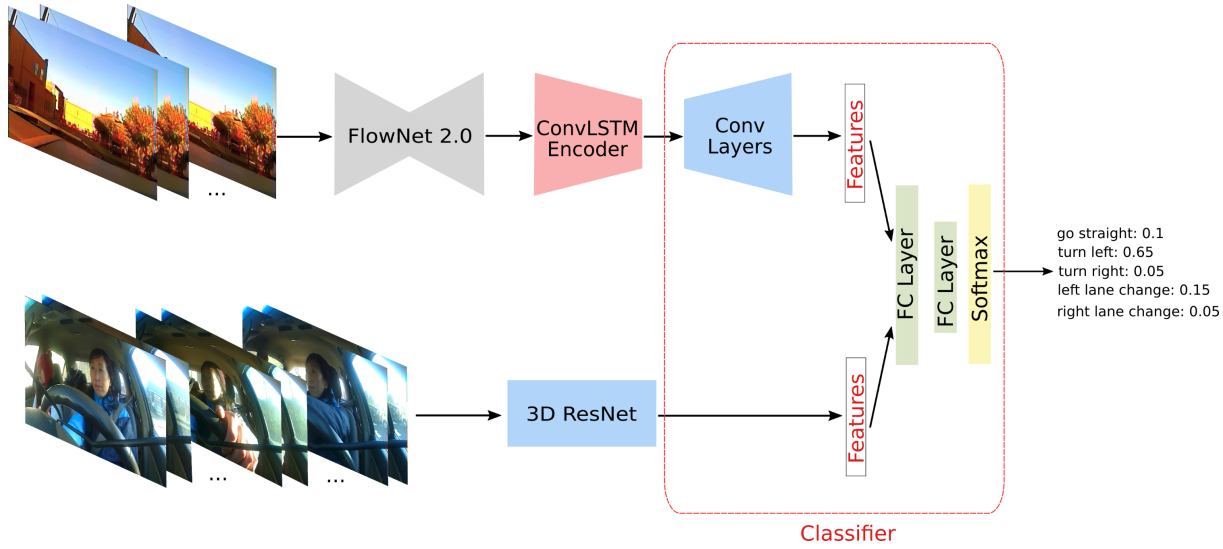


Fig. 1: The overview of our framework. The upper branch depicts the feature extraction from out-cabin videos: FlowNet 2.0 extracts the optical flow from the consecutive frames; then the traffic motion is captured by a ConvLSTM-based encoder. The bottom branch represents the feature extraction from in-cabin videos based on the 3D ResNet-50 network. The red frame in the end refers to the classifier, where a decoder (marked as "Conv Layers") for outside features is integrated. This novel classifier architecture allows features from in- and outside of the cabin to be considered jointly.

To approach these aims, we propose a deep learning framework, which combines the information from the driver monitoring videos with the outside view. This framework is shown in Fig. 1. In our framework, a ConvLSTM [8] based encoder (shown in upper branch) extracts the motion information, which is interpreted in optical flow images. Meanwhile, the 3D ResNet-50 (shown in bottom branch) acquires features from the driver video. The motion decoder for outside motion features is integrated in the classifier. This novel classifier leverages features from both sides, i.e., driver and scene, jointly to produce a maneuver anticipation.

The contribution of our work is manifold: (1) we encode the traffic scene motion using a ConvLSTM-based auto-encoder, (2) propose a deep net framework investigating features from two incoming streams (in- and outsides) jointly, without using any manual-encoded or hand-crafted information, (3) achieve a state-of-the-art maneuver anticipation performance with less parameters compared to the previous work [11], and (4) experimentally validate that the in- and outside videos contain complementary information.

The remaining of this paper is organized as follows: In Section II, we first discuss related works. Our proposed methods and modules mentioned in Fig. 1 are explained in detail in Section III. In Section IV, we introduce the dataset used for training and evaluation of our method and discuss our evaluation results. Finally, we summarize our main findings and conclude this paper.

II. RELATED WORK

Maneuver intention can be detected from drivers' behaviors, such as looking at the outside mirrors or out of the windows. Therefore, popular methods from the domain of human action recognition are suitable and have been applied to tackle this challenge. An action consists of spatial and temporal information. As widely known, features in the spatial domain can be captured by Deep Convolutional Neural Networks (CNNs), while Recurrent Neural Network (RNN) architectures and Long Short-Term Memory (LSTM) cells are well-known for comprehending the logic hidden in time series. LSTM and RNN techniques are therefore often combined with 2D CNNs in video processing applications to deal with both spatial and temporal information, for example as in [8].The formulation from [8] is shown in Eq. 1 with a minor modification, since it contains no bias component.

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \cdot c_{t-1}) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \cdot c_{t-1}) \\
g_t &= tanh(W_{xc} * x_t + W_{hc} * h_{t-1}) \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \cdot c_t) \\
h_t &= o_t \cdot \tanh(c_t)
\end{aligned}
\tag{1}
$$

In the above Eq. 1, subscript t implies the time sequence. x_t is the input. i_t, g_t, f_t and o_t are the gates in the cell. c_t is the cell state and h_t is the hidden state. All the Ws refer to the weights in a convolutional operation. $*$ denotes the convolution operation, while $\cdot$ refers to the element-wise multiplication. σ and $\tanh$ are sigmoid and hyperbolic tangent functions, respectively, which are also applied element-wise. The features learned by ConvLSTM can be used for regression or classification problems. For instance, the authors from [8] built an encoding-forecasting structure to predict the future frame using ConvLSTM cells.

One essential element of video analyzing is motion understanding. Motion describes changes in both temporal and spatial spaces and is often estimated on an image plane based on the optical flow. This technique has been researched for decades since [16]. It calculates the motion of individual

pixels in consecutive frames, which can be then aggregated to interpret the motion of objects. Optical flow is for example widely used in automobile applications [13], since it serves as an extra feature. The extraction of optical flow has been regarded as an optimization problem in the past with various approaches for optical flow estimation such as energy-based method [17], or region-based matching [18]. However, with the rapid development of deep learning, CNN-based networks achieved very impressive results. [19], [20] are only two representative networks for this problem performing in an end-to-end style, where the networks take two consecutive frames as input and output the optical flow.

As previously mentioned, there are multiple works aiming at the driver maneuver anticipation [1], [9], [10], [11], [12]. However, none of the previous work solved driver intention prediction with information from both video (in and out of the car) streams, since the traffic on road is too complex for hand-crafting explicit features. Therefore, several works, such as [1], [9], [10], [12], use manual-encoded feature vectors. On the other hand, training CNNs with outside videos in an end-to-end fashion did not show satisfactory results [11], since there was not enough on-road video data related to maneuver anticipation for training a CNN-based deep network.

In contrast to the above mentioned approaches, we propose to use the outside video stream and the driver observation data jointly for intention anticipation. In the following sections, we introduce our method that leverages information from both videos towards an accurate intention anticipation.

III. METHODOLOGY

A. Future Frame Prediction

Based on ConvLSTM, we propose a network trained in an encoder-decoder manner for motion prediction and feature extraction. Due to its inherent convolutional capability, this structure is able to tackle the spatio-temporal sequence forecasting problem [8]. The details of this architecture are shown in Fig. 2. $h_{i,j}$ is the hidden state and $c_{i,j}$ is the cell state. The subscript i denotes the time step and j indicates the layer number. All the states with $i = 0$ are initialized by the network at the beginning.

The input is a clip of five optical flow images X_i ($i < 5$, $i \in \mathbb{Z}$). The rationale for choosing five as the input length is to gain an uniformly sampled clip for one second (30 frames) up to five second (150 frames). More specifically, "uniformly" means that the interval L between each input is equal. The output of the decoder is the predicted frame in the L-frame future. The decoder is in fact a point-wise convolutional layer here, which differs our architecture from other previous work [8], [26]. In this way, motion information of the five-frame input, which can be used for future motion prediction, is compacted by the encoder. The encoder is regarded as the motion feature extractor, thus, the role of the decoder should be weakened.

The convolution information of the network is shown as in Table. I. In the third column, the size of the output of every layer is shown. The size has four dimensions: the first dimension is the time step; the second one is the channel number, and the last two refer to the height and width of the input image, respectively. Every ConvLSTM cell takes one frame at one time step, so the first dimension changes to one after the input layer. Additionally, it is worth mentioning that the output from the encoder is the feature needed for maneuver anticipation.

TABLE I: The convolution information about the future motion prediction module

Layer	Kernel Size / Stride	Output size
Input		5×3×h×w
Layer 0	(3,3)/(1,1)	1×128×h×w
Layer 1	(3,3)/(1,1)	1×64×h×w
Layer 2	(3,3)/(1,1)	1×64×h×w
Layer 3	(3,3)/(1,1)	1×32×h×w
Conv	(1,1)/(1,1)	1×3×h×w

B. Maneuver Anticipation Framework

The proposed method makes use of two input sources: inside and outside videos, as shown in Fig. 1. For the traffic videos, the FlowNet 2.0 first takes original frames to produce optical flow images. Then, the optical flow images are fed into the ConvLSTM encoder described in the last section. The output from the encoder is then the 3D dimension feature (32×112×176), which will be processed by multiple convolutional blocks (Conv-Block) before fusion. At the same time, the other branch, a 3D ResNet-50, deals with the

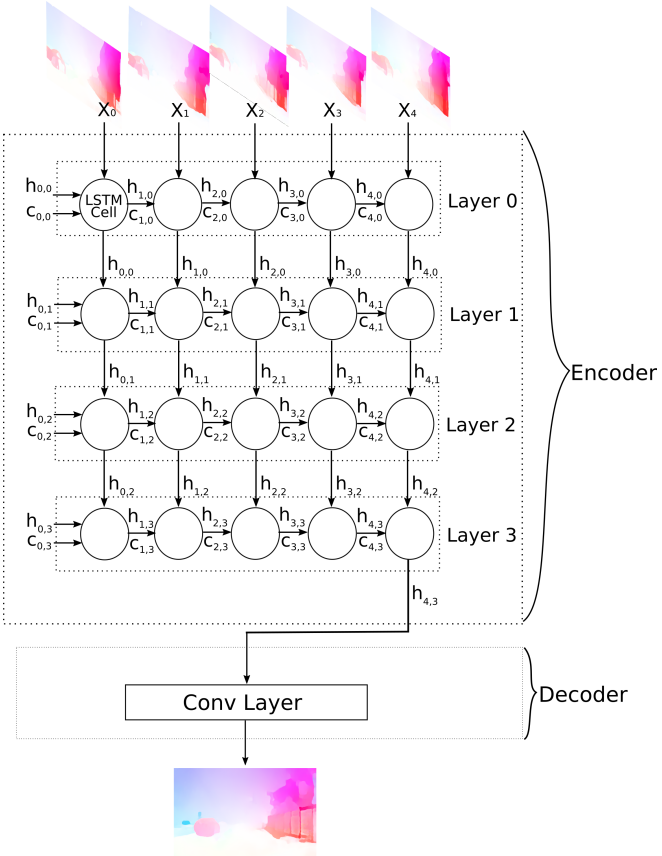


Fig. 2: Architecture of the proposed future motion prediction module.

TABLE II: The architecture of the proposed classifier, which considers joint features from in- and outside videos. The first column indicates the feature source, the second column shows the name of the layer, and the third column is the output size after the layer. The features are combined in the "Concatenate" layer.

Feature	Layer	Output size
Outside	Conv-Block 0	64×37×59
	Conv-Block 1	128×12×20
	Conv-Block 2	256×4×7
	Conv-Block 3	512×1×2
	Concatenate	3072×1
Both	FC 0	3072×2048
Both	FC 1	2048× N_{cls}
Both	Softmax	N_{cls}

driver videos. The main body is consistent with the original network in [15]. Additionally, we added a dropout layer after the average pooling layer in the end to prevent overfitting. The feature we extracted is the input of the last FC layer in ResNet-50, which is a 2048-dimension vector. The input of the ResNet-50 is a 16-frame clip.

The novelty of the proposed classifier is that the decoder for outside features is trained jointly with features of inside videos. Its explicit structure is listed in Table. II. The Conv-Block is for decoding the outside motion. The structure inside one Conv-Block is shown in Fig. 3, where "ReLU" refers to the activation function and "BN" represents the Batch Normalization (BN) layer. There is also a ReLU and a BN between the last two FC layers. The output size after every layer is shown in the third column. In the end, N_{cls} represents the number of classes, which is five in our case.

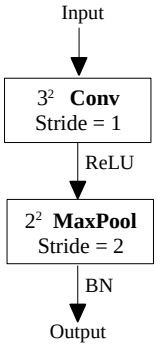


Fig. 3: The architecture inside "Conv-Block"

IV. RESULTS AND DISCUSSIONS

A. Dataset

The Brain4Cars [1] dataset includes driver observation videos (1088px × 1920px, 25 fps) and videos of the outside scenes (480px × 720px, 30 fps) recorded simultaneously. There are five classes of maneuvers in the dataset: *go straight, left lane change, left turn, right lane change, right turn*.

According to the Brain4cars dataset, the video covers the behavior before the actual maneuver occurs, i.e., no maneuver is performed during the video. In this work, we also study the early detection capability of our models. Therefore, we take every second as a dividing line. In the model evaluation, we give the frames before time step T, here $T \in (-5, -4, -3, -2, -1)$. The $-$ represents the time (in second) before the maneuver happens. The shorter videos cover a shorter time period before the maneuver starts. Since the videos have different lengths, we have different amount of input material when we study early prediction. Moreover, samples with no simultaneous recordings of the inside and outside view are considered as invalid and not further used in our study. The number of valid video samples for training the whole framework relatively to the covered time period before a maneuver is shown in Table III.

TABLE III: The number of the valid samples relatively to the video length

video length [s]	> 4	> 3	> 2	> 1	> 0
samples	490	542	563	573	585

We use a 5-fold cross-validation for all the experiments in this work, which also aligns with other previous works using the Brain4cars dataset [1], [9], [10], [11], [12].

B. Out-cabin Motion Extraction

For the outside motion feature extraction, we trained the encoder/decoder module presented in III-A. To achieve a generalized solution, we added a temporal augmentation in training: a 5-frame clip is randomly and uniformly cut and given as the input to the network. The target is the L-th frame after the last one in the clip. In the spatial domain, they are first resized to a smaller size (112×176), yet keeping the original scale. Additionally, we employ the Mean Square Error (MSE) as the loss function and Stochastic Gradient Descent (SGD) as the optimizer. The weight decay is set to 0.001 and momentum to 0.9. The whole training takes 60 epochs with the learning rate of 0.1.

For evaluation, we first studied how far into the future the model is able to predict. More specifically, we evaluated our model with respect to the interval of $L \in (5, 10, 15, 20, 25, 30)$ frames. As the output of the decoder is the predicted motion in the L-th frame after the last input, a larger interval represents a further future. The maximal interval value is 30 (requiring thus 150 frames), which reaches the maximal video length (5s) in the dataset. On the other hand, an interval less than 5 frames (0.33s) is too short, and thus not considered here. The target frame is the last frame in the video, whereas the metric for comparison is the MSE. The average MSE with respect to different intervals is shown in the Fig. 4.

Please note that the MSE value is multiplied by 1000 to make the differences more clear. Our results show that it is difficult for the model to predict a far future frame: The model does not learn properly when the interval is larger than 20 frames (0.67s). In order to have relatively precise motion features, we choose the model with L of 5. After setting the interval L to 5, we evaluated our model with regard to different time periods of the video. More specifically,

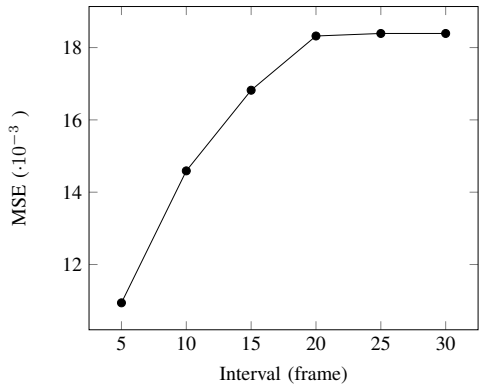


Fig. 4: MSE for different interval values

the input frames are all included in the time period before T ($T \in (-4, -3, -2, -1, 0)$), and the last frame of every second is the target frame. To quantify the comparison between the target and predicted image, we employed three metrics: MSE, Structural Similarity (SSIM) index, and Peak Signal-to-Noise Ratio (PSNR). The results of prediction are shown in Table IV. For the PSNR and SSIM, higher values are better. The results of five folds are shown in the form: "Average (Avg) $\pm$ Standard Error (SE)".

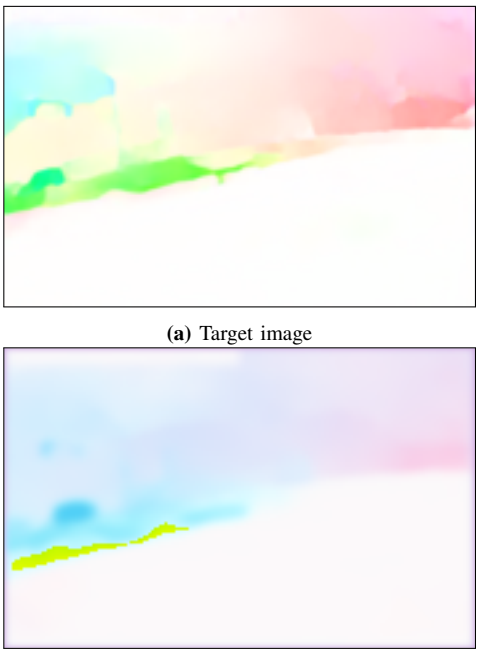
Our results show that the best maneuver prediction is achieved from video information 4 to 5 seconds before the actual maneuver occurs. Thus, motion changes are not massive earlier on before -3 second. In case of large motion changes (e.g., when the car is turning), it is hard for the encoder to catch the whole change. Accordingly, in the third and the last second before a maneuver, the outside motion changes noticeably. However, from $-2s$ to $-1s$, motion keeps changing but not as distinct as its contiguous time steps. In general, the important traffic motion changes can be observed within three seconds before the maneuver, which also corresponds to the early detection results in the Section IV-D, where the encoder was emplyed to extract the outside motion features.

TABLE IV: Results of future motion prediction.

prediction at [s]	MSE ($\cdot 10^{-3}$)	SSIM	PSNR
-4	9.13 ± 0.42	0.909 ± 0.001	21.77 ± 0.16
-3	9.42 ± 0.40	0.906 ± 0.002	21.49 ± 0.10
-2	10.75 ± 0.61	0.904 ± 0.002	21.35 ± 0.18
-1	9.97 ± 0.22	0.900 ± 0.001	21.27 ± 0.05
0	10.73 ± 0.46	0.898 ± 0.002	21.08 ± 0.10

Fig. 5b shows an example of the predicted frame using the proposed encoder/decoder module compared to the target image in 5a. From the visual image results, it is apparent that the major problem is the color disorder. The area in light yellow and the green color is mistaken by light blue in the output. According to the optical flow color coding [21], the direction changes 90 degree (from bottom side to right side) from the light yellow to blue, and the green is in between. This detailed motion is difficult for the encoder to catch.

Using the features extracted from the outside videos by the ConvLSTM-encoder alone can also produce a prediction among five classes. The results are presented in Table V, whereas a comparison to related approaches is provided in Table VI.



(a) Target image

(b) Predicted image

Fig. 5: The comparison of target and the predicted image

C. In-cabin Action Recognition

We employ the 3D ResNet-50 for the inside feature extraction, since the 3D ResNet has shown high performance in human action recognition tasks [15]. However, end-to-end training requires a large amount of the dataset, which is not the case for Brain4cars. Hence, we use the Kinetics-pretrained 3D ResNet-50 [15] and fine-tune the model with Brain4cars inside videos.

To prevent overfitting, we added spatial and temporal data augmentation. With regard to spatial augmentation, we added a random crop (but with the focus on the driver side), a random scale and a horizontal flip. It is worth noticing that the label also needs to change accordingly when it is related to the direction (left/right). For temporal augmentation, we randomly but uniformly cut a short clip from every second. The short clips constitutes a 16-frame clip as the input to the 3D ResNet-50, and the input size is 112×112. One extra dropout layer is added before the last FC layer when training. We use a dropout rate of 0.5 an cross entropy loss as out loss function. The model is trained for 60 epochs, with learning rate starting with 0.1 and a decay rate of 0.1 after the 30th and 50th epoch. The optimizer is the SGD with the momentum and weight decay of 0.9 and 0.001, respectively. In out evaluation, we use the frames from the end of every second before T ($T \in (-4, -3, -2, -1, 0)$) to compose the 16-frame input for the 3D ResNet.

The main body of trained 3D ResNet-50 is used as the feature extractor. The feature before the last FC layer is fed into the final classifier. The results of using only this module (inside video) for classification are shown in Table V, whereas the comparison to related approaches is given in Table VI.

D. Feature Fusion

After training the ConvLSTM model and 3D ResNet-50 model separately, the features from inside and outside video are extracted by the two trained modules. The obtained outside feature is a volume with the shape of $32 \times 112 \times 176$, and the inside feature is a 2048-size vector. They are fed into the classifier introduced in the section III-B. We conducted the evaluation procedure with regard to different time periods as in both modules.

The performance indicators are accuracy and the F_1-score. The F_1-score takes both precision (Pr) and recall (Re) of a classifier into consideration (Eq. 2). n refers to the number of classes, and Ω is the set of all the classes that our model can recognize, which includes four maneuvers plus "no maneuver" class. TP_i indicates the amount of correctly recognized samples of class i. P_i and N_i are the number of samples that are predicted as class i and that are labeled as class i, separately.

$$
\begin{aligned}
Pr &= \frac{1}{n} \sum_{i \in \Omega} \frac{TP_i}{P_i} \\
Re &= \frac{1}{n} \sum_{i \in \Omega} \frac{TP_i}{N_i} \\
F_1 &= \frac{2 \cdot Pr \cdot Re}{Pr + Re}
\end{aligned}
\tag{2}
$$

Table V shows the results of accuracy and F_1 in % for different times before the occurrence of a maneuver using different data sources. Both accuracy and F_1 increase as the time approaches the beginning of maneuver, despite of different data sources. Intuitively, the early stage of all the maneuvers (or no maneuver) is similar, which is "going straight". In this case, the longer period the model observes, the more accurate the decision it can make. According to these results, early detection is possible. For example, 71.72% of the maneuvers are correctly predicted two seconds before the maneuver happens when using both video streams.

The best results are achieved by using both video sources in all different time periods. Only using outside videos gives the worst results when compared to other two data sources. The reason for the poor performance of outside data is that the auto-encoder only provides the motion feature of one future frame. However, the inside feature contains the information over a long time period. Moreover, we can see the decisive motion occurs ordinarily within three seconds before maneuvers. Especially from -4 to -2, the improvement of accuracy and F_1 are substantial.

The inside videos always provide good results, but it is still slightly inferior to the joint two-stream input. It is important to see that outside video feature does not depress the performance of the inside video feature, but improves it. Therefore, the information from both inside and outside videos are complementary. Besides, as the outside video become more informative, its effect is more apparent. The differences of accuracy and F_1 between inside only and both sides increase steadily after -3 seconds. Fig. 7 and Fig. 8 illustrate the differences among using different data sources in relation to various time periods more clearly. Additionally, Fig. 6 shows the confusion matrix of three models using different data sources. Prediction is made based on time period [-5,0]. From this, an improvement of all classes can be observed when using two video streams.

TABLE V: The results of using proposed framework with different input data sources. The results of five folds are shown in the form: "Avg ± SE".

Inside video	Time period	Acc (%)	F_1 (%)
	[-5,-4]	56.49 ±0.02	48.19 ±0.03
	[-5,-3]	63.63 ±0.02	58.46 ±0.02
	[-5,-2]	70.48 ±0.02	68.63 ±0.03
	[-5,-1]	75.73 ±0.01	73.09 ±0.01
	[-5,0]	77.40 ±0.02	75.49 ±0.02
Outside video	Time period	Acc (%)	F_1 (%)
	[-5,-4]	44.08 ±0.01	38.91 ±0.03
	[-5,-3]	44.22 ±0.01	38.75 ±0.01
	[-5,-2]	50.43 ±0.01	46.98 ±0.01
	[-5,-1]	59.53 ±0.01	62.37 ±0.01
	[-5,0]	60.87 ±0.01	66.38 ±0.03
In- & outside	Time period	Acc (%)	F_1 (%)
	[-5,-4]	59.13 ±0.02	53.35 ±0.02
	[-5,-3]	64.93 ±0.02	60.33 ±0.01
	[-5,-2]	72.07 ±0.02	70.56 ±0.02
	[-5,-1]	79.92 ±0.02	78.90 ±0.01
	[-5,0]	83.98 ±0.01	84.30 ±0.01

We compare our results with the ones from work [11] in Table. VI, since we all use the end-to-end training and investigate the performance with three different data sources. We compare the accuracy, F_1 and the number of parameters of our models. The results listed here are all from zero time-to-maneuver and in 5-fold cross-validation.

Our model surpasses the model in [11] except using only inside view. It is because the 3D ResNet-101 is used in [11], which has almost two times more parameters than 3D ResNet-50 in our work. We choose to use a smaller ResNet in order to avoid outfitting problems when fine tuning a very large network with a small dataset. Moreover, a low resource-cost model is preferable for automobile applications. Our framework outperforms the previous work with much less parameters in using two-stream input: It achieves 83.98% of accuracy and 84.30% of F_1 averagely within five folds, surpassing the previous work by 8.48 percentage points in accuracy and 11.1 percentage points in F_1. When only con-

	Straight	L Lane	L Turn	R Lane	R Turn
Straight	.80	.07	.02	.07	.04
L Lane	.12	.71	.04	.08	.05
L Turn	.02	.09	.77	.02	.10
R Lane	.07	.07		.80	.06
R Turn	.04	.13	.04	.07	.72

(a) Inside videos

	Straight	L Lane	L Turn	R Lane	R Turn
Straight	.64	.19		.17	
L Lane	.31	.55		.12	.02
L Turn	.02	.05	.88	.03	.02
R Lane	.35	.23	.02	.37	.03
R Turn		.04	.02	.05	.89

(b) Outside videos

	Straight	L Lane	L Turn	R Lane	R Turn
Straight	.87	.05	.01	.05	.02
L Lane	.12	.75	.03	.07	.03
L Turn		.05	.88	.04	.03
R Lane	.08	.06	.01	.80	.05
R Turn	.02	.02		.02	.94

(c) In and outside videos

Fig. 6: The confusion matrix of using different video streams. The prediction is made at the last second before the occurrence of a maneuvers.

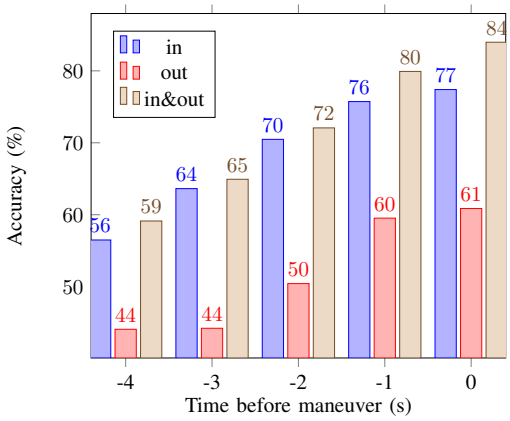


Fig. 7: Accuracy: comparison using different data sources.

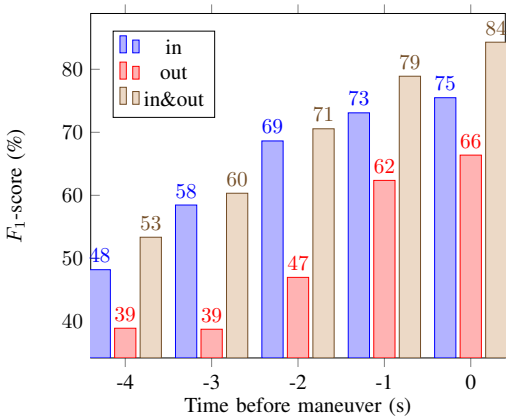


Fig. 8: F_1-score: comparison using different data sources.

sidering outside videos, our models surpasses theirs by 7.67 percentage points and 22.98 percentage points in accuracy and F_1, respectively. It achieves to extract useful features from outside with much less parameters. More importantly, our model does not confront the same problem that the outside videos weaken the classifier performance. In other words, our results show that the information from outside videos are also valuable.

We also conduct an experiment using similar threshold policy as in [1], [9] on our model which uses two-stream video: If the probability is NOT greater than the threshold, then *"go straight"* is predicted. As shown in Fig. 9, the performance gets worse when this threshold is larger than

TABLE VI: Comparison of our proposed framework with other method. The results of five folds are shown in the form: "Avg $\pm$ SE". In order to show a clear difference, we use "m" to represent the number of parameters in FlowNet2.0, which is a common module in both methods.

Method	Data Source	Acc (%)	F_1 (%)	Param.(M)
[11]	inside only	83.1 $\pm$2.5	81.7 $\pm$2.6	85.26+m
	outside only	53.2 $\pm$0.5	43.4 $\pm$0.9	85.26+m
	in-&out-side	75.5 $\pm$2.4	73.2 $\pm$2.2	170.52+m
our	inside only	77.40 $\pm$0.02	75.49 $\pm$0.02	46.22
	outside only	60.87 $\pm$0.01	66.38 $\pm$0.03	5.41+m
	in-&outside	**83.98 $\pm$0.01**	**84.30 $\pm$0.01**	**57.92+m**

0.4 in all lengths of input videos, since the model is trained on a balanced loss function and learns motion features of all five maneuvers. It always gives a relatively confident prediction with a probability over 0.4. For our model, no threshold policy is necessary.

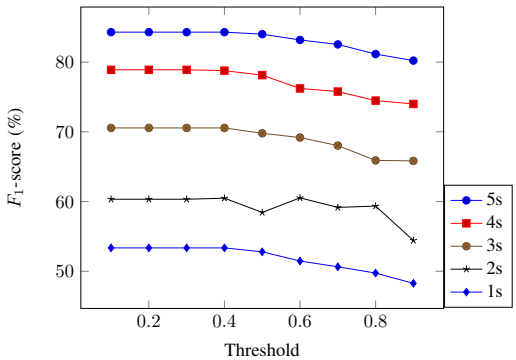


Fig. 9: Effect of using thresholds. Two-stream input with different video lengths (from 1 to 5 seconds).

V. CONCLUSION AND FUTURE WORK

In this work, we propose a framework that considers both inside and outside cabin motion features to anticipate the driver maneuver intention. We propose to extract the outside traffic motion using a ConvLSTM-based auto-encoder. These motion features are decoded by a novel classifier architecture, which considers the in- and outside motions jointly. Our model is trained in end-to-end style, without using

any manual-encoded or hand-crafted features. Our results show that dual input (driver observation and driving scene videos) surpasses by far related approaches based on single input analyses. Additionally, we validate experimentally that both inside and outside videos convey valuable and complementary information. This conclusion suggests that both traffic scenes and driver behaviors should be taken into consideration when anticipating maneuver intention.

For our future work, we plan to improve the performance of the outside motion decoder in the classifier by training a more delicate decoder which can interpret the motion covering a longer time period. In this way, the module would gain a perspective of the entire outside motion. Moreover, accurately predicting the motion of the further future is another aim for our future work.

ACKNOWLEDGMENT

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-Number 2064/1 – Project number 390727645.

REFERENCES

[1] Jain, Ashesh and Koppula, Hema S and Raghavan, Bharad and Soh, Shane and Saxena, Ashutosh, *Car that knows before you do: Anticipating maneuvers via learning temporal driving models*, Proceedings of the IEEE International Conference on Computer Vision, pages 3182–3190, 2015

[2] *Road traffic injuries*, February 7. 2020. Accessed on: Feb. 9, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Morando, Mark Mario and Tian, Qingyun and Truong, Long T and Vu, Hai L, *Studying the safety impact of autonomous vehicles using simulation-based surrogate safety measures*, Journal of advanced transportation, vol.2018, 2018, Hindawi

[4] Fox, M, *Self-driving cars safer than those driven by humans: Bob Lutz*, CNBC, [Online]. Available: www. cnbc. com, 2014

[5] Teoh, Eric R and Kidd, David G, *Rage against the machine? Google's self-driving cars versus human drivers*, Journal of safety research, vol.63, page 57–60, 2017, Elsevier

[6] *Most Common Causes for Road Accidents in Britain Revealed*, July 2. 2018. Accessed on: Feb.9,2020. [Online]. Available: https://www.regtransfers.co.uk/content/common-causes-for-road-accidents-in-britain/

[7] Ramanishka, Vasili and Chen, Yi-Ting and Misu, Teruhisa and Saenko, Kate, *Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, page 7699–7707, 2018

[8] Xingjian, SHI and Chen, Zhourong and Wang, Hao and Yeung, Dit-Yan and Wong, Wai-Kin and Woo, Wang-chun, *Convolutional LSTM network: A machine learning approach for precipitation nowcasting*, Advances in neural information processing systems, page 802–810, 2015

[9] Jain, Ashesh and Soh, Shane and Raghavan, Bharad and Singh, Avi and Koppula, Hema S and Saxena, Ashutosh, *Brain4Cars: Sensory-Fusion Recurrent Neural Models for Driver Activity Anticipation*

[10] Zhou, Dong and Ma, Huimin and Dong, Yuhan, *Driving maneuvers prediction based on cognition-driven and data-driven method*, 2018 IEEE Visual Communications and Image Processing (VCIP), page 1–4, 2018, IEEE

[11] Gebert, Patrick and Roitberg, Alina and Haurilet, Monica and Stiefelhagen, Rainer, *End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks*, 2019 IEEE Intelligent Vehicles Symposium (IV), page 969–974, 2019, IEEE

[12] Tonutti, Michele and Ruffaldi, Emanuele and Cattaneo, Alessandro and Avizzano, Carlo Alberto, *Robust and subject-independent driving manoeuvre anticipation through Domain-Adversarial Recurrent Neural Networks*, Robotics and Autonomous Systems, vol.115, page 162–173, 2019, Elsevier

[13] Menze, Moritz and Geiger, Andreas, *Object scene flow for autonomous vehicles*, Proceedings of the IEEE conference on computer vision and pattern recognition, page 3061–3070, 2015

[14] Carreira, Joao and Zisserman, Andrew, *Quo vadis, action recognition? a new model and the kinetics dataset*, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, page 6299–6308, 2017

[15] Hara, Kensho and Kataoka, Hirokatsu and Satoh, Yutaka, *Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?*, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, page 6546–6555, 2018

[16] Horn, Berthold KP and Schunck, Brian G, *Determining optical flow*, Techniques and Applications of Image Understanding, vol.281, page 319–331, 1981, International Society for Optics and Photonics

[17] Adelson, Edward H and Bergen, James R, *Spatiotemporal energy models for the perception of motion*, Josa a, vol.2, page 284–299, 1985, Optical Society of America

[18] Anandan, Padmanabhan, *A computational framework and an algorithm for the measurement of visual motion*, International Journal of Computer Vision, vol.2, page 283–310, 1989, Springer

[19] Dosovitskiy, Alexey and Fischer, Philipp and Ilg, Eddy and Hausser, Philip and Hazirbas, Caner and Golkov, Vladimir and Van Der Smagt, Patrick and Cremers, Daniel and Brox, Thomas, *Flownet: Learning optical flow with convolutional networks*, Proceedings of the IEEE international conference on computer vision, page 2758–2766, 2015

[20] Ilg, Eddy and Mayer, Nikolaus and Saikia, Tonmoy and Keuper, Margret and Dosovitskiy, Alexey and Brox, Thomas, *Flownet 2.0: Evolution of optical flow estimation with deep networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, page 2462–2470, 2017

[21] Baker, Simon and Scharstein, Daniel and Lewis, JP and Roth, Stefan and Black, Michael J and Szeliski, Richard, *A database and evaluation methodology for optical flow*, International journal of computer vision, vol.92, page 1–31, 2011, Springer

[22] Zhou, Bolei and Khosla, Aditya and Lapedriza, Agata and Oliva, Aude and Torralba, Antonio, *Learning deep features for discriminative localization*, Proceedings of the IEEE conference on computer vision and pattern recognition, page 2921–2929, 2016

[23] Mikolov, Tomáš and Karafiát, Martin and Burget, Lukáš and Černockỳ, Jan and Khudanpur, Sanjeev, *Recurrent neural network based language model*,Eleventh annual conference of the international speech communication association, 2010

[24] Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V, *Sequence to sequence learning with neural networks*, Advances in neural information processing systems, page 3104–3112, 2014

[25] Wu, Zuxuan and Wang, Xi and Jiang, Yu-Gang and Ye, Hao and Xue, Xiangyang, *Modeling spatial-temporal clues in a hybrid deep learning framework for video classification*, Proceedings of the 23rd ACM international conference on Multimedia, page 461–470, 2015

[26] Srivastava, Nitish and Mansimov, Elman and Salakhudinov, Ruslan, *Unsupervised learning of video representations using lstms*, International conference on machine learning, page 843–852, 2015

[27] Braunagel, Christian and Kasneci, Enkelejda and Stolzmann, Wolfgang and Rosenstiel, Wolfgang, *Driver-activity recognition in the context of conditionally autonomous driving*, 2015 IEEE 18th International Conference on Intelligent Transportation Systems, page 1652–1657, 2015

[28] Braunagel, Christian and Geisler, David and Rosenstiel, Wolfgang and Kasneci, Enkelejda, *Online recognition of driver-activity based on visual scanpath classification*, IEEE Intelligent Transportation Systems Magazine, 9 (4), page 23–36, 2017

[29] Braunagel, Christian and Rosenstiel, Wolfgang and Kasneci, Enkelejda, *Ready for take-over? A new driver assistance system for an automated classification of driver take-over readiness*, IEEE Intelligent Transportation Systems Magazine, 9:(4), page 10–22, 2017

Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring

Conference Proceedings:
2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)

Author: Yao Rong; Zeynep Akata; Enkelejda Kasneci

Publisher: IEEE

Date: 20-23 Sept. 2020

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Where and What: Driver Attention-based Object Detection

YAO RONG, University of Tübingen, Germany
NAEMI-REBECCA KASSAUTZKI, University of Tübingen, Germany
WOLFGANG FUHL, University of Tübingen, Germany
ENKELEJDA KASNECI, University of Tübingen, Germany

Human drivers use their attentional mechanisms to focus on critical objects and make decisions while driving. As human attention can be revealed from gaze data, capturing and analyzing gaze information has emerged in recent years to benefit autonomous driving technology. Previous works in this context have primarily aimed at predicting "where" human drivers look at and lack knowledge of "what" objects drivers focus on. Our work bridges the gap between pixel-level and object-level attention prediction. Specifically, we propose to integrate an attention prediction module into a pretrained object detection framework and predict the attention in a grid-based style. Furthermore, critical objects are recognized based on predicted attended-to areas. We evaluate our proposed method on two driver attention datasets, BDD-A and DR(eye)VE. Our framework achieves competitive state-of-the-art performance in the attention prediction on both pixel-level and object-level but is far more efficient (75.3 GFLOPs less) in computation.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Computer vision**; • **Human-centered computing**;

Additional Key Words and Phrases: deep learning, gaze prediction, eye tracking, object detection, driver attention, gaze mapping

ACM Reference Format:
Yao Rong, Naemi-Rebecca Kassautzki, Wolfgang Fuhl, and Enkelejda Kasneci. 2022. Where and What: Driver Attention-based Object Detection. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA, Article 146 (May 2022), 22 pages. https://doi.org/10.1145/3530887

1 INTRODUCTION

Human attentional mechanisms play an important role in selecting task-relevant objects effectively in a top-down manner, which can solve the task efficiently [36, 39, 49]. To visualize human attention for these tasks in a general way, a Gaussian filter is applied on fixation points to form a *saliency* map [23], thus highlighting the visual attention area. Due to the effectiveness and irreplaceability of human attention in solving visual tasks, visual attention is also being studied in artificial intelligence research (e.g., [57]). Many computer vision applications embrace human gaze information, for instance in classification tasks [28, 41], computer-aided medical diagnosis systems [16, 42], or important objects selection/cropping in images and videos [43, 44, 50, 52]. To better understand how the human brain processes visual stimuli, knowing not only *where* humans are looking at, but

Authors' addresses: Yao Rong, yao.rong@uni-tuebingen.de, University of Tübingen, Sand 14, Tübingen, Germany, 72076; Naemi-Rebecca Kassautzki, University of Tübingen, Sand 14, Tübingen, Germany, 72076, naemi-rebecca.kassautzki@student.uni-tuebingen.de; Wolfgang Fuhl, University of Tübingen, Sand 14, Tübingen, Germany, 72076, wolfgang.fuhl@uni-tuebingen.de; Enkelejda Kasneci, University of Tübingen, Sand 14, Tübingen, Germany, 72076, enkelejda.kasneci@uni-tuebingen.de.

2573-0142/2022/5-ART146 $15.00
https://doi.org/10.1145/3530887

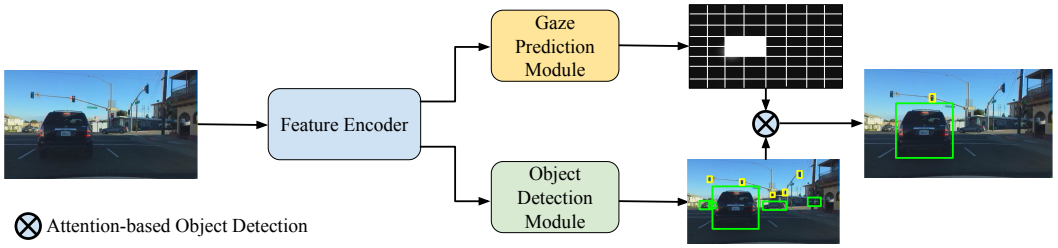


Fig. 1. Overview of our proposed critical object detection framework. The **feature encoder** extracts features from the input image. The **gaze prediction module** predicts driver attention in a grid-based saliency map and the **object detection module** detects all the objects in the traffic using extracted features. The **attention-based objects** are detected and returned to users based on the predicted saliency map and detected objects.

also *what* object is essential, i.e., gaze-object mapping [4]. This mapping is needed in many research projects, especially in analytics of student learning process [21] or human cognitive functions [35].

In autonomous driving applications, successful models should be able to mimic “gaze-object mapping" of humans, which includes two challenges: Driver gaze prediction and linking the gaze to objects. It is practical to predict driver gaze since sometimes no eye tracker is available or no human driver is required in the higher level of autonomous vehicles. For instance, Pomarjanschi et al. [37] validates that highlighting potentially critical objects such as a pedestrian on a head-up display helps to reduce the number of collisions. In this case, a model capable of predicting these critical objects can be used as a “second driver" and give warnings that assist the real driver. For fully autonomous cars, it is essential to identify these task-relevant objects efficiently to make further decisions and also explain them [17]. Recently, there is a growing research interest in predicting human drivers’ gaze-based attention [11, 34, 54]. These existing works predict pixel-level saliency maps, however, they lack semantic meaning of the predicted attention, i.e., the model only predicts *where* drivers pay attention, without knowing *what* objects are inside those areas.

To bridge the research gap between driver gaze prediction and semantic object detection existing in the current research landscape of autonomous driving applications, we propose (1) to predict where and what the drivers look at. Furthermore, we aim (2) at a model that is efficient in computation, since resources on self-driving cars are limited. Specifically, we designed a novel framework for efficient attention-based object detection based on human driver gaze. Our approach provides not only pixel-level attention saliency maps, but also the information of objects appearing in attention areas, as illustrated in Fig. 1. A feature encoder is first used in our framework to encode the information in the input image. Then, the extracted features are used to predict gaze and detect objects in the image at the same time. Since obtaining accurate high-level (object) information is our final goal, instead of low-level (pixel) accuracy in saliency map prediction, we predict salient areas in a grid-based style to save computational costs while still maintaining high performance in the critical object detection task.

Our contributions can be summarized as follows: (1) We propose a framework to predict objects that human drivers pay attention to while driving. (2) Our proposed grid-based attention prediction module is very flexible and can be incorporated with different object detection models. (3) We evaluate our model on two datasets, BDD-A and DR(eye)VE, showing that our model is computationally more efficient and achieves comparable performance in pixel- and object-level prediction compared to other state-of-the-art driver attention models. For the sake of reproducibility, our code is available at https://github.com/yaorong0921/driver-gaze-yolov5.

2 RELATED WORK

In the following, we first discuss previous works of gaze-object mapping used in applications other than driving scenarios and we discuss the novelty of our proposed method for solving this task. Then, we introduce the related work with a special focus on the driver attention prediction in the context of saliency prediction for human attention, followed by the introduction of several object detectors our framework is based on. Thanks to deep learning techniques, there exists a plethora of works in the past decades for visual saliency models and object detectors (see [6, 58] for review). It is impracticable to thoroughly discuss these works in the two branches, therefore we only present the works which are closely related to our work.

Gaze-Object Mapping. Previous works [20, 53] set out to reduce tedious labelling by using gaze-object mapping, which annotates objects at the fixation level, i.e., the object being looked at. One popular algorithm checks whether a fixation lies in the object bounding box predicted by deep neural network-based object detector [4, 21, 29] such as YOLOv4 [5]. Wolf et al. [53] suggest to use object segmentation using Mask-RCNN [12] as object area detection. These works train their object detectors with limited object data and classes to be annotated. Panetta et al. [35], however, choose to utilize a bag-of-visual-words classification model [9] over deep neural networks for object detection due to insufficient training data. Barz et al. [3] propose a "cropping-classification" procedure, where a small area centered at the fixation is cropped and then classified by a network pretrained on ImageNet [10]. This algorithm from [3] can be used in Augmented Reality settings for cognition-aware mobile user interaction. In the follow-up work [4], the authors compare the mapping algorithms based on image cropping (IC) with object detectors (OD) in metrics such as precision and recall, and the results show that IC achieves higher precision but lower recall scores compared to OD.

However, these previous works are often limited in object classes and cannot be used to detect objects in autonomous driving applications, since a remote eye tracker providing precious fixation estimation is required for detecting attended objects. Unlike previous gaze-object mapping methods, a model in semi-autonomous driving applications should be able to predict fixation by itself, for instance, giving safety hints at critical traffic objects as a "second driver" in case human drivers oversee them. In fully autonomous driving, where no human driver fixation is available, a model should mimic human drivers' fixation. Therefore, our framework aims to showcase a driver attention model achieving predicting gaze and mapping gaze to objects simultaneously, which is more practical in autonomous driving applications.

Gaze-based Driver Attention Prediction. With the fast-growing interest in (semi-)autonomous driving, studying and predicting human drivers' attention is of growing interest. There are now studies showing improvement in simulated driving scenarios by training models in an end-to-end manner using driver gaze, so that models can observe the traffic as human drivers [25, 30]. Based on new created real-world datasets, such as DR(eye)VE [34] and BDD-A [54], a variety of deep neural networks are proposed to predict pixel-wise gaze maps of drivers (e.g., [15, 33, 34, 45, 54]). The DR(eye)VE model [34] uses a multi-branch deep architecture with three different pathways for color, motion and semantics. The BDD-A model [54] deploys the features extracted from AlexNet [19] and inputs them to several convolutional layers followed by a convolutional LSTM model to predict the gaze maps. An attention model is utilized to predict driver saliency maps for making braking decisions in the context of end-to-end driving in [1]. Two other well-performing networks for general saliency prediction are ML-Net [8] and PiCANet [26]. ML-Net extracts features from different levels of a CNN and combines the information obtained in the saliency prediction. PiCANet is a pixel-wise contextual attention network that learns to select informative context locations for

each pixel to produce more accurate saliency maps. In this work, we will also include these two models trained on driver gaze data in comparison to our proposed model. Besides these networks, which are focused on predicting the driver gaze map, other models are extended to predict additional driving-relevant areas. While Deng et al. [11] use a convolutional-deconvolutional neural network (CDNN) and train it on eye tracker data of multiple test person, Pal et al. [33] propose to include distance-based and pedestrian intent-guided semantic information in the ground-truth gaze maps and train models using this ground-truth to enhance the models with semantic knowledge.

Nevertheless, these models cannot provide the information of objects that are inside drivers' attention. It is possible to use the existing networks for detecting attended-to objects, but this would have the disadvantage that predicting gaze maps on pixel-level introduces unnecessary computational overhead if we are just interested in the objects. Hence, going beyond the state of the art, we propose a framework combining gaze prediction and object detection into one network to predict visual saliency in the grid style. Based on a careful experimental evaluation, we illustrate the advantages of our model in having high performance (saliency prediction and object detection) and saving computational resources.

Object Detection. In our framework, we use existing object detection models for detecting objects in driving scenes and providing feature maps for our gaze prediction module. In the context of object detection, the *You only look once* (YOLO) architecture has played a dominant role in object detection since its first version [38]. Due to its speed, robustness and high accuracy, it is also applied frequently in autonomous driving [31, 46]. YOLOv5 [14] is one of the newest YOLO networks that performs very well. Since YOLOv5 differs from traditional YOLO networks and it does not use Darknet anymore, we also consider Gaussian YOLOv3 [7]. Gaussian YOLOv3 is a variant of YOLOv3 that uses Gaussian parameters for modeling bounding boxes and showed good results on driving datasets. For comparison, we also tried an anchor free object detection network CenterTrack [59], which regards objects as points. By using the feature maps of the object detection network such as YOLOv5 to predict gaze regions, we save the resources of an additional feature extraction module.

3 METHODOLOGY

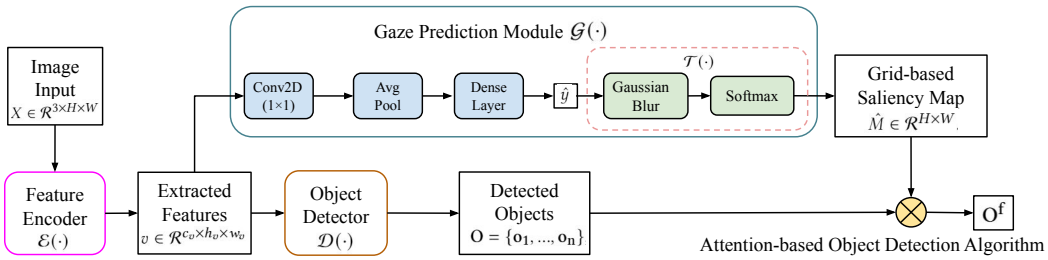


Fig. 2. Overview of our proposed driver attention-based object detection framework.

State-of-the-art driver gaze prediction models extract features from deep neural networks used in image classification or object recognition, e.g., AlexNet [19] or VGG [47], and use decoding modules to predict precise pixel-level saliency maps. We propose a new approach as shown in Fig. 2 to predict what objects drivers attend to based on a grid-based saliency map prediction. The object detector and attention predictor share the same image features and run simultaneously in a resource-efficient manner. In this section, we first introduce our attention-based object detection framework in Sec. 3.1, including the gaze prediction module and object detection algorithm, etc.

Implementation details of our model, such as the specific network architecture of network layers are discussed in Sec. 3.2.

3.1 Attention-based Object Detection

The framework is formalized as follows: Given an RGB image input from driving scenarios $X \in \mathcal{R}^{3\times H\times W}$ where H and W refer to the height and width, an image feature encoder $\mathcal{E}(\cdot)$ encodes the input image X into feature v. This feature can be a feature map $v \in \mathcal{R}^{c_v\times h_v\times w_v}$ where h_v,w_v and c_v represent the height, width and number of channels of the feature map. v is the input of the gaze prediction module $\mathcal{G}(\cdot)$, which first predicts a grid-vector $\hat{y} = \mathcal{G}(v)$. Then, a transformation operation $\mathcal{T}(\cdot)$ is applied on $\hat{y}$ to turn it into a 2-dimensional saliency map $\hat{M} \in \mathcal{R}^{H\times W}$. Similarly, the object detection module $\mathcal{D}(\cdot)$ predicts a set of objects appearing in the image $\mathbf{O} = \{\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_n}\}$, where each $\mathbf{o_i}$ contains the bounding box/class information for that object and n is the total number of objects. Based on $\hat{M}$ and $\mathbf{O}$, we run our attention-based object detection operation $\bigotimes$ to get the set of focused objects $\mathbf{O^f}$, which can be denoted as $\hat{M} \bigotimes \mathbf{O} = \mathbf{O^f}$ and $|\mathbf{O^f}| \leq n$. Fig. 2 demonstrates different modules in our framework.

Gaze Prediction Module. To reduce the computational cost, we propose to predict the gaze saliency map in grids, i.e., we alter the saliency map generation problem into a multi-label prediction problem. Concretely, we transform the target saliency map $M \in \mathcal{R}^{H\times W}$ into a grid-vector $y \in \mathcal{R}^{n\cdot m}$, where n and m are the numbers of grid cells in height and width dimension, respectively. Each entry of the grid-vector y is a binary value. The index of entry corresponds to the index of a region in the gaze map. 1 means that the region is focused by the driver, while 0 means not. Here, we obtain a grid-vector y from a saliency map M using the following procedure: (1) We binarize the M to M' with a value of 15% of the maximal pixel value (values larger than it will be set to 1, otherwise to 0). (2) For each grid cell (j-th entry in the y), we assign a "probability" of being focused as $p = \frac{\sum M'_j}{\sum M'}$, where $\sum M'_j$ is the summation of all pixel values in the j-th grid cell while $\sum M'$ is the sum of all pixels. (3) If the probability of being focused is larger than the threshold $\frac{1}{n\cdot m}$, the entry of this region will be set to 1, otherwise to 0. Fig. 3 shows an example of this procedure.

Given the grid setting n and m, the encoded feature $v = \mathcal{E}(X)$ and the grid-vector y transformed from the ground-truth saliency map M, we train the gaze prediction module $\mathcal{G}(\cdot)$ using the binary cross-entropy loss:

$$L(\hat{y}, y) = -\frac{1}{K}\sum_{i=1}^{K} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot (1 - log(\hat{y}_i)) \tag{1}$$

where $\hat{y} = \mathcal{G}(v)$ and $K = n \cdot m$ represents the number of grid cells.

To get a 2D saliency map, we conduct $\hat{M} = \mathcal{T}(\hat{y})$. More specifically, each entry in $\hat{y}$ represents a grid cell in the 2D map (see Fig. 3) and we fill each grid with its entry value. The size of each grid cell is $\frac{H}{n} \times \frac{W}{m}$, therefore a 2D matrix in the size of $n \times m$ is constructed. Then we apply a Gaussian blur and softmax to smooth the 2D matrix and use it as the predicted saliency map $\hat{M}$. The upper branch in Fig. 2 shows the procedure of predicting a grid-based saliency map.

Attention-based Object Detection Algorithm. An object detector $\mathcal{D}(\cdot)$ takes v as input and predicts all objects' information $\mathbf{O}$: the classes and bounding box. Our feature encoder $\mathcal{E}(\cdot)$ together with $\mathcal{D}(\cdot)$ form an entire object detection network. To train a good object detector, a large image dataset with densely annotated (bounding boxes and classes) information is required. Since there are some well-trained publicly available object detection models, e.g., YOLOv5 [14], we use their pretrained parameters in our $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$. More details about the architecture design will be discussed in the next section. Please note that we do not require extra training on $\mathcal{E}(\cdot)$ or $\mathcal{D}(\cdot)$, which makes

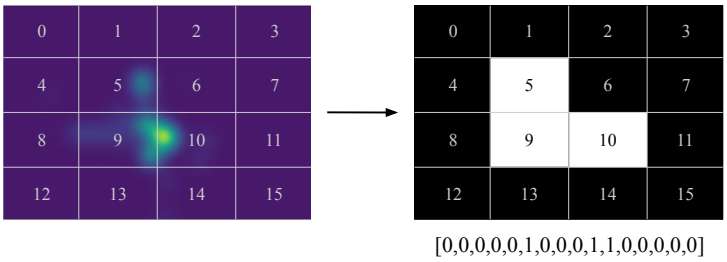


Fig. 3. Illustration of transforming a saliency map into a grid-vector. The used grid here is 4×4. Grid cells 5, 9 and 10 reach the threshold, therefore the grid-vector y for the saliency map M is $[0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0]$.

our whole framework fast to train. Given all objects' information $\mathbf{O}$ and a saliency map $\hat{M}$, the attention-based object detection operation $\bigotimes$ works as follows: for each object $\mathbf{o_i} \in \mathbf{O}$, we use the maximum pixel value inside its bounding box area on $\hat{M}$ as the probability of being focused for $\mathbf{o_i}$. A threshold Th for the probability can be set to detect whether $\mathbf{o_i}$ is focused on by drivers. Th can be chosen by users according to their requirements for different metrics, such as precision or recall. A separate discussion regarding the effect of Th can be found in Sec. 4.

3.2 Model Details

We use three pretrained object detection networks as our feature encoder $\mathcal{E}(\cdot)$, i.e., YOLOv5 [14], Gaussian YOLOv3 [7] and CenterTrack [59], to validate the efficiency and adaptability of our gaze prediction. Specifically, we deploy the layers in the YOLOv5 framework (size small, release v5.0) before the last CSP-Bottleneck (Cross Stage Partial [51]) layer within the neck (PANet [27]). Meanwhile, we use the remaining part of the model (i.e., the detector layer) as the object detector $\mathcal{D}(\cdot)$. Similarly, we use the partial network of YOLOv3 (first 81 layers) as $\mathcal{E}(\cdot)$, and use the "keypoint heatmaps" for every class of CenterTrack [59]. Tab. 1 lists the concrete dimension of extracted v. Furthermore, this table also presents the dimension of the output after each layer in the gaze prediction module. The convolutional layer with the kernel size 1×1 shrinks the input channels to 16 when using YOLO backbones, while to one channel when the CenterTrack features are used. To reduce the computational burden for the dense layer, an average pooling layer is deployed to reduce the width and height of the feature maps. Before being put into the dense layer, all the features are reshaped to vectors. The dense layer followed by the sigmoid activation function outputs the $\hat{y} \in \mathcal{R}^{n \cdot m}$.

Table 1. Network architecture details when using different object detectors. Column "Feature Encoder" shows the used backbone for extracting feature v and the dimension of v. Column "Gaze Prediction" demonstrates the dimension of output after each layer.

Feature Encoder $\mathcal{E}(\cdot)$		Gaze Prediction $\mathcal{G}(\cdot)$		
Backbone	v	Conv	Avg Pooling	Dense Layer
YOLOv5 [14]	$512 \times 12 \times 20$	$16 \times 12 \times 20$	$16 \times 6 \times 10$	number of grid cells
Gaussian YOLOv3 [7]	$1024 \times 13 \times 13$	$16 \times 13 \times 13$	$16 \times 7 \times 7$	number of grid cells
CenterTrack [59]	$80 \times 72 \times 128$	$1 \times 72 \times 128$	$1 \times 18 \times 32$	number of grid cells

4 EXPERIMENTAL RESULTS

In this section, we first introduce experimental implementation including analysis of the datasets BDD-A and DR(eye)VE, evaluation metrics and the details of how we train our proposed gaze prediction module on the BDD-A dataset. After the implementation details, we show and discuss the evaluation results of our whole framework on attention prediction as well as attention-based object detection compared to other state-of-the-art driver attention prediction networks. To further validate the effectiveness of our network, we tested and evaluated our framework on several videos from the DR(eye)VE dataset [2].

4.1 Implementation Details

4.1.1 *Datasets.*

BDD-A. The BDD-A dataset [54] includes a total of 1426 videos, each is about ten seconds in length. Videos were recorded in busy areas with many objects on the roads. There are 926 videos in the training set, 200 in the validation set and 300 in the test set. We extracted three frames per second and after excluding invalid gaze maps, the training set included 30158 frames, the validation set included 6695 frames and the test set 9831. Tab. 2 shows the statistics of the ground-truth "focused on" objects on the test set. In each image frame, there are on average 7.99 cars detected (denoted as "Total"), whereas 3.39 cars of those attract the driver's attention (denoted as "Focused"). 0.94 traffic lights can be detected in each frame, but only 0.18 traffic lights are noticed by the driver. This is due to the fact that drivers mainly attend to traffic lights that are relative to their driving direction. In total, there are 10.53 objects and approximately 40% (4.21 objects) fall within the driver's focus. Therefore, to accurately detect these focused objects is challenging.

Table 2. Traffic-related class analysis on BDD-A test set: The values in the table show the average number of objects in one video frame. "Total" means detected objects while "focused" means attended objects by the human driver. "-" refers to a number smaller than 0.001. "Sum" includes also non-traffic objects.

Object	**Person**	**Bicycle**	**Car**	**Motorcycle**	**Bus**	**Truck**
Total	0.78	0.03	7.99	0.03	0.18	0.48
Focused	0.24	0.02	3.39	0.01	0.11	0.25
Object	**Traffic light**	**Fire Hydrant**	**Stop Sign**	**Parking Meter**	**Bench**	***Sum***
Total	0.94	0.02	0.05	0.004	0.002	10.53
Focused	0.18	0.002	0.008	-	-	4.21

DR(eye)VE. The DR(eye)VE dataset [2] contains 74 videos. We used five videos (randomly chosen) from the test set (video 66, 67, 68, 70 and 72), which cover different times, drivers, landscapes and weather conditions. Each video is 5 minutes long and the FPS (frames per second) is 25, resulting in 7500 frames for each video. After removing frames with invalid gaze map records, our test set includes 37270 frames in total. We run a pretrained YOLOv5 network on all five videos and obtained the results shown in Table 3. Compared to the BDD-A dataset in Table 2, DR(eye)VE incorporates a relatively monotonous environment with fewer objects on the road. On average, there are 3.24 objects in every frame image. 39% of the objects are attended by drivers, which is similar to the BDD-A dataset.

4.1.2 *Evaluation Metrics.*

We evaluated the models from three perspectives: object detection (object-level), saliency map generation (pixel-level) and resource costs. To compare the quality of generated gaze maps, we used the Kullback–Leibler divergence (D_{KL}) and Pearson's Correlation Coefficient (CC) metrics as in previous works [33, 34, 54]. We resized the predicted and ground-truth saliency maps to 36×64

Table 3. Traffic-related class analysis on DR(eye)VE dataset (test set): The value is the average number of objects in each video frame. "Total" means detected objects while "focused" means attended objects by the human driver. "-" refers to the number smaller than 0.001. "Sum" includes also non-traffic objects.

Object	**Person**	**Bicycle**	**Car**	**Motorcycle**	**Bus**	**Truck**
Total	0.07	0.009	2.35	0.003	0.026	0.09
Focused	0.02	0.004	1.06	-	0.01	0.04
Object	**Traffic light**	**Fire Hydrant**	**Stop Sign**	**Parking Meter**	**Bench**	***Sum***
Total	0.46	-	0.02	0.005	0.003	3.24
Focused	0.07	-	0.002	0.003	-	1.26

keeping the original width and height ratio following the setting of Xia et al. [54]. Since saliency maps predicted by different models were in different sizes, we scaled them to the same size (36×64) as suggested by Xia et al. [54] to fairly compare them. For the object detection evaluation, we first decided the ground-truth "focused" objects by running our attention-based object detection on all the objects (detected by the YOLOv5 model) and the ground-truth gaze saliency maps, $M \bigotimes \mathbf{O}$, i.e., used the maximal value inside the object (bounding) area as the probability. If that probability was larger than 15%, this object was recognized as the "focused on" object. The 15% was chosen empirically to filter out the objects that were less possible than a random selection (averagely ten objects in one frame shown in Tab. 2). For the evaluation, we regarded each object as a binary classification task: the object was focused by the driver or not. The evaluation metrics used here were Area Under ROC Curve (AUC), precision, recall, F_1 score and accuracy. Except for AUC, all the metrics require a threshold Th, which will be discussed in Sec. 4.2. Finally, to quantitatively measure and compare the computational costs of our models, we considered the number of trainable parameters and the number of floating point operations (GFLOPs) of the networks.

4.1.3 *Training Details.*

All experiments were conducted on one NVIDIA CUDA RTX A4000 GPU. The proposed gaze prediction module was trained for 40 epochs on the BDD-A training set using the Adam optimizer [18] and validated on the validation set. The learning rate started from 0.01 and decayed with a factor of 0.1 after every 10 epochs. The feature encoder and the object detector were pretrained[1] and we did not require further fine-tuning for the object detection.

4.2 Results on BDD-A

4.2.1 *Quantitative Results.*

Different Grids. We first conducted experiments on different grid settings in the gaze prediction module: from 2×2 ($n = m = 2$) to 32×32 ($n = m = 32$) increasing by a factor of 2. We used YOLOv5 as our backbone for all grid settings here. The evaluation between different grids is shown in Tab. 4. "Pixel-level" refers to the evaluation of the saliency map using D_{KL} and CC metrics. "Object-level" refers to results of attention-based object detection. We set the threshold Th for detecting attended regions to 0.5 to compare the performance between different settings fairly. This evaluation shows that the performance increases when the grids become finer. Nevertheless, we can see that the advantage of 32×32 grids over 16×16 grids is not significant and the AUC is almost equal. To save computational costs, we chose the 16×16 grids as our model setting for all further experiments.

[1]Pretrained parameters for YOLOv5 can be found at https://github.com/ultralytics/yolov5; for YOLOv3 at https://github.com/motokimura/PyTorch_Gaussian_YOLOv3 and for CenterTrack at https://github.com/xingyizhou/CenterTrack.

Table 4. Comparison of using different grid settings on object- and pixel-level performance (Th=0.5). For all metrics except D_{KL}, a higher value indicates the better performance. The best result is marked in bold.

	Object-level					Pixel-level	
	AUC	*Prec (%)*	*Recall (%)*	*F_1 (%)*	*Acc (%)*	*D_{KL}*	*CC*
2×2	0.58	43.86	88.97	58.75	50.05	2.35	0.18
4×4	0.76	52.43	**91.50**	66.66	63.40	1.61	0.41
8×8	0.84	57.87	89.16	70.18	69.71	1.27	0.55
16×16	**0.85**	71.98	73.31	**72.64**	77.92	1.15	0.60
32×32	**0.85**	**75.47**	68.79	71.97	**78.58**	**1.13**	**0.62**

Table 5. Comparison of different Th using 16×16 grids on attention-based object detection. Results are shown in % and for all metrics, a higher value indicates better performance. The best result is marked in bold.

	Prec	**Recall**	**F_1**	**Acc**
0.3	63.76	**83.33**	72.24	74.39
0.4	68.11	78.36	**72.88**	76.68
0.5	71.98	73.31	72.64	77.92
0.6	75.81	68.09	71.74	**78.55**
0.7	**79.61**	62.04	69.73	78.47

Different Thresholds. The effect of different Th on attention-based object detection is listed in Tab. 5. Our results show that a lower Th yields better performance on the recall score, while a higher Th improves the precision score. The best F_1 score is achieved when Th is equal to 0.4, and for the best accuracy Th is set to 0.6. When setting Th to 0.5, we obtain relatively good performance in F_1 (72.64%) and in the accuracy (77.92%). Th is a hyperparameter that users can decide according to their requirements for the applications. For example, if high precision is preferred, Th can be set to a higher value.

Comparison with other Models. We compared our three proposed models based on YOLOv5, Gaussian YOLOv3 and CenterTrack with four existing saliency models: BDD-A [54], DR(eye)VE [34], ML-Net [8] and PiCANet [26][2]. We examined the performance from three perspectives: object detection, gaze saliency map generation and resource cost. For the object detection, we used the same object detector (YOLOv5) to detect all objects in images, then run our attention-based object detection algorithm $\bigotimes$ based on generated saliency maps from each model. The "Baseline" refers to the average BDD-A training set saliency map as illustrated in Fig. 4 (b). For a fair comparison of the Th-dependent object-level scores precision, recall, F_1 and accuracy, we computed for each model the threshold Th, which gives the best ratio of the true positive rate (TPR) and the false positive rate (FPR). Specifically, we created for each model the ROC curve (Receiver Operating Characteristic) on the BDD-A test set and determined the Th, which corresponds to the point on the curve with the smallest distance to (0,1): $argmax(\sqrt{TPR \cdot (1 - FPR)})$. The ROC curves and the values of Th for each model can be found in appendix A. Tab. 6 shows the results of our comparison with the different models. (More results of using other Th can be found in appendix B.1.)

The AUC scores show that our two YOLO models can compete on object level with the other models, even though PiCANet performs slightly better. Although our models were not trained for pixel-level saliency map generation, the D_{KL} and CC values show that our YOLOv5 based model with D_{KL} of 1.15 and CC of 0.60 is even on pixel-level comparable to the other models (under our experiment settings). In object detection, our two YOLO-based models achieve 0.85 in the AUC, which is slightly inferior to PiCANet of 0.86. Nevertheless, they have better performance in F_1 and accuracy scores than other models.

Moreover, our gaze prediction model shares the backbone (feature encoder) with the object detection network and requires mainly one extra dense layer, which results in less computational costs. For instance, our YOLOv5 based model requires 7.52M parameters in total and only 0.25M from them are extra parameters for the gaze prediction, which results in the same computational

[2]All models were trained on the BDD-A training set. Trained parameters of the BDD-A model were downloaded from https://github.com/pascalxia/driver_attention_prediction and the rest were from https://sites.google.com/eng.ucsd.edu/sage-net.

Table 6. Comparison with other gaze models on the BDD-A dataset. On object-level, all models are evaluated with detected objects of YOLOv5. Our three models use 16×16 grids. Pixel-level values in brackets are the results reported from the original work [33, 54]. * indicates that the backbone is pretrained on COCO [24], † on ImageNet [10] and ‡ on UCF101 [48]. The resource required for the gaze prediction is listed in the last column.

	Object-level					Pixel-level		Resource	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	D_{KL}	*CC*	*Param.(M)*	*GFLOPs*
Baseline	0.82	66.10	74.22	69.92	74.47	1.51	0.47	0.0	0.0
BDD-A [54] †	0.82	66.00	74.33	69.92	74.43	1.52 (1.24)	0.57 (0.59)	3.75	21.18
DR(eye)VE [34] ‡	0.85	70.04	74.94	72.41	77.16	1.82 (1.28)	0.57 (0.58)	13.52	92.30
ML-Net [8]†	0.84	70.48	73.75	72.08	77.15	1.47 (1.10)	0.60 (0.64)	15.45	630.38
PiCANet [26]†	0.86	70.23	77.67	73.76	77.91	1.69 (1.11)	0.50 (0.64)	47.22	108.08
Ours (CenterTrack)*	0.83	68.93	72.83	70.83	76.01	1.32	0.56	19.97	28.57
Ours (YOLOv3)*	0.85	70.25	74.72	72.41	77.24	1.20	0.59	62.18	33.06
Ours (YOLOv5)*	0.85	70.54	75.30	72.84	77.55	1.15	0.60	7.52	17.0

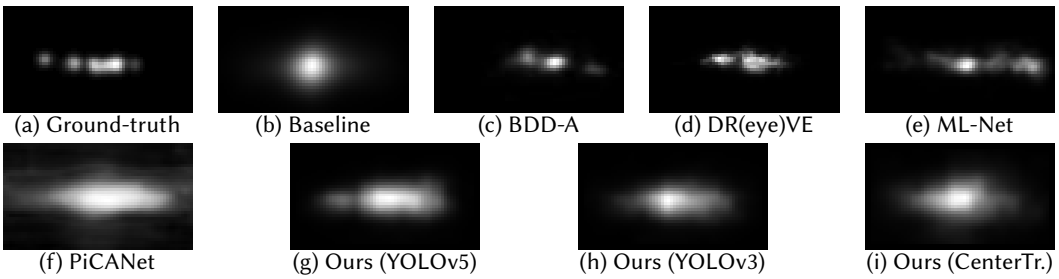


Fig. 4. Comparison of predicted driver attention saliency maps using different models. (a) Ground-truth driver attention map; (b) The baseline saliency map (center-bias); (c-f) Predictions using models [8, 26, 34, 54]; (g-i) Predictions using our framework with different backbones.

cost as a YOLOv5 network (17.0 GFLOPs). In general, the advantage of our framework is that the gaze prediction almost does not need any extra computational costs or parameters than the object detection needs. Other models need an extra object detection network to get the attention-based objects in their current model architectures. Nevertheless, we list the needed resources of each model only for the saliency prediction in Tab. 6 for a fair comparison. To achieve a similar object detection performance, for example, DR(eye)VE needs 13.52M parameters and 92.30 GFLOPs to compute only saliency maps, which are more than our YOLOv5 framework requires for the object detection task and saliency map prediction together.

4.2.2 *Qualitative Results.*

We demonstrate the qualitative results of the saliency map prediction using different models in Fig. 4. Our framework uses the backbones from YOLOv5, YOLOv3 and CenterTrack. We see that BDD-A, DR(eye)VE and ML-Net provide a more precise and concentrated attention prediction. However, BDD-A and ML-Net highlight a small area at the right side wrongly instead of an area at the left side, while our predictions (g) and (h) focus on the center part as well as the right side. Although our predictions are based on grids, they are less coarse than the ones of PiCANet.

Fig. 5 shows one example of attention-based predicted objects using different models. The predicted objects are framed with bounding boxes. The frame is taken from a video, where a vehicle drives towards a crossroad and passes waiting vehicles that are on the right lane of the road. Comparing (i) and (a), we see that the human driver pays attention to several objects but not

(a) Ground-truth driver attention (b) Ours (YOLOv5) (c) Ours (YOLOv3)

(d) Ours (CenterTrack) (e) BDD-A (f) DR(eye)VE

(g) ML-Net (h) PiCANet (i) All objects

Fig. 5. Comparison of attention-based object detection using different models. (a) Ground-truth attention; (b-d) Predictions using our framework with different backbones; (e-h) Predictions using models [8, 26, 34, 54]; (i) Object detection without driver attention.

most of the objects. Our models based on features from YOLOv5 as well as CenterTrack backbones predict all waiting vehicles as focused by drivers (in (b) and (d)), matching with the ground-truth (in (a)). BDD-A prediction focuses on a car on the oncoming lane and a church clock, missing a waiting car in the distance. Moreover, always predicting gaze at the vanishing point is a significant problem for driving saliency models. From this example, we can deduce that our model does not constantly predict the vanishing point in the street, whereas DR(eye)VE, ML-Net and PiCANet predict the object around the center point as critical.

We also present two failed predictions of our YOLOv5 based model in Fig. 6. In the first row, the vehicle is changing lanes from the left to the middle to pass two cyclists. Our model correctly notices the cars in front of the vehicle as well as the cyclists. Directly in front of the cyclists, our model predicts wrongly parked cars to be critical compared to the ground-truth. Nevertheless, this is a good example for the effect of attention-based object detection. The vehicles in front and the cyclists, which might make it necessary to react, are detected, while the cars parked two lanes away are not detected. In the second row, a vehicle drives towards a crossroad with a traffic light turning red. Our model correctly predicts the vehicle braking in front on the same lane and a car parked on the right. But additionally, our model considers a cyclist on the right of the scene as critical. Although the cyclist is wrongly predicted, it shows that the predictions of our model are not limited to the center part of an image.

4.3 Results on DR(eye)VE

4.3.1 Quantitative Results.

We tested our model on the DR(eye)VE dataset without further training to validate its generalization ability. We ran our YOLOv5 model in 16×16 grids and compared it with DR(eye)VE, BDD-A, ML-Net

Fig. 6. Comparison of our prediction, ground-truth in attention-based object detection and not using attention-based object detection on BDD-A test set. (Failed cases.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

Table 7. Comparison with other gaze models on DR(eye)VE dataset. On object-level, all models are evaluated with detected objects of YOLOv5. Our models uses 16×16 grids. * indicates that the backbone is pretrained on COCO [24], † on ImageNet [10] and ‡ on UCF101 [48].

	Object-level					**Pixel-level**	
	AUC	*Prec. (%)*	*Recall (%)*	F_1 *(%)*	*Acc (%)*	*KL*	*CC*
Baseline	0.86	65.18	77.79	70.93	77.94	2.00	0.40
BDD-A [54] †	0.84	71.63	73.34	72.48	78.38	2.07	0.46
DR(eye)VE [34] ‡	0.86	68.90	79.39	73.77	78.09	2.79	0.47
ML-Net [8]†	0.87	69.74	79.73	74.40	78.71	2.17	0.45
PiCANet [26]†	0.88	73.90	81.48	77.50	81.64	2.36	0.41
Ours (YOLOv5)*	0.88	75.33	78.73	76.99	81.74	1.78	0.51

and PiCANet. As in the experiments on BDD-A, we computed the threshold individually with the ROC curves shown in appendix A and evaluated the models on object-level with metrics *AUC*, precision, recall, F_1 and accuracy and on pixel-level with D_{KL} and *CC*. The results are shown in Tab. 7. The bottom-up models ML-Net and PiCANet achieved in our experimental setting better results than the top-down networks DR(eye)VE and BDD-A. Our model and PiCANet achieved the best results on object-level ($AUC = 0.88$) and outperformed all other models on pixel-level ($D_{KL} = 1.78$, $CC = 0.51$). Achieving good performance on DR(eye)VE shows that our model is not limited to the BDD-A dataset.

4.3.2 *Qualitative Results.*

Fig. 7 shows two examples of our attention-based object prediction model on the DR(eye)VE dataset. The frames in the first row belong to a video sequence where the driver follows the road in a

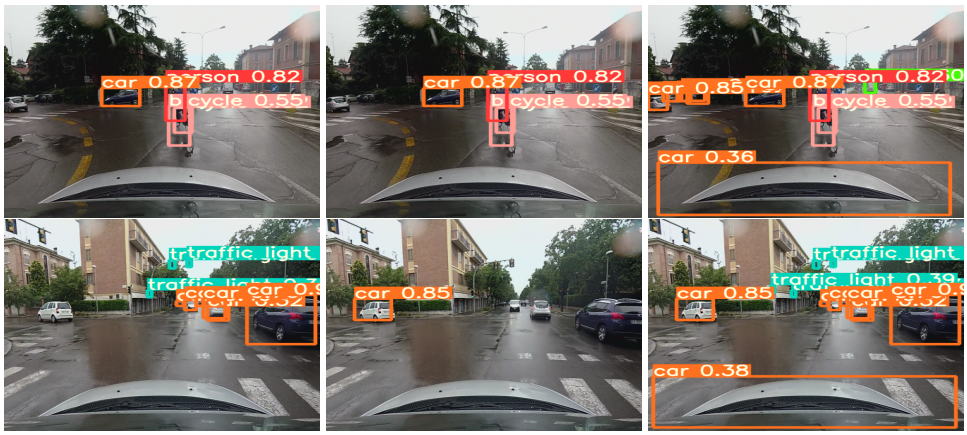


Fig. 7. Comparison of our prediction, ground-truth in attention-based object detection and not using attention-based object detection on the DR(eye)VE testset ($Th = 0.4$ to better illustrate the wrongly predicted attention region in the failed case). (Second line is failed case.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

left curve. Our model (left) detects the cyclist driving in front of the car and a vehicle waiting on the right to merge. Other cars further away were not predicted as focused, thus it matches the ground-truth (middle). In the second row, we can see a frame where the driver wants to turn left. Our model (left) predicts the cars and traffic lights on the road straight ahead, whereas the ground-truth (middle) covers a car turning left. This example underlines the difficulty of predicting drivers' attention when it depends on different driving goals [56].

5 DISCUSSION

In this section, we first show our LSTM-variant architecture and discuss the results to address the challenges of using temporal information in this task. Then, we deliberate other limitations of the current project.

5.1 Modelling with LSTM-Layer

To extend our framework into a video-based prediction, we added one LSTM-layer (Long Short-Term Memory [13]) with 256 as the size of the hidden state before the dense layer in the gaze prediction network. The input for this network is an eight-frame video clip. We tested our extended architecture using the same configuration described in the last section (i.e., 16×16 grids with Th of 0.5) and achieved the following results on the BDD-A dataset:

Object Detection: AUC = 0.85, Precision = 73.13%, Recall = 70.44%,
F_1 score = 71.76%, Accuracy = 77.83%

Saliency Prediction: D_{KL} = 1.17, CC = 0.60

The above results are similar to our model without the LSTM-layer, both achieved $AUC = 0.85$ and $CC = 0.60$. It is worth mentioning that the sequence length (from 2 to 16) had no significant influence on the performance. (See appendix B.4 for more results.) Similarly, [54] also observes that using LSTM-layers cannot improve the performance in driver gaze prediction but rather introduces center biases into the prediction. In summary, more frames do not increase the information gain. One possible reason behind this bias is that using an LSTM-layer ignores the spatial information, since the extracted features given to the LSTM-layer are reshaped to vectors. Therefore, in the

context of our future work, we would like to analyze the integration of other modules that include temporal information, such as the convolutional LSTM (convLSTM) [55]. Using convLSTM can capture the temporal information of each spatial region and predict the new features for it based on the past motion information inside the region. For example, [40, 54] validate that convLSTM helps capture the spatial-temporal information for driver attention/action predictions. Another proposal is to use 3D CNN to get the spatial-temporal features. For instance, [34] deploys 3D convolutional layers that takes a sequence of frames as input in predicting the driver's attention.

5.2 Limitations and Future Work

One limitation of current projects is that all current models have a central bias in their prediction. This effect stems from the ground-truth data because human drivers naturally look at the center part of the street, creating very unbalanced data: 74.2% of all focused objects on BDD-A come from the central bias area as shown in the baseline in Figure 4. The central bias reflects natural human behavior and is even enhanced in the saliency models proposed by Kümmerer et al. [22, 23]. Although our model predicts objects in the margin area of the scene as shown in our qualitative examples, the center is often prioritized. Our model has an F_1 score of 81.7% inside of the center area, while it only reaches 34.8% in F_1 outside of the center area. PiCANet, which achieves the best result among all models, has better F_1 scores outside (44.0%) and inside of the center (82.7%), however, its performance inside of the center is dominant. We intend to improve the model prediction outside of the center but still keep the good performance in the center area in the future. In the context of autonomous driving, it would be also essential to test the generalization ability on other datasets, which are not limited to just the gaze map data. Since drivers also rely on peripheral vision, they do not focus on every relevant object around them. Using other datasets that additionally highlight objects based on semantic information (e.g., [33]) could increase the applicability for finding task-relevant objects.

All models in the experiments are trained on saliency maps derived from driver gaze. These salient features are related to regions of interest where a task-relevant object should be located, thus reflecting top-down features [32]. However, these features are currently extracted from the visual information given by camera images. The context of driving tasks can still be enhanced by adding more input information, since human top-down feature selection mechanisms require comprehensive understanding of the task that is outside the realm of visual perception. Concretely, the driver's attention can be affected by extrinsic factors such as road conditions, or intrinsic factors such as driver intentions based on driving destinations. These factors, along with traffic information, form the driver attention as well as gaze patterns. Unfortunately, the current dataset used for our model training does not provide this additional input. For the future work, we will consider incorporating GPS and Lidar sensor information, which can provide more insights of tasks to better predict driver attention.

6 CONCLUSION

In this paper, we propose a novel framework to detect human attention-based objects in driving scenarios. Our framework predicts driver attention saliency maps and detects objects inside the predicted area. This detection is achieved by using the same backbone (feature encoder) for both tasks, and the saliency map is predicted in grids. In doing so, our framework is highly computation-efficient. Comprehensive experiments on two driver attention datasets, BDD-A and DR(eye)VE, show that our framework achieves competitive results in the saliency map prediction and object detection compared to other state-of-the-art models while reducing computational costs.

7 ACKNOWLEDGMENTS

We acknowledge the support by Cluster of Excellence - Machine Learning: New Perspectives for Science, EXC number 2064/1 - Project number 390727645.

REFERENCES

[1] Ekrem Aksoy, Ahmet Yazıcı, and Mahmut Kasap. 2020. See, Attend and Brake: An Attention-based Saliency Map Prediction Model for End-to-End Driving. *arXiv preprint arXiv:2002.11020* (2020).
[2] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. 2016. Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In *CVPRW*.
[3] Michael Barz, Sebastian Kapp, Jochen Kuhn, and Daniel Sonntag. 2021. Automatic recognition and augmentation of attended objects in real-time using eye tracking and a head-mounted display. In *ACM ETRA*. 1–4.
[4] Michael Barz and Daniel Sonntag. 2021. Automatic Visual Attention Detection for Mobile Eye Tracking Using Pre-Trained Computer Vision Models and Human Gaze. *Sensors* 21, 12 (2021), 4143.
[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).
[6] Ali Borji. 2018. Saliency prediction in the deep learning era: Successes, limitations, and future challenges. *arXiv preprint arXiv:1810.03716* (2018).
[7] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. 2019. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *ICCV*. 502–511.
[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep multi-level network for saliency prediction. In *ICPR*.
[9] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *ECCVW*, Vol. 1. Prague, 1–2.
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.
[11] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and B. Manjunath. 2019. How Do Drivers Allocate Their Potential Attention? Driving Fixation Prediction via Convolutional Neural Networks. *T-ITS* (2019).
[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*. 2961–2969.
[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[14] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. 2021. *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models.* https://doi.org/10.5281/zenodo.4679653
[15] lv Kai, Hao Sheng, Zhang Xiong, Wei Li, and Liang Zheng. 2020. Improving Driver Gaze Prediction With Reinforced Attention. *IEEE Transactions on Multimedia* (2020).
[16] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data* 8, 1 (2021), 1–18.
[17] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *ECCV*. 563–578.
[18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS* (2012).
[20] Puneet Kumar, Mathias Perrollaz, Stéphanie Lefevre, and Christian Laugier. 2013. Learning-based approach for online lane change intention prediction. In *IV*. IEEE, 797–802.
[21] Niharika Kumari, Verena Ruf, Sergey Mukhametov, Albrecht Schmidt, Jochen Kuhn, and Stefan Küchemann. 2021. Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4. *Sensors* 21, 22 (2021), 7668.
[22] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. 2014. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045* (2014).
[23] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* (2016).
[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.
[25] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bertram E Shi. 2019. A gaze model improves autonomous driving. In *ACM ETRA*.

[26] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*. 3089–3098.
[27] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *CVPR*. 8759–8768.
[28] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. 2021. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*. 3794–3803.
[29] Eduardo Manuel Silva Machado, Ivan Carrillo, Miguel Collado, and Liming Chen. 2019. Visual Attention-Based Object Detection in Cluttered Environments. In *SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*. IEEE, 133–139.
[30] Alexander Makrigiorgos, Ali Shafti, Alex Harston, Julien Gerard, and A Aldo Faisal. 2019. Human visual attention prediction boosts learning & performance of autonomous driving agents. *arXiv preprint arXiv:1909.05003* (2019).
[31] Brilian Tafjira Nugraha, Shun-Feng Su, et al. 2017. Towards self-driving car using convolutional neural network and road lane detector. In *ICACOMIT*.
[32] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. 2003. Top-down control of visual attention in object detection. In *ICIP*, Vol. 1. IEEE, I–253.
[33] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. 2020. "Looking at the Right Stuff"-Guided Semantic-Gaze for Autonomous Driving. In *CVPR*.
[34] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. 2018. Predicting the Driver's Focus of Attention: the DR(eye)VE Project. *TPAMI* (2018).
[35] Karen Panetta, Qianwen Wan, Aleksandra Kaszowska, Holly A Taylor, and Sos Agaian. 2019. Software architecture for automating cognitive science eye-tracking data analysis and object annotation. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 268–277.
[36] Robert J Peters and Laurent Itti. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*. IEEE, 1–8.
[37] Laura Pomarjanschi, Michael Dorr, and Erhardt Barth. 2012. Gaze guidance reduces the number of collisions with pedestrians in a driving simulator. *ACM TiiS* 1, 2 (2012), 1–14.
[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*. 779–788.
[39] Raymond D Rimey and Christopher M Brown. 1994. Control of selective perception using bayes nets and decision theory. *IJCV* 12, 2 (1994), 173–207.
[40] Yao Rong, Zeynep Akata, and Enkelejda Kasneci. 2020. Driver intention anticipation based on in-cabin and driving scene monitoring. In *ITSC*. IEEE, 1–8.
[41] Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. 2021. Human Attention in Fine-grained Classification. In *BMVC*.
[42] Khaled Saab, Sarah M Hooper, Nimit S Sohoni, Jupinder Parmar, Brian Pogatchnik, Sen Wu, Jared A Dunnmon, Hongyang R Zhang, Daniel Rubin, and Christopher Ré. 2021. Observational supervision for medical image classification using gaze data. In *MICCAI*. Springer, 603–614.
[43] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *CHI*. 771–780.
[44] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. 2015. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*.
[45] Mohsen Shirpour, Steven S Beauchemin, and Michael A Bauer. 2021. Driver's Eye Fixation Prediction by Deep Neural Network.. In *VISIGRAPP*.
[46] Martin Simony, Stefan Milzy, Karl Amendey, and Horst-Michael Gross. 2018. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *ECCV*.
[47] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01* (2012).
[49] Richard S Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine learning* 3, 1 (1988), 9–44.
[50] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. In *CVPR*. 4129–4138.
[51] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In *CVPRW*. 390–391.
[52] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. 2018. Salient object detection driven by fixation prediction. In *CVPR*. 1711–1720.
[53] Julian Wolf, Stephan Hess, David Bachmann, Quentin Lohmeyer, and Mirko Meboldt. 2018. Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research* 11,

6 (2018).
[54] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. 2018. Predicting driver attention in critical situations. In *ACCV*.
[55] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, Vol. 28.
[56] Alfred L Yarbus. 1967. Eye movements during perception of complex objects. In *Eye Movements and Vision*. Springer, 171–211.
[57] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020. Human Gaze Assisted Artificial Intelligence: A Review. In *IJCAI*, Vol. 2020. NIH Public Access, 4951.
[58] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems* 30, 11 (2019), 3212–3232.
[59] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *ECCV*.

A VISUALIZATION OF THE ROC CURVES

In Fig. 8 and Fig. 9 we show the ROC curves and computed thresholds for all models on the BDD-A and DR(eye)VE test sets.

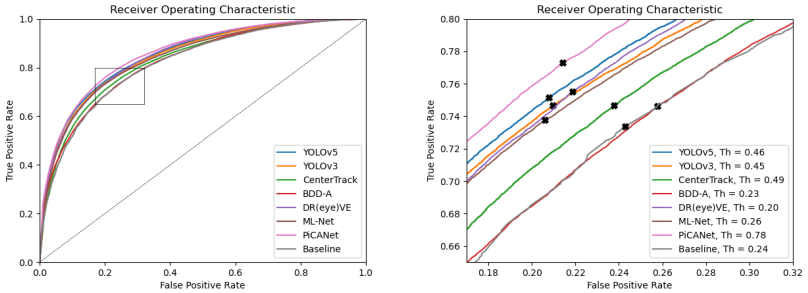


Fig. 8. ROC curves and computed thresholds on the BDD-A test set. On the right, the curves are zoomed in and the points that belong to the computed thresholds are marked.

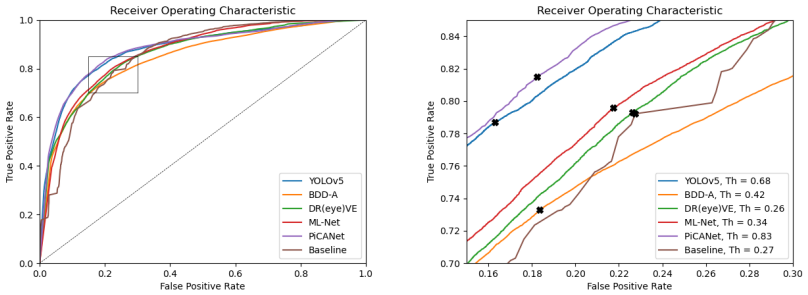


Fig. 9. ROC curves and computed thresholds on the DR(eye)VE test set. On the right, the curves are zoomed in and the points that belong to the computed thresholds are marked.

B MORE QUANTITATIVE RESULTS

B.1 Results of Other Thresholds on BDD-A

Our models always achieve high F_1 scores in different Th, indicating that our models have relatively good performance in precision and recall scores at the same time. PiCANet is more unbalanced

in recall and precision compared to other models. The accuracy scores are influenced by the Th values, however, the highest accuracy 78.55% is achieved by our YOLOv5-based model when Th is set to 0.6.

Table 8. Comparison of different models with Th = 0.3 on BDD-A dataset. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	68.88	69.43	69.16	75.24
DR(eye)VE	75.32	66.42	70.59	77.87
ML-Net	72.84	70.43	71.61	77.68
PiCANet	43.61	99.36	60.61	48.36
Ours (CenterTrack)	61.19	83.11	70.49	72.17
Ours (YOLOv3)	63.97	82.15	71.93	74.36
Ours (YOLOv5)	63.76	83.33	72.24	74.39

Table 9. Comparison of different models with Th = 0.4 on BDD-A dataset. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	72.68	63.44	67.75	75.85
DR(eye)VE	78.99	59.95	68.16	77.61
ML-Net	77.50	62.79	69.37	77.83
PiCANet	47.15	98.26	63.72	55.27
Ours (CenterTrack)	65.52	77.86	71.15	74.76
Ours (YOLOv3)	68.16	77.02	72.32	76.43
Ours (YOLOv5)	68.11	78.36	72.88	76.68

Table 10. Comparison of different models on BDD-A dataset with Th = 0.5. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	75.84	57.44	65.37	75.67
DR(eye)VE	81.84	54.38	65.35	76.94
ML-Net	80.96	55.75	66.03	77.07
PiCANet	51.26	95.98	66.83	61.90
Ours (CenterTrack)	69.29	72.19	70.71	76.09
Ours (YOLOv3)	72.14	72.23	72.18	77.74
Ours (YOLOv5)	71.98	73.31	72.64	77.92

Table 11. Comparison of different models on BDD-A dataset with Th = 0.6. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	78.84	51.41	62.23	75.05
DR(eye)VE	84.13	49.57	62.39	76.10
ML-Net	83.53	48.80	61.61	75.68
PiCANet	56.31	92.30	69.95	68.29
Ours (CenterTrack)	73.13	66.12	69.45	76.74
Ours (YOLOv3)	75.71	66.68	70.91	78.12
Ours (YOLOv5)	75.81	68.09	71.74	78.55

Table 12. Comparison of different models on BDD-A dataset with Th = 0.7. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	81.90	45.57	58.56	74.21
DR(eye)VE	86.14	44.34	58.55	74.89
ML-Net	85.85	42.31	56.69	74.15
PiCANet	63.10	85.88	72.75	74.28
Ours (CenterTrack)	76.91	59.52	67.11	76.67
Ours (YOLOv3)	79.39	60.44	68.63	77.91
Ours (YOLOv5)	79.61	62.04	69.73	78.47

B.2 Results of Other Thresholds on DR(eye)VE

Our model achieves the best F_1 socre of 76.94% and accuracy of 81.9%, while the best F_1 score and accuracy scores among other models are 74.24% and 79.68% respectively, which validates the good performance of our model in the attention-based object detection task.

Table 13. Comparison of different models on DR(eye)VE dataset with Th = 0.3. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	65.94	78.83	71.81	75.98
DR(eye)VE	70.34	76.95	73.50	78.46
ML-Net	67.98	81.77	74.24	77.98
PiCANet	42.34	98.98	59.31	47.31
Ours (YOLOv5)	58.08	91.25	70.98	71.04

Table 14. Comparison of different models on DR(eye)VE dataset with Th = 0.4. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	70.82	74.16	72.45	78.12
DR(eye)VE	73.57	71.54	72.54	78.98
ML-Net	71.85	76.23	73.97	79.19
PiCANet	46.83	95.81	62.91	56.16
Ours (YOLOv5)	62.81	89.19	73.71	75.31

Table 15. (ADDED) Comparison of different models on DR(eye)VE dataset with Th = 0.5. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	74.58	69.53	71.97	78.98
DR(eye)VE	76.21	66.46	71.01	78.94
ML-Net	75.48	71.02	73.19	79.80
PiCANet	51.30	93.92	66.36	63.05
Ours (YOLOv5)	68.33	85.83	76.08	79.06

Table 16. Comparison of different models on DR(eye)VE dataset with Th = 0.6. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	77.34	65.54	70.95	79.17
DR(eye)VE	79.25	61.67	69.37	78.86
ML-Net	78.43	66.84	72.17	80.00
PiCANet	56.95	92.19	70.40	69.92
Ours (YOLOv5)	71.90	82.26	76.73	80.64

Table 17. Comparison of different models on DR(eye)VE dataset with Th = 0.7. Results are shown in % and for all metrics, a higher value indicates better performance.

	Prec	Recall	F_1	Acc
BDD-A	79.74	61.61	69.51	79.03
DR(eye)VE	81.88	57.21	67.35	78.48
ML-Net	80.70	62.61	70.51	79.68
PiCANet	62.88	89.49	73.86	75.42
Ours (YOLOv5)	76.09	77.80	76.94	81.90

B.3 Results of Our YOLOv3- and CenterTrack-based Models

For a fair comparison, we computed object-level metrics with the detected objects of YOLOv5 for all models in Sec. 4. In Tab. 18, we show the object-level results for our 16×16 grids YOLOv3 and CenterTrack based models using their detected objects.

Table 18. Comparison of different models on BDD-A dataset with own detected objects (Th = 0.5). For all metrics a higher value indicates better performance.

	AUC	Prec (%)	Recall (%)	F_1 (%)	Acc (%)
CenterTrack	0.83	69.80	74.62	72.13	75.33
YOLOv3	0.84	70.23	73.42	71.79	76.22

B.4 Results of Different Input Sequence Lengths of LSTM

In Tab. 19 the results for different input sequence lengths are shown, when adding one LSTM layer with hidden size 256 before the dense layer of our YOLOv5 based 16×16 grids model. All sequence length achieve very similar results.

Table 19. Comparison of different input sequence lengths when using one LSTM layer. Our model uses the 16×16 grids. For all metrics except D_{KL}, a higher value indicates the better performance. (Th = 0.5)

	Object-level					**Pixel-level**	
	AUC	*Prec. (%)*	*Recall (%)*	*F_1 (%)*	*Acc (%)*	*KL*	*CC*
2	0.85	72.40	72.68	72.54	78.00	1.16	0.60
4	0.85	72.58	73.02	72.80	78.18	1.16	0.60
6	0.85	72.52	73.04	72.78	78.16	1.18	0.60
8	0.85	73.13	70.44	71.76	77.83	1.17	0.60
16	0.85	71.84	73.39	72.61	77.86	1.18	0.60

C MORE QUALITATIVE RESULTS

C.1 LSTM

In Fig. 10 there are two examples of predicted gaze maps with LSTM module (middle) in comparison with predicted gaze maps without LSTM module (left) and ground-truth (right). The LSTM module contains one layer with hidden size 256 and the input sequence length is 8. We see that the results with LSTM module enhance the prediction of the center area, which has sometimes advantages and sometimes disadvantages, thus the AUC is the same (0.85).

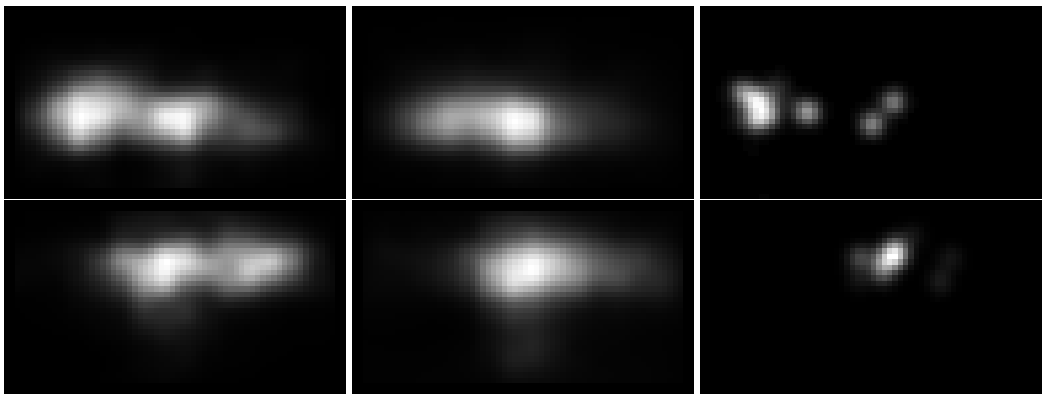


Fig. 10. Comparison of predicted gaze maps without and with LSTM and ground-truth **Left**: Our prediction without LSTM; **Middle**: Our prediction with LSTM; **Right**: Ground-truth.

C.2 BDD-A Dataset

In Fig. 11 there are two more examples of our YOLOv5 based model on BDD-A dataset. In the first row, our model predicts correctly the car on the two lanes leading straight ahead and ignoring parked cars two lanes away and another car on a turn lane. In the second row, our model predicts a traffic light in the middle of the scene, and two parked cars which could be critical if the driver would drive straight ahead. Since the driver turns left, the ground-truth covers objects on the turning road.

Fig. 11. Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.5) and not using attention-based object detection on BDD-A test set. (Second line is failed case.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

C.3 DR(eye)VE Dataset

Fig. 12 and Fig. 13 are two more examples of predicted objects with our YOLOv5 based model on DR(eye)VE dataset. In Fig. 12 we see that our model predicts correctly the cars on the road and ignores the parked cars two lanes away. In Fig. 13 our model predicts the cyclist next to the vehicle and a car waiting to the right, while the ground-truth focuses objects which the driver will pass later. One reason could be that the driver sees the objects next to him with peripheral view.

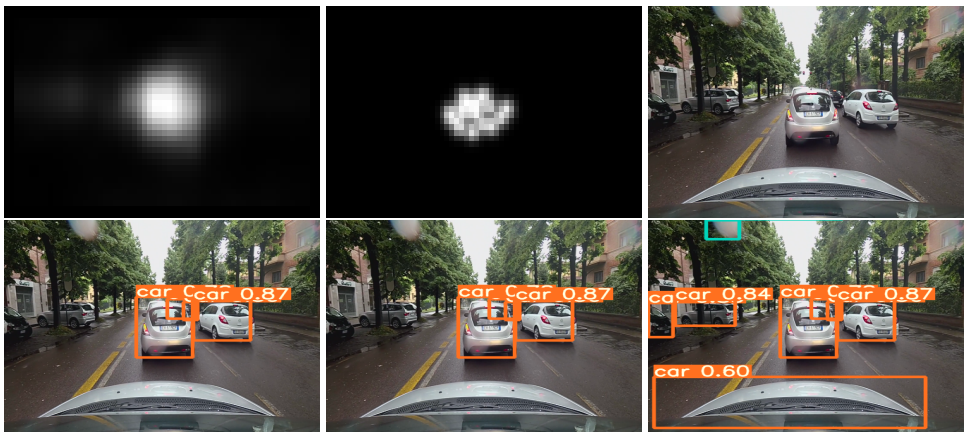


Fig. 12. Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.4) and not using attention-based object detection on DR(eye)VE test set. **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

Received November 2021; revised January 2022; accepted April 2022

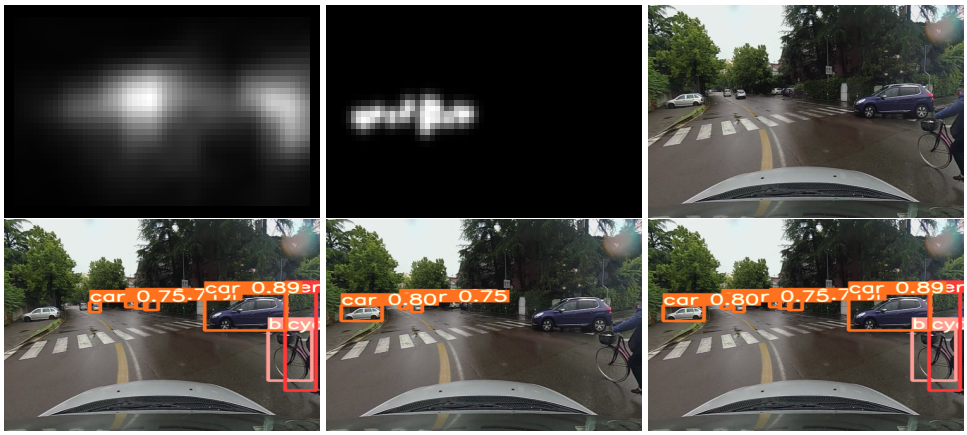


Fig. 13. Comparison of our prediction, ground-truth in attention-based object detection (Th = 0.4) and not using attention-based object detection on DR(eye)VE test set. (Failed case.) **Left**: Our prediction; **Middle**: Ground-truth; **Right**: Object detection without driver attention. Better view in colors.

Publication Rights & Licensing Policy

Updated on January 1, 2023

Introduction

ACM embraces a not-for-profit business model that aims to assure sustainable revenue for the continued operation and enhancement of the ACM Publications Program and the ACM Digital Library, while making ACM Publications available to the widest possible global audience of computing professionals and students.

For over half a century, ACM has requested that authors transfer copyright of their articles, so that ACM could act as a steward of their published work, manage the publication process, respond to requests related to third-party rights and permissions, and defend their published Works against misconduct such as plagiarism or copyright infringement. Since that time, ACM's copyright and permissions policies have been widely used as a model by other scholarly publishers in adapting their own policies to the ever-changing realities of electronic dissemination and open access publication.

Over the years, ACM has made regular updates to its copyright policy, which is now in its 10th iteration, to ensure we are acting in the best interest of our authors and the global computing community, such as we did in 2013 when ACM introduced Exclusive, Non-exclusive, and Creative Commons licensing options as part of our ACM eRights process for authors. Many of these changes were done to support our authors with options enabling them to comply with government open science mandates around the world and to retain the underlying intellectual property of their Work.

Today, every ACM author of a scholarly Work accepted by an ACM Publication has the option of retaining the copyright of their Work and granting ACM a license to publish that Work in the ACM Digital Library. For Corresponding Authors affiliated with ACM Open participating institutions or Corresponding Authors not affiliated with an ACM Open institution, but who are willing to pay a reasonably priced Article Processing Charge (APC), there is an additional option to select an appropriate **Creative Commons** license to facilitate sharing and reuse of their Works, so the community may build on their Work without the need to obtain additional permissions from ACM or the Author, provided proper attribution is given.

As ACM continues to transition its entire scholarly Publication program to an Open Access model, the use of **Creative Commons** licensing is becoming more prevalent. In fact, many of the large government Open Science mandates around the world require the use of a Creative Commons or equivalent license when research grant recipients publish Work funded by those governments. Many private research funders are following suit (i.e. - Gates Foundation, Welcome Trust, etc.).

With its stated goal of sustainably transitioning to a fully Open Access Publisher around the end of 2025 and in response to calls for greater copyright retention and intellectual property ownership by ACM's authorship, ACM is now taking the most significant step forward since the creation of its Copyright Policy in 1994 by effectively sensetting the existing Copyright Policy and replacing it with this new Publication Rights and Licensing Policy. ACM will continue to register and hold copyright and other intellectual property rights of ACM Journals, Magazines, Conference Proceedings, Newsletters, Books, and other ACM Publications, but after January 1, 2023 ACM will no longer hold copyright in any of the newly published articles in ACM Publications.

What is Changing?

ACM will continue to require authors to assign publication rights to ACM as a condition of publishing the work. This is necessary to protect both ACM's authors and ACM against infringement and misconduct by third parties.

During the June 2022 meeting of the ACM Publications Board, the Board took perhaps the most significant "copyright-related" step taken in its history by voting to end the “Copyright Transfer” option **starting January 1, 2023**. After January 1, 2023, when authors’ Works are accepted into any of ACM’s Publications and enter the ACM Rights System via the link in their Acceptance Email, the "Corresponding Author” will no longer be given the option (currently listed as the 3rd of 3 options) of transfering copyright to ACM. For published Works prior to that date where copyright has been transferred by the Author to ACM, ACM will continue to be the copyright holder for such Works.

After January 1, 2023, there will be two remaining options, as follows:

- **Institutional Paid Open Access / Permissions Release** - This is the Open Access option. Wording may vary slightly depending on whether the Corresponding Author is affiliated with an ACM Open participating institution or not. If not, they will be given the option to pay an Article Proceeding Charge (APC). This option is the default when the Corresponding Author is affiliated with an ACM Open participating institution. Authors selecting this option will retain all rights to their Work and agree to grant ACM a non-exclusive permission to publish their Work in the ACM Digital Library and have the additional option of displaying a Creative Commons license on the published version of their Work in the ACM Digital Library.
- **Closed Access / Exclusive License to Publish** - This is the Closed Access option. Authors selecting this option will retain all rights to their Work and grant ACM an exclusive license to publish their Work in the ACM Digital Library.

Creative Commons Licensing Options

If the Corresponding Author of a Work accepted into an ACM Publication is either affiliated with an **ACM Open participating institution** or has decided to pay the **Open Access Article Processing Charge** (APC), the Corresponding Author will be given the additional option of applying a Creative Common license to govern how their Work may be shared and reused. Most US and European funding agencies prefer the use of the CC-BY 4.0 License, although

authors should check with their specific funder to learn if their funder has any firm requirements on the version of Creative Commons license they must use as part of the publishing process.

The current ACM Policy is to allow authors the option of selecting their preferred version. ACM currently offers 6 Creative Commons license options, including:

- **CC-BY 4.0 License** - This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.
- **CC-BY 4.0-SA -** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.
- **CC-BY 4.0-NC -** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.
- **CC-BY 4.0-NC-SA -** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.
- **CC-BY 4.0-ND -** This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use.
- **CC-BY 4.0-NC-ND -** This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator.

Creative Common Zero (CC-0) License

There is one additional CC License that ACM Authors may apply to their research artifacts (i.e. - data, code, etc.) called **CC-0**. CC-O allows creators to give up their copyright and put their Works in the worldwide public domain. CC-0 is no longer offered in the ACM Rights system for ACM Publications, because it places the Work in the public domain and is irreversible, which could create problems for the author and ACM as the Publisher in the future. However, when ACM Authors are depositing their research artifacts either in the ACM DL or a third-party site such as GITHUB, some authors may wish to assign a CC-0 license to those research artifacts. ACM cautions the use of CC-0 unless the author has given significant consideration to this and would like to give away their copyright and allow unrestricted use of their research artifacts to the public. When ACM Authors choose to apply a CC-0 license to their research artifacts, they should indicate this alongside the artifact(s) wherever that artifact is hosted inside or outside the ACM Digital Library.

Defending Authors Against Misconduct

One of the major changes with the removal of the copyright transfer option is that regardless of which option the Author selects, ACM commits to defending their published Work in the ACM Digital Library against infringement and misconduct without the requirement to hold copyright on the published Work. In

practice, ACM has been doing this for years, but is formalizing this commitment in this new Policy. When an ACM Author agrees to have ACM serve as the Publisher of Record for their accepted Work, protecting that Work against various forms of infringement and misconduct by third parties is one of the services ACM commits to provide to the Author. In return, ACM Authors agree to abide by all of **ACMs Publications Policies** and cooperate with ACM staff, volunteers, and advisers in their investigations and process to adjudicate allegations of infringement and misconduct.

Requirement to Grant ACM Exclusive or Non-Exclusive Publication Rights (applies to Journal, Conference, and Magazine articles)

ACM requires that authors have the authority to grant publication rights to ACM or that they obtain the necessary authorization to execute the grant of publication rights and that they complete ACM's Rights Management Process as a pre-condition for publishing their Work with ACM. Such grant applies to any medium used by ACM for publication (i.e.- print, online, etc.). If Authors are uncertain about their having the authority to grant these rights as a result of their employer's intellectual property rights requirements or working for a government employer with specific requirements, they should always check with their employer before completing ACM's Rights Assignment process. Authors should also take note of the following:

- Authors should incorporate the appropriate Copyright or License notice and ACM citation of the publication into copies they personally maintain on non-ACM servers.
- The author's grant of publication rights applies only to the Work as a whole, and not to any embedded objects owned by third parties. An author who embeds an object, such as an art image that is copyrighted by a third party, must obtain that party's permission to include the object, with the understanding that the entire work may be distributed as a unit in any medium.
- The requirement to obtain third-party permission does not apply if the author embeds only a link to the copyright holder's object. Other requirements for third-party permissions can be found below under the section called 3rd Party Permissions.
- Authors who wish to embed a component of another ACM-copyrighted or licensed work, e.g., an excerpt, a table, or a figure, must obtain an explicit permission (there is no fee) from ACM.

Self-Archiving and Posting Rights

All ACM published authors of magazine articles, journal articles, and conference papers retain the right to post the pre-submitted (also known as "pre-prints"), submitted, accepted, and peer-reviewed versions of their work in any and all of the following sites:

- Author's Homepage
- Author's Institutional Repository
- Any Repository legally mandated by the agency or funder funding the research on which the work is based
- Any Non-Commercial Repository or Aggregation that does not duplicate ACM tables of contents. Non-Commercial Repositories are defined as Repositories owned by non-profit organizations that do not charge a fee to access deposited articles and that do not sell advertising or otherwise profit from serving scholarly articles.

Authors should include an appropriate citation and attribution statement on all Submitted or Accepted versions of the Work similar to the following:

- **"© {Owner/Author | ACM} {Year}. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in {SourcePublication}, http://dx.doi.org/10.1145/{number}."**

For the avoidance of doubt, an example of a site ACM authors may post all versions of their work to, with the exception of the final published "Version of Record", is arXiv. ACM does request authors, who post to ArXiv or other permitted sites, to also post the published version's Digital Object Identifier (DOI) alongside the pre-published version on these sites, so that easy access may be facilitated to the published "Version of Record" upon publication in the ACM Digital Library.

Examples of sites ACM authors may not post their work to are ResearchGate, Academia.edu, Mendeley, or Sci-Hub, as these sites are all either commercial or in some instances utilize predatory practices that violate copyright, which negatively impacts both ACM and ACM authors.

Current ACM Publications Policy is that ACM sponsored and ICPS conferences may not impose embargoes on authors posting pre-prints of submissions on arXiv or disqualify such submissions that have already been posted on arXiv at the time of submission or during the peer review process. This policy was most recently reaffirmed by the ACM Publications Board in 2019. This Policy is currently under reconsideration by the ACM Publications Board and it is expected that this policy will either be reaffirmed or updated by December 31, 2022.

Requirements for ACM Books Authors

Unlike other types of ACM Publications listed above, ACM Books authors shall continue to be given the option of signing either a Copyright Transfer & Publishing Agreement or Exclusive License to Publish Agreement. The reason for this is that there are fundamental differences in how books are published, marketed, sold, and distributed via the ACM Digital Library, 3rd party channels, and in print that relate primarily to commercial considerations, financial remuneration for ACM Books authors, and posting or self-archiving policies for ACM Books, which differs from ACM's general posting and self-archiving policy for journal, conference, and magazine authors. For more information, please see the **Publishing Policies related to ACM Books authors**.

Definitive Versions of Record, Official Publication Dates, and Corrections to the Version of Record

Preserving the scholarly record "as published" is a critical component of maintaining the community and public's trust in scientific publications in general and trust in ACM specifically. As a result, ACM is committed to the publication and long term digital preservation of published works in the ACM Digital Library and via several third-party digital preservations initiatives, including **CLOCKS** and **Portico**. ACM will create and maintain a definitive Version of Record (VoR) of all ACM published works and share these with our digital preservation providers. There are instances where VoRs are hidden in the ACM Digital Library for legal or public safety reasons, to comply with other ACM Publications Policies, such as in connection with the implementation of ACMs **Name Change Policy**, when **Retractions** are made, or when Corrected Versions of Record (CVoR) are added to ACM Digital Library citation pages when errata or corrigenda are created in connection with a published work. ACM will provide the reason for the Correction on the article's Digital Library citation page. ACM does not alter works once published. There are times, however, when it is appropriate to publish a revised or corrected version of a work; doing so requires the approval of the responsible editor. Please see ACM's **Publications Policy on the Withdrawal, Correction, Retraction, and Removal of Works from ACM Publications and ACM DL**

Persistent Unique Identifiers for Every ACM Article

The **DOI (Digital Object Identifier)** is the scholarly publishing standard (ISO 26324) identifier for articles published by ACM in the ACM Digital Library. Every article in the ACM Digital Library shall have one and only one DOI.

The official publication date of an ACM published article will be considered the date on which the article’s official Version of Record (VoR) is published online in the ACM Digital Library, and the official VoR of an ACM article shall be the final peer reviewed, accepted, edited, tagged, and identified (using a DOI or other standardized identifier) definitive version that appears in ACM Publications (i.e. - journals, magazines, conference proceedings, newsletters, books, etc.) inside the ACM Digital Library.

For the avoidance of doubt, only the official VoR or in CVoR shall be considered the “Published” version of the Work for purposes of attribution, rights & permissions, prior art, investigations into potential ethics & plagiarism violations or other forms of infringement, and relevant open access embargo periods. If a new Work is substantially developed, i.e., it contains at least 25% new substantive material, it is considered a new Derivative Work or Major Revision. It is important to note that word counts are not an absolute measure, but rather a useful guid, and in general the author must use their discretion when determining if a new article is to be considered a new Derivative Work, a Minor Revision, or a Major Revision. The owner/author controls all rights in the new Work and may do as they wish with it. That said, it is commonly accepted practice that for new Derivative or Major Revision Works, the author should incorporate a citation to the previous work.

For example:

**"This work is based on an earlier work: TITLE, in PUBLICATION, {VOL#, ISS#, (DATE)} © Author, {YEAR}.
http://dx.doi.org/10.1145/{number}"**

If the work is a *Minor Revision, the copyright or exclusive publishing license remains with ACM and the Owner should use best efforts to display the ACM citation,

**"© {Owner/Author {YEAR}. This is a minor revision of the work published in PUBLICATION, {VOL#, ISS#, (DATE)}
http://dx.doi.org/10.1145/{number}"**

The appropriate notice should appear both within the document and in the metadata associated with the document. Instructions for how to do this will be found in the instructions for authors in ACM's various publications.

Solicited Works

From time to time, ACM solicits works for publication. Examples are columns, invited works, award lectures, and keynote speeches. ACM asks authors of such works not to distribute copies or post these works on their Home Pages until ACM has published them. Authors who wish to circulate before publication should get permission from ACM. ACM considers lectures and speeches to be published at the time they are given.

PERMISSIONS

ACM grants gratis permission for individual digital or hard copies made without fee for use in academic classrooms and for use by individuals in personal research and study. Further reproduction or distribution requires explicit permission and possibly a fee.

ACM is now a signatory of the **STM Permission Guidelines Initiative**, which supports an approach to research based on common decency, respect, fairness and mutual trust. These Guidelines are built to allow Signatory STM Publishers to use limited amounts of material in other original published works without charge, and with a minimum of effort needed for permissions clearance. ACM joined the initiative in 2022 to lower the burden on authors to obtain third party permissions when authoring works for ACM and third party publishers.

All copies should carry the original citation, the appropriate copyright and notice of permission on the first page or initial screen of the document. (See **§2.2 Copyright Notice**.)

Most permission requests should go through ACM's automated rights system available in the ACM Digital Library and pointed to by **permissions@acm.org**. Requests that cannot be handled through the online system will take longer to resolve: requestors may expect a response to their inquiry within seven business days.

Fair Use for Educational Purposes

Definition of classroom use: Copying and distributing single works by a university/college instructor, where no fee is charged to the students, and the distribution is limited to students enrolled in a university/college course and their instructors.

- **Course Material** - Permission granted without fee if the course material is produced without charge to the student. (See Commercially produced Course Packs below.)

- **Electronic Reserves** - Permission granted without fee provided the library or institution has an authentication mechanism for controlled access to the server and a license to the ACM-published work. A college, university or other accredited institution may place a copy of a definitive Version of Record of the work in its library's electronic reserves for the duration of its educational needs for that work, provided that access is limited to its enrolled students (including those in its distance learning programs), faculty, and staff. Those institutions without a current license to the work should contact **permissions@acm.org**.

- **Distance Learning** - Permission granted without fee for distance learning students enrolled at the institution. They have the same access rights to those ACM copyrighted materials licensed by their institution as any other student. Since institutional access is authenticated by IP address, it is up to the institution to provide a proxy server for its remote users, and to register the IP address of that proxy with ACM.

- **Interlibrary Loan (ILL)** - Permission granted without fee for an institution with an ACM Digital Library license to download and print works for Interlibrary Loan. The Digital Library may be used as the source for the printed copy. The loan of the work is limited to printed copies, as part of normal library functions.

- **Walk-Ins** - Permission granted without fee for access to all ACM publications, print or electronic, by all members of the community which a subscribing library is charted to serve.

- **Open Access / Creative Commons Material** - Permission is granted without fee, provided proper attribution is given to the Author(s) and Publisher at the time of use.

Commercial Republication

Definition of commercial republication: Any use that is not personal or non-profit educational use. Includes reprinting by trade and scholarly publishers, and use in corporate settings and their web sites, both internal and external. No direct profit need be realized from the publication or sale of ACM material.

Commercial use normally requires a license and payment of release fees. All reproductions other than those listed in this document require specific permission and a fee payable to ACM. This includes republishing in textbooks, commercially-produced course packs sold to students, anthologies, and other edited publications, and posting or other electronic distributions, unless use is done in connection with the **STM Permission Guidelines Initiative**.

- **Commercially Produced Course Packs** - Use of copyrighted or licensed material in course packs sold to students requires an appropriate license. Send requests to **permissions@acm.org** or go to **http://www.copyright.com**.

- **Print permission** - A grant of permission involves consultation with the lead author of the work, the publisher's agreement to pay the required fees, and prominent display of the proper credit acknowledgment.

- **Electronic permission** - Rules for commercial distribution will apply unless the request falls under educational use as defined above. Fees for internal and external commercial posting of ACM published material are tied to the term of the license. All postings must include pointers to the correct Citation Page in the ACM Digital Library.

- **Multiple copies** - Producing multiple copies of ACM copyrighted or licensed works for distribution to more than ten peers, co-workers, clients, etc. requires a transactional license from the CCC and payment of the required per copy fee Send requests to **permissions@acm.org** or go to**http://www.copyright.com**.

- **Software** - Owners/Authors of software grant ACM a non-exclusive permission to publish and manage all rights and permissions themselves.

3rd Party Permissions

Lastly, another major change relating to how ACM handles rights and permissions is that ACM has adopted **STM Permissions Guidelines**, which simplifies the process for third parties (including researchers) to reuse ACM published content in new works under development. This is a broad-based publisher initiative that includes the vast majority of publishers in computing literature. Other signatories of these guidelines are listed **here**. It is our goal to simplify the process of publishing with ACM, and we welcome your feedback after the above steps have been implemented.

ACM publications staff will monitor requests for permission not handled by ACM's automated permissions system which is accessed via the ACM Digital Library. Persons granted permission to copy an ACM published work should display the appropriate Publication Notice followed by: "Included here by permission."

Edited Collections

Edited collections such as conference proceedings and newsletters are copyrighted as a whole by ACM. Going forward after January 1, 2023, authors will retain the copyright of individual components of those Works, such as articles, letters-to-the-editor, abbreviated works, etc. For these individual components, ACM will obtain either an exclusive or non-exclusive permission to publish (conveyed tacitly or by the ACM Permission Form) that permits publication in both print and online forms, and also grants ACM the right to transform the work into any formats as necessary for use within the ACM Digital Library or other media.

No ACM-copyrighted or exclusively licensed collection may be posted for open distribution without prior permission from ACM and before it has been included in the ACM Digital Library. Approved distributions must include a notice of this permission along with the copyright notice for the Work.

Links

ACM treats links as citations (references to objects) rather than as incorporations (embedding of objects). Permission is not needed to create links to citations in The ACM Digital Library or Online Guide to Computing Literature. ACM encourages the widespread distribution of links to the definitive Version of Records of its copyrighted works in the ACM Digital Library and does not require that authors obtain prior permission to include such links in their new works.

However, someone who creates a work or a service whose pattern of links substantially duplicates an ACM-copyrighted volume or issue should get prior permission from ACM. One example: the creator of "A Table of Contents for the Current Issue of TODS" -- consisting of citations and active links to author-versions of the works in the latest issue of TODS -- needs ACM permission because that creator is reproducing an ACM-copyrighted work. If all the links in the "Table of Contents" pointed to the ACM-held definitive Version of Records, ACM would normally give permission because then the new work advertises an ACM work. To avoid misunderstandings, consult with ACM before duplicating an ACM work via links.

If an author wishes to embed a copyrighted object---rather than a link---in a new work, that author needs to obtain the copyright holder's permission.

Distributions From non-ACM Servers

Service providers do not need to obtain prior permission from ACM to locate and dispense links to the ACM-held definitive Version of Records of works, but they do need permission if they are making, collecting, or distributing copies of ACM-copyrighted or licensed works.

Other Related Policies

Conference Publication Policy

Please see the **Conference Publication Policy** for additional expectations related spcifically to ACM Conference Publications.

Inappropriate Content Policy

Please see **ACM's Inappropriate Content Policy**

Submitting and Investigating Potential Violations of this Policy

See **Policy on Submitting and Investigating Claims**

Confidentiality Policy

See **Confidentiality Policy**

Communicating Results of Investigations

See **Policy on Communicating Results of Investigations**

Appealing Violation Decisions

See **Appealing Policy Violation Decisions**

Contact ACM

The ACM Director of Publications should be contacted for any:

- Questions about the interpretation of this policy
- Questions about appeals of decisions
- Requests for deviations from, or extensions to, this policy
- Reporting of egregious behavior related to this policy, including purposeful evasion of the policy or false reporting

Mailing address:
ACM Director of Publications
Association for Computing Machinery
1601 Broadway, 10th Floor
New York, NY 10019-7434
Phone: +1-212-626-0659
Or via email:
scott.delman@hq.acm.org

A Consistent and Efficient Evaluation Strategy for Attribution Methods

Yao Rong [*1] **Tobias Leemann** [*1] **Vadim Borisov** [1] **Gjergji Kasneci** [1] **Enkelejda Kasneci** [1]

Abstract

With a variety of local feature attribution methods being proposed in recent years, follow-up work suggested several evaluation strategies. To assess the attribution quality across different attribution techniques, the most popular among these evaluation strategies in the image domain use pixel perturbations. However, recent advances discovered that different evaluation strategies produce conflicting rankings of attribution methods and can be prohibitively expensive to compute. In this work, we present an information-theoretic analysis of evaluation strategies based on pixel perturbations. Our findings reveal that the results are strongly affected by information leakage through the shape of the removed pixels as opposed to their actual values. Using our theoretical insights, we propose a novel evaluation framework termed Remove and Debias (ROAD) which offers two contributions: First, it mitigates the impact of the confounders, which entails higher consistency among evaluation strategies. Second, ROAD does not require the computationally expensive retraining step and saves up to 99 % in computational costs compared to the state-of-the-art. We release our source code at `https://github.com/tleemann/road_evaluation`.

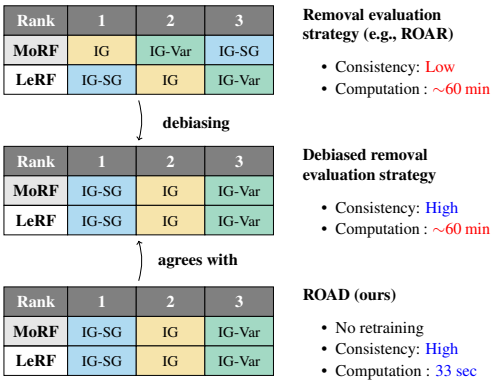


Figure 1. Comparison between previous removal and retraining evaluation strategies (**Top**) and ours (**Bottom**). Previously, rankings of different attribution methods, Integrated Gradients (IG) (Sundararajan et al., 2017) and its two variants SmoothGrad (IG-SG) (Smilkov et al., 2017), SmoothGrad^2 (IG-SQ) (Hooker et al., 2019), are highly inconsistent with respect to hyperparameters such as the removal orders Most Relevant First (MoRF) and Least Relevant First (LeRF). Our ROAD strategy achieves a consistent ranking using only 1% of the previously required resources.

1. Introduction

Explainable Artificial Intelligence (XAI) has become a widely discussed research topic (Adadi & Berrada, 2018). Specifically, feature attribution methods (Springenberg et al., 2015; Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017; Selvaraju et al., 2017) that quantify the importance of input features to a model's decision are widely used. Such local explanations can help to analyze and debug predictive models (Bhatt et al., 2020b; Adebayo et al., 2020), e.g., in the medical domain (Eitel et al., 2019), in recommender systems (Afchar & Hennequin, 2020), and many other applications. With an increasing number of feature attribution methods proposed in the literature, the need for sound strategies to evaluate these methods is also increasing (Nguyen & Martínez, 2020; Hase & Bansal, 2020; Yeh et al., 2019; Hooker et al., 2019).

Evaluation strategies, proposed to compare different attribution methods, commonly follow an ablation approach by perturbing the input features, e.g., image pixels, deemed most or least important. Specifically, perturbing pixels assigned high importance should decrease predictive quality whereas perturbing unimportant pixels, should hardly affect the predictions. These measures aim to capture the *fidelity* of explanations (Tomsett et al., 2020), i.e., how well the explanation genuinely reflects the prediction of the

[*]Equal contribution [1]Department of Computer Science, University of Tübingen, Tübingen, Germany. Correspondence to: Yao Rong <yao.rong@uni-tuebingen.de>, Tobias Leemann <tobias.leemann@uni-tuebingen.de>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

underlying model. Fidelity based on a single data sample is known as local fidelity, while global fidelity is measured on the whole data set (Tomsett et al., 2020).

The outcome of evaluation strategies is highly sensitive to parameters such as the perturbation function and order. Depending on the order chosen, i.e., *most relevant pixels first* or *least relevant pixels first*, such removal strategies often lead to highly contradictory results. For instance, local attribution methods that seem to perform well in one order may perform rather poorly in the other (Tomsett et al., 2020; Haug et al., 2021; Hooker et al., 2019). This inconsistency makes it hard for researchers to impartially compare between different attribution methods and it is not well understood where the inconsistencies stem from. Moreover, for conducting the global fidelity check, a retraining step is required by some methods (Hooker et al., 2019), which is prohibitively expensive in practice (Tomsett et al., 2020). These two drawbacks and our improvements are illustrated in Figure 1.

In this paper, we aim to overcome these shortcomings and make the evaluation more consistent and efficient. To this end, we propose a new debiased strategy that compensates for confounders causing inconsistencies. Furthermore, we show that in the debiased setting, we can skip the retraining without significant changes in the results. This results in drastic efficiency gains as shown in the lower part of Figure 1. We argue that it is crucial for the community to have sound evaluation strategies that do not suffer from limited accessibility due the required compute capacity. Specifically, we make the following contributions:

- We examine the mechanisms underlying the evaluation strategies based on perturbation by conducting a rigorous information-theoretic analysis, and formally reveal that results can be significantly confounded.
- To compensate for this confounder, we propose the Noisy Linear Imputation strategy and empirically prove its efficiency and effectiveness. The proposed strategy significantly decreases the sensitivity to hyperparameters such as the removal order.
- We generalize our findings to a novel evaluation strategy, ROAD (RemOve And Debias), which can be used to objectively and efficiently evaluate several attribution methods. Compared to previous evaluation strategies requiring retraining, e.g., Remove and Retrain (ROAR) (Hooker et al., 2019), ROAD saves 99 % of the computational costs.

2. Related Work

There is a plethora of works on different explanation techniques (Tjoa & Guan, 2020), especially attribution methods that assign importance scores to each input features. Popular approaches have been proposed by Springenberg et al. (2015); Lapuschkin et al. (2015); Ribeiro et al. (2016); Kasneci & Gottron (2016); Sundararajan et al. (2017); Fong & Vedaldi (2017); Shrikumar et al. (2017); Smilkov et al. (2017); Petsiuk et al. (2018); Adebayo et al. (2018); Chen et al. (2018); Xu et al. (2020); Covert et al. (2021), and many more.

With the growing number of attribution methods, various scholars have presented desiderata that explanations should fulfill (Bhatt et al., 2020a; Nguyen & Martínez, 2020; Fel et al., 2021; Afchar et al., 2021; Nauta et al., 2022). Doshi-Velez & Kim (2017) consider two subcategories in this field, namely *human-grounded* metrics relying on human judgment and *functional-grounded* metrics. The latter do not require a human-generated ground truth that can be hard or even impossible to obtain. Metrics of this type frequently rely on the idea that if the most important part of the image is changed, the output probability of the given black-box model should also change in return. Examples include the Sensitivity-n measure proposed by Ancona et al. (2017) and the infidelity and max-sensitivity metrics by Yeh et al. (2019). Samek et al. (2016) and Petsiuk et al. (2018) also propose to perturb the pixels in the input image according to the importance scores. However, Hooker et al. (2019) show that the perturbation introduces artifacts and results in a distribution shift, putting these no-retraining approaches in question. They propose the Remove and Retrain (ROAR) framework with an extensive model retraining step to adapt to the distribution shift. Therefore, we distinguish between evaluation methods with *retraining* and *no-retraining* approaches. ROAR has been adopted in several recent studies (Hartley et al., 2020; Izzo et al., 2020; Meng et al., 2021; Schramowski et al., 2020; Srinivas & Fleuret, 2019) and variations are being proposed in concurrent work (Shah et al., 2021).

Only few papers have used and compared different evaluation strategies for attribution methods and a sound theoretical explanation for the differences between them is still missing. Sturmfels et al. (2020) assess different baselines for feature attribution applying the Integrated Gradient method (Sundararajan et al., 2017). They also observe that changing the hyperparameter settings can lead to varying results. Haug et al. (2021) draw the same conclusion for attributions on tabular data. Tomsett et al. (2020) compute the consistency among different, no-retraining evaluation strategies and report an alarmingly low agreement. In this work, we conduct a rigorous analysis of reasons for existing inconsistency and provide a solution to reduce it, which is not studied in previous works. Moreover, our solution also reduces high computational costs caused by retraining.

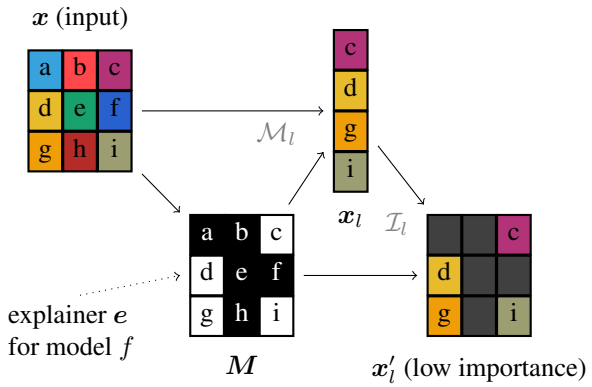


Figure 2. Our analytical model of feature removal evaluation (MoRF order shown): The input image $\boldsymbol{x}$ (9 pixels a–i) is explained by an explanation method that returns a mask $\boldsymbol{M}$ indicating important pixels (black). The remaining, less important pixel values $\boldsymbol{x}_l$ can be extracted from the image using the masking operator $\mathcal{M}_l$ and transformed via the imputation operator $\mathcal{I}_l$ to an imputed variant of the input $\boldsymbol{x}'_l$, which determines the evaluation outcome. This model allows to separate the information in the feature *values* from that contained in the binary mask $\boldsymbol{M}$.

3. Preliminaries

In this section, we formally define the pixel-perturbation strategies considered by the following analysis.

3.1. Retraining Evaluation Strategies

We consider a pixel removal strategy, where pixels are successively replaced by imputed values. Consistent with the literature (Tomsett et al., 2020; Samek et al., 2016), we consider two removal orders: **MoRF** (Most Relevant First) or **LeRF** (Least Relevant First), where the subsequent removal starts with the most important pixels for the former and the least important ones for the latter. We now provide a formal definition of MoRF with retraining, i.e., the ROAR benchmark, that will be used throughout our analysis. We always use the MoRF order in the analysis presented in this paper. However, an analogous analysis of its counterpart LeRF is possible without much additional effort and can be found in the appendix.

To ease our derivations, we describe the procedure by a series of operations that can be analyzed independently. A classifier $f : \mathbb{R}^d \to \{1, \dots, c\}$ maps inputs $\boldsymbol{x} \in \mathbb{R}^d$ to labels $C \in \{1, \dots, c\}$, where c is the number of classes. A feature attribution explanation for the prediction assigns each input dimension an importance value. In the MoRF setting, the features are ordered in a descending order of importance. Subsequently, the k most important features per instance are selected for removal, where $0 \leq k \leq d$ is successively increased during the benchmark. However, for the moment we consider only one fixed value of k. Thus,

C	Class label random variable
I	Mutual information
$\mathcal{I}$	Imputation operator
$\boldsymbol{M}$	Binary mask in $\{0,1\}^d$
$\mathcal{M}$	Mask selection operator (takes out relevant features)
$\boldsymbol{x}$	Input features in $\mathbb{R}^d$
$\boldsymbol{x}_l$	Low importance features only in $\mathbb{R}^{d-k}$
$\boldsymbol{x}'_l$	Imputed low importance features in $\mathbb{R}^d$

Table 1. Overview of the notation used in this work.

we can model the explanation $\boldsymbol{e}_k$ as a choice of features via a binary mask $\boldsymbol{M} = \boldsymbol{e}_k(f, \boldsymbol{x}) \in \{0,1\}^d$, with the corresponding value set to one, if the corresponding feature is among the top-k, and to zero otherwise. Furthermore, suppose $\mathcal{M}_l : \{0,1\}^d \times \mathbb{R}^d \to \mathbb{R}^{d-k}$ to be the selection operator for the least important dimensions indicated in the mask and $\boldsymbol{x}_l = \mathcal{M}_l(\boldsymbol{M}, \boldsymbol{x})$ to be a vector containing only the remaining features as shown in Figure 2. We suppose that the features preserve their internal order in $\boldsymbol{x}_l$, i.e., features are ordered ascendingly by their original input indices. This definition allows to separately consider the information flow in the feature *mask* $\boldsymbol{M}$ and that in the feature *values* $\boldsymbol{x}_l$.

The ROAR approach measures the accuracy of a newly trained classifier f' on modified samples $\boldsymbol{x}'_l := \mathcal{I}_l(\boldsymbol{M}, \boldsymbol{x_l})$, where $\mathcal{I}_l : \{0,1\}^d \times \mathbb{R}^{d-k} \to \mathbb{R}^d$ is an imputation operator that redistributes all inputs in the vector $\boldsymbol{x}_l$ to their original positions and sets the remainder to some filling value. In the special case of zero imputation, $\boldsymbol{x}'_l = \mathcal{I}_l(\boldsymbol{M}, \mathcal{M}_l(\boldsymbol{M}, \boldsymbol{x})) = (1 - \boldsymbol{M}) \odot \boldsymbol{x}$. This means the top-$k$ features are discarded. For a better evaluation result, the accuracy should drop quickly with increasing k, indicating that the most influential features were successfully removed.

3.2. Information Theory

We now briefly revisit the central concepts of information theory that will be handy for our analysis and introduce the notation. The fundamental quantity in information theory is the entropy H of a discrete random variable X with support $\operatorname{supp}\{X\}$,

$$H(X) := -\sum_{x \in \operatorname{supp}\{X\}} P(X = x) \log P(X = x). \quad (1)$$

The entropy corresponds to the information gained through observation of a realization of this variable. If the random variable considered can be easily inferred, we use $p(x)$ as a shorthand for $P(X = x)$. Furthermore, we denote the joint entropy between random variables X and Y by $H(X, Y)$, which is equivalent to the entropy of their joint distribution. In accordance with Cover & Thomas (2006), we always

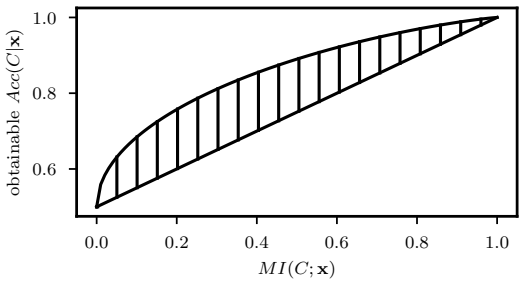


Figure 3. Relation between Mutual Information (MI) and obtainable accuracy for the two-class problem with equal class priors. The knowledge of the MI $I(\boldsymbol{x}; C)$ implies strong bounds for the obtainable accuracy. This connection permits to use MI as a surrogate for the obtainable accuracy in the perturbation strategy in our analysis. Figure adapted from Meyen (2016).

separate random variables by comma to denote the joint distribution of multiple of variables.

The conditional entropy $H(X|Y)$ is the expected amount of information left in a variable, given the observation of a condition Y. The most central concept in our analysis will be mutual information (MI), i.e., the amount of information in one random variable shared with another. For example, by $I(\boldsymbol{x}; C) := H(C) - H(C|\boldsymbol{x})$, we denote the MI between the complete feature vector and the class variable C. We separate arguments by a semicolon and allow single random variables or sets of random variables as arguments to all the defined quantities. For sets, we always consider the joint distribution of their member variables. Please confer Cover & Thomas (2006) for a more profound introduction. We provide a short overview of our notation in Table 1.

4. Analysis

In this section, we show that the pixel perturbation strategies are susceptible to a previously unknown confounder: The binary mask itself can leak class information that might in not be present in the feature values. After making the connection between the accuracy and mutual information as a theoretical tool in Section 4.1, we formally derive the confounder and identify this leakage on real data in Section 4.2. We subsequently show how to mitigate it through Minimally Revealing Imputation in Section 4.3.

4.1. On the Relation Between Accuracy and Mutual Information

To begin our analysis of the presented strategies and their underlying mechanisms, we first establish the relation between classification accuracy and the mutual information. It is well-known that the classification performance of an optimal classifier in the Bayesian sense (assigning the class with the highest posterior) is dependent on the MI between features and labels (Hellman & Raviv, 1970; Vergara & Estévez, 2014; Meyen, 2016). Nevertheless, the relationship is not a function, but comes in form of upper and lower bounds of the obtainable accuracy. For the simple two-class problem, the bounds are shown in Figure 3 (cf. Appendix A.1 for derivations). They impose strong limits on the optimal classification performance, if the mutual information $I(\boldsymbol{x}; C)$ is known.

For the pixel removal strategies that use retraining, this allows us to analyze the frameworks using MI as a surrogate for the attainable accuracy because higher MI almost always leads to higher accuracy. In the MoRF setting with retraining, $I(\boldsymbol{x}'_l; C)$ will play a key role, because it quantifies the information left in the least important features and thus determines obtainable accuracy which is the outcome of the evaluation. Low mutual information $I(\boldsymbol{x}'_l; C)$ results in a sharp drop in accuracy and good benchmarking results:

$$\downarrow I(\boldsymbol{x}'_l; C) \;\; \Rightarrow \uparrow \text{MoRF benchmark.}$$

Therefore, in the MoRF setting low mutual information of $\boldsymbol{x}'_l$ and C is desirable[1].

4.2. Class Information Leakage through Masking

We demonstrate that it is easily possible to leak class information only through the mask's shape and to harshly manipulate the evaluation score. Therefore, we start by separating the influence of the mask from that of the feature values. Our derivation relies on the multi-information $I(C; \boldsymbol{x}'_l; \boldsymbol{M})$, which is defined by Vergara & Estévez (2014) as follows:

$$I(C; \boldsymbol{x}'_l; \boldsymbol{M}) = I(C; \boldsymbol{x}'_l|\boldsymbol{M}) - I(C; \boldsymbol{x}'_l) \tag{2}$$
$$I(C; \boldsymbol{x}'_l; \boldsymbol{M}) = I(C; \boldsymbol{M}|\boldsymbol{x}'_l) - I(C; \boldsymbol{M}). \tag{3}$$

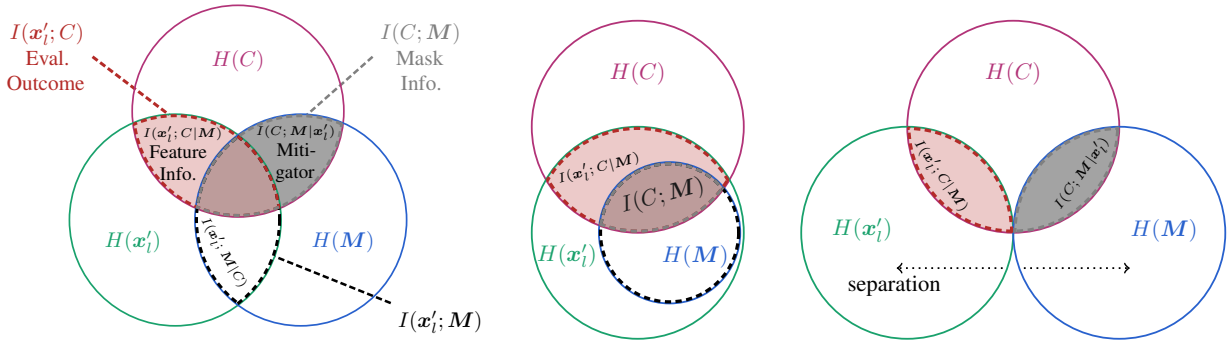
Setting Equation (2) and Equation (3) equal, we arrive at the identity:

$$\underbrace{I(\boldsymbol{x}'_l; C)}_{\text{Eval. Outcome}} = \underbrace{I(C; \boldsymbol{x}'_l|\boldsymbol{M})}_{\text{Feature Info.}} + \underbrace{I(C; \boldsymbol{M})}_{\text{Mask Info.}} - \underbrace{I(C; \boldsymbol{M}|\boldsymbol{x}'_l)}_{\text{Mitigator}}. \tag{4}$$

The quantities involved are visualized in Figure 4a. The first term "Feature Information" is the class information contained in the features (and not in the mask) that we wish to estimate. The second term "Mask Information" shows that class-discriminative information in the mask can have a high impact on the result. This influence can be compensated by the "Mitigator" term.

Class Information Leakage If the Mask Information term is superior to the Mitigator, $I(C; \boldsymbol{M}) > I(C; \boldsymbol{M}|\boldsymbol{x}'_l)$,

[1] In LeRF, a higher accuracy and thus higher $I(\boldsymbol{x}'_l; C)$ is beneficial



(a) General Case (b) Invertible Imputation (c) Minimally Revealing Imputation

Figure 4. The Evaluation Outcome $I(\boldsymbol{x}_l'; C)$ (red area), is confounded by the Mask Information $I(C; \boldsymbol{M})$ (gray area) when there is some overlap (a). Only the Feature Information $I(\boldsymbol{x}_l'; C|\boldsymbol{M})$, the part of the Outcome not overlapping (light red area), should actually be assessed. In the worst case (which we term Invertible Imputation), the Mask Information is entirely contained in the Outcome (b). Separating the information in the imputed image $\boldsymbol{x}_l'$ and the mask $\boldsymbol{M}$ allows to reduce the overlap and the influence (c).

the evaluation outcome is unfairly increased to a value not justified by the selected features. We term this phenomenon *Class Information Leakage*, as some discriminative information is "leaked" through the used binary mask $\boldsymbol{M}$.

The Mitigator can entirely vanish when the mask is perfectly inferable from the imputed image $\boldsymbol{x}_l'$. This results in a non-compensated effect of Class Information Leakage. We define this imputation operation as follows:

Condition 4.1. *Invertible Imputation. Let* $\mathcal{I}_l : \{0,1\}^d \times \mathbb{R}^{d-k} \to \mathbb{R}^d$ *be the imputation operator that takes the least important features as an input. We suppose that there are inverse functions* $\mathcal{I}_{l,M}^{-1}$ *and* $\mathcal{I}_{l,x}^{-1}$*, such that*

$$\boldsymbol{x}_l' = \mathcal{I}_l\left(\boldsymbol{M}, \boldsymbol{x}_l\right) \Leftrightarrow \boldsymbol{M} = \mathcal{I}_{l,M}^{-1}(\boldsymbol{x}_l') \wedge \boldsymbol{x}_l = \mathcal{I}_{l,x}^{-1}(\boldsymbol{x}_l').$$

If, for instance, the pixels removed are set to some reserved value indicating their absence, the imputation operator is invertible, as the mask can be reconstructed. Therefore, $H(\boldsymbol{M}|\boldsymbol{x}_l')=H\left(\mathcal{I}_{l,M}^{-1}(\boldsymbol{x}_l')|\boldsymbol{x}_l'\right)=0$. In this case, also the Mitigator $I\left(C; \boldsymbol{M}|\boldsymbol{x}_l'\right) = 0$, because it is bounded by $0 = H(\boldsymbol{M}|\boldsymbol{x}_l') \geq I\left(C; \boldsymbol{M}|\boldsymbol{x}_l'\right) \geq 0$. The Feature Information term is constrained to be positive. Thus, the Mask Information has a non-negligible impact on the Evaluation Outcome because a higher Mask Information term will always increase it. This case is depicted in Figure 4b.

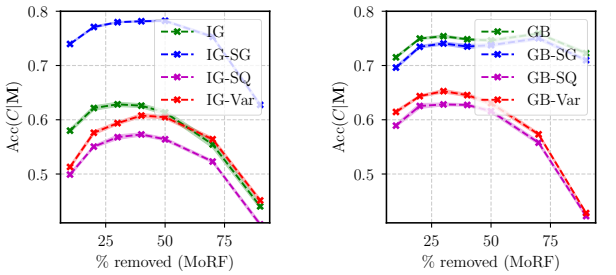
We can create a simple example that shows how evaluation scores are influenced: Imagine a two-class problem that consists of detecting whether an object is located on the left or the right side of an image. A reasonable attribution method masks out pixels on the left or the right depending on the location of the object. In this case, the retraining step can lead to a classifier that infers the class just from the location of the masked out pixels and obtain high accuracy. This explanation map will be rated far worse in MoRF (no accuracy drop) than it might actually be. In the context of amortized explanation methods, a similar finding has been made by Jethani et al. (2021). We theoretically showed that this problem also arises in evaluation strategies and empirically demonstrate that the leakage is significant for popular attribution methods on real data in Section 5.1.

4.3. Reduction of Information Leakage

To tackle this problem, we follow an intuitive approach: If we cannot guarantee that there is no class information contained in the mask itself, we have to stop it from leaking the class information into the imputed images. Therefore, we make sure that the mask used cannot be easily inferred from the imputed image. We would like to set $I(\boldsymbol{x}_l'; \boldsymbol{M}) = 0$, i.e., the mask is independent of the imputed vector allowing to separate the effects as shown in Figure 4c. Unfortunately, this is not possible in general: If both should be dependent on the class label, they will also have to share a minimal amount of information (that regarding the class). However, we can demand conditional independence and make $I(\boldsymbol{x}_l'; \boldsymbol{M})$ as small as possible.

Condition 4.2. *Minimally Revealing Imputation. Let* $\mathcal{I}_l : \{0,1\}^d \times \mathbb{R}^{d-k} \to \mathbb{R}^d$ *be the infilling operator that takes the least important features as an input. Suppose* $\boldsymbol{x}_l'$ *and* $\boldsymbol{M}$ *are independent given the class information* $I\left(\boldsymbol{x}_l'; \boldsymbol{M}|C\right) = 0$ *and* $I\left(\boldsymbol{x}_l'; \boldsymbol{M}\right) \approx 0$.

In this case, $I(C; \boldsymbol{M}) - I\left(C; \boldsymbol{M}|\boldsymbol{x}_l'\right) = I\left(\boldsymbol{x}_l'; \boldsymbol{M}\right) - I\left(\boldsymbol{x}_l'; \boldsymbol{M}|C\right) \approx 0$, which implies $I(C; \boldsymbol{M}) \approx I\left(C; \boldsymbol{M}|\boldsymbol{x}_l'\right)$ (also cf. Figure 4c), indicating that the Mitigator effectively compensates the Mask Information term.



(a) Integrated Gradients (IG)

(b) Guided Backpropagation (GB)

Figure 5. Accuracy of a trained classifier only using the binary masks $\boldsymbol{M}$ without feature values as input on the CIFAR-10 data set. Binary masks $\boldsymbol{M}$ were computed for different variants of IG and GB. Only the masks contain enough information to reach an accuracy of almost up to 80 % (compared to 85 % with full images) highlighting that the feature values do not play an important role in the evaluation. This underlines the necessity to compensate for this confounder.

5. Debiasing Evaluation Strategies for Local Attribution Methods

With the theoretical analysis in Section 4, we can better understand where the biases come from, and thus mitigate them. Building on the derivations, we now show the strong impact of the Class Information Leakage introduced in Section 4.2 on a real-world data set to highlight the necessity to compensate for this confounder. We explain how we reduce its influence by proposing a novel imputation operator termed *Noisy Linear Imputation*.

5.1. Extent of Class Information Leakage

To empirically confirm our findings, we performed experiments on CIFAR-10 (Krizhevsky et al., 2009). We use the same attribution methods as in Hooker et al. (2019): Integrated Gradients (IG) (Sundararajan et al., 2017) and Guided Backprop (GB) (Springenberg et al., 2015) serve as base explanations, and three ensembling strategies for each are used in addition: SmoothGrad (SG) (Smilkov et al., 2017), SmoothGrad^2 (SQ) (Hooker et al., 2019) and VarGrad (Var) (Adebayo et al., 2018). In total, we consider eight attribution methods and provide details and parameters in the supplementary material.

We empirically show that with fixed value imputation with the global mean, the explanation masks are leaking class information. This takes two steps: (1) We show that the Mask Information $I(C; \boldsymbol{M})$ is extremely high. (2) We verify that the Mitigator is small by testing the *Invertible Imputation* Condition, which implies that class information is leaked into the evaluation outcome through $I(C; \boldsymbol{M})$.

To assess the class information in the mask, we train a ResNet-18 (He et al., 2016) that uses only binary masks $\boldsymbol{M}$ (no pixel values $\boldsymbol{x}_l$) to predict the class. As we discussed previously, the accuracy of a classifier can be used as a surrogate for the calculation of MI, which is prohibitively expensive for high-dimensional data. The curves[2] are shown in Figure 5. Stunningly, the mask alone results in high accuracy curves that reach almost 80 % for IG-SG, only some percent below the accuracy of the classifier on the full inputs. This allows us to conclude that the Mask Information $I\left(C; \boldsymbol{M}\right)$ is almost as high as our Evaluation Outcome $I\left(C; \boldsymbol{x}'_l\right)$.

To show that the Mitigator is almost zero which leads to class information leakage, we test the *Invertible Imputation* condition. Therefore, the inverse function $\mathcal{I}^{-1}_{l,M}$ that predicts the imputation mask from the imputed image is required (having this function, finding $\mathcal{I}^{-1}_{l,x}$ is trivial). For the fixed value imputation, an approximate inverse is simple: Setting all pixels in the mask to 0 if the corresponding image pixel has the filling value (which has to be inferred from the distribution). For a stronger verification, we train an imputation predictor network consisting of three convolutional layers, which predicts for each pixel if it was imputed or original. As Figure 6e (blue curve) shows, the miss-classification rate when using fixed value imputation is almost zero, i.e., the network can easily recognize the pixels that were imputed. According to our analysis, in this setting close to *Invertible Imputation*, the Mitigator will be negligibly small.

This leads us to the conclusion that the mask-related leakage fundamentally influences many previous evaluations using fixed value imputation (Shrikumar et al., 2017; Petsiuk et al., 2018; Hooker et al., 2019) and it is essential to stop the information leakage through the masks.

5.2. Debiasing with Noisy Linear Imputation

To reduce the Class Information Leakage, we propose a better-suited imputation operator $\mathcal{I}_l$ that adheres to the *Minimally Revealing Imputation* condition we derived. The remaining process is left unchanged and stays as depicted in Figure 2. However, we face three requirements: (1) We have to get closer to the theoretical condition of Minimally Revealing Imputation. (2) The imputation strategy needs to be highly efficient, since the imputation module has to be run for each image in the data set. (3) We wish to have as few hyper-parameters as possible (preferably none to rule out another confounding factor).

We devise a new strategy called *Noisy Linear Imputation*, which fulfills the above goals. In this way, our model addresses some of the fundamental problems of existing

[2]Standard Errors are indicated by shaded areas in all figures. However, they are often hardly visible due to their low magnitude.

strategies. Intuitively, we search a way to make more subtle imputations that cannot be easily recognized and result in lower $I\left(\boldsymbol{x}_l'; \boldsymbol{M}\right)$. To this end, we suppose that each pixel can be approximated by the weighted mean of its neighbors (cf. Figure 6d) as image pixels are highly correlated[3]:

$$\begin{aligned}\boldsymbol{x}_{i,j} = w_d \left(\boldsymbol{x}_{i,j+1} + \boldsymbol{x}_{i,j-1} + \boldsymbol{x}_{i+1,j} + \boldsymbol{x}_{i-1,j}\right) \\ + w_i \left(\boldsymbol{x}_{i+1,j+1} + \boldsymbol{x}_{i-1,j+1} + \boldsymbol{x}_{i+1,j-1} + \boldsymbol{x}_{i-1,j-1}\right)\end{aligned}$$

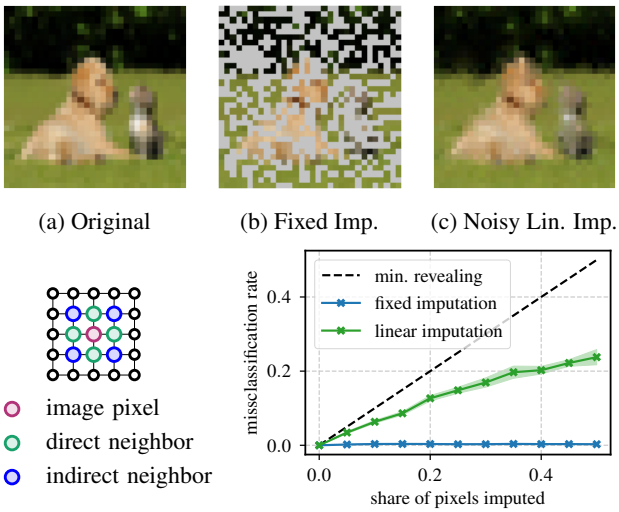
where w_d, w_i are constant coefficients for direct neighbors and indirect, diagonal neighbors. When setting up a single equation for each removed pixel we arrive at an equation system. For known pixels, we directly plug in their values and only consider each removed pixel as an unknown variable. When neighboring pixels are removed, the equations become connected and cannot be solved independently. Nevertheless, the resulting system is sparse and can be efficiently solved, even for a large number of missing pixels. To choose the neighbor weights for the linear interpolation, we draw inspiration from the graph structure (see Figure 6d): Indirect neighbors have distance 2 from the original node in the graph and direct neighbors have distance 1. Hence, we gave the direct neighbors twice the weight of the diagonal ones. Because the weights need to some up to 1 for a weighted interpolation, this leads to $w_d{=}\frac{1}{6}$ and $w_i{=}\frac{1}{12}$. We add a small random noise ($\sigma = 0.1$) to the solution to ensure that the linear dependency cannot be learned by the model.

Figure 6 (top) provides an example of an imputed sample. From the imputed version in Figure 6c, inference on the mask is significantly harder than the one imputed with fixed values as in Figure 6b. We again train the imputation predictor for verification and show the results in Figure 6e. We confirm that our strategy lies significantly closer to the optimal, Minimally Revealing Imputation. Admittedly, there are even more sophisticated imputation strategies, for example building on Generative Adversarial Networks (GANs) such as Generative Adversarial Imputation Nets (GAIN) proposed by Yoon et al. (2018). However, our strategy already achieves considerable improvements and is highly efficient, because it does not require training of a GAN model. For completeness, we include additional experiments with GAN imputation in Appendix B.

6. Experiments

Having established that our Noisy Linear Imputation fulfills its purpose, in this section, we show that it entails even more benefits in practice. We first highlight how it makes results among different evaluation strategies more consistent in Section 6.1. We then present another considerable advantage in Section 6.2: its agreement with a no-retraining evaluation

[3]In fact, for direct and indirect neighbors, $\rho{=}0.89$ and $\rho{=}0.82$ respectively on CIFAR-10



(a) Original (b) Fixed Imp. (c) Noisy Lin. Imp.

(d) Graphical model used to derive our imputation

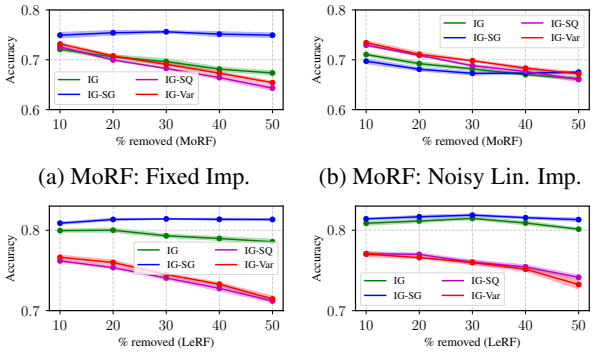
(e) Misclassification rate of the imputation predictor for different shares of randomly imputed pixels (CIFAR-10)

Figure 6. The considered imputation operators. When 50 % of the original image (a) are removed, they can either be imputed by a fixed value (b) or by our proposed Noisy Linear strategy (c,d). Training of an imputation predictor (e) shows that it is much harder to tell which pixels are original and which were imputed when using our proposed imputation model. This is closer to the optimal, minimally revealing imputation (black). Hence, by using imputed samples of this kind, Class Information Leakage is reduced.

strategy is sufficiently high, so that the retraining step is no longer required. We name this debiased and no-retraining evaluation framework ROAD (RemOve And Debias). All experiments in this section were conducted on CIFAR-10 using the eight attribution methods mentioned. We also use Food-101 (Bossard et al., 2014), a large-scale dataset of high-resolution images, to validate the generalizability of our method. To this end, we train over 1000 models from scratch on data imputed using the strategies, explanations and removal percentages. Since the results on Food-101 also support the findings from CIFAR-10, we include them in Appendix D.

6.1. Consistency under Removal Orders

As we aim for evaluation strategies that are less prone to the hyperparameter setting and allow for a consistent ranking, we study the consistency of evaluation results under the different removal orders MoRF and LeRF. Figure 7 depicts the obtained curves (using "Retrain"). For a clear view, we only show four curves of attribution methods based on IG with retraining and up to 50% pixels are removed. We include the full curves for the IG with its derivatives as well as GB with derivatives in Appendix C. The results using the common fixed value imputation shown in Figure 7a and



(a) MoRF: Fixed Imp. (b) MoRF: Noisy Lin. Imp.

(c) LeRF: Fixed Imp. (d) LeRF: Noisy Lin. Imp.

Figure 7. Consistency comparison using fixed value vs. Noisy Linear Imputation. The higher accuracy is better in LeRF, while the lower is better in MoRF. Comparing (a) and (c), fixed value imputation gives different rankings in MoRF and LeRF orders: IG-SG is the best in LeRF but the worst in MoRF. Comparing (b) and (d), Noisy Linear Imputation changes the outcome considerably and yields a consistent ranking in MoRF and LeRF.

Figure 7c. The results with our Noisy Linear Imputation are shown in Figure 7b and Figure 7d. In MoRF, a sharp drop in the beginning indicates a better attribution method, while a slight drop is desirable in LeRF. Hence, using fixed imputation, the ranking in MoRF is IG, IG-Var, IG-SQ, IG-SG, whereas the ranking in LeRF is IG-SG, IG, IG-SQ, and IG-Var. We see, for instance, that IG-SG is the worst in MoRF and the best in LeRF. When using the Noisy Linear Imputation, the inconsistency vanishes. The ranking in MoRF is: IG-SG, IG, IG-SQ, and IG-Var, which is the same as in LeRF.

We quantitatively compute the consistency among all eight attribution methods with and without retraining. Concretely, we compute the ranks (from 1=best to 8=worst) of our explanation methods for each percentage of perturbed pixels. We then calculate the Spearman Rank correlation between different evaluation strategies. As shown in Table 2, the correlation score of the fixed value imputation is -0.01 when using retraining and 0.01 when no retraining is applied. This indicates no consistency in the rankings. When we deploy our Noisy Linear Imputation, the results change drastically: The correlation score is improved to 0.61 and 0.58 with and without retraining, respectively. This might imply that the information leakage is responsible for a major share of the inconsistency.

6.2. Efficiency

When we apply our Noisy Linear Imputation, we additionally reduce the difference between evaluation with and without retraining. This can be attributed to the reduced distribution shift incurred when using an almost *Minimally Revealing Imputation*. If all pixels were perfectly imputed,

Retrain MoRF vs. LeRF		No-Retrain MoRF vs. LeRF	
fixed	lin	fixed	lin
-0.01±0.01	**0.61**±0.01	0.01±0.00	**0.58**±0.01

Table 2. Spearman rank correlation between evaluation strategies. There is almost no agreement between MoRF and LeRF when using fixed imputation (as in previous works). When using our imputation ("lin"), consistency across MoRF and LeRF orders increases drastically.

MoRF Retain vs. No-Retr.		LeRF Retain vs. No-Retr.	
fixed	lin	fixed	lin
0.15±0.01	**0.84**±0.01	0.09±0.01	**0.94**±0.01

Table 3. Spearman rank correlation between evaluation with and without retraining. Our Noisy Linear Imputation ("lin") also results only in marginal differences between "Retrain" and "No-Retrain". We conclude that the retraining step is no longer necessary.

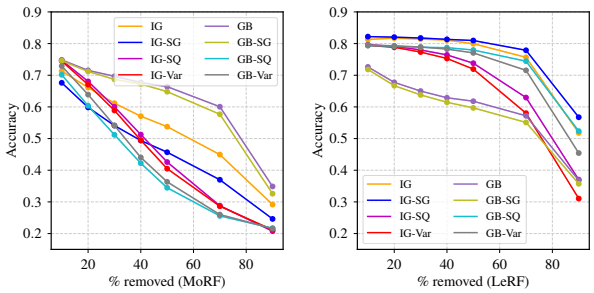
the resulting image would not be out-of-distribution. Since we are interested in the rankings of attribution methods, we again compute Spearman correlation between the rankings obtained with and without retraining and show it in Table 3. The order remains almost always intact between the "Retrain" with Noisy Linear Imputation and the "No-Retrain" variant with Noisy Linear Imputation resulting in a rank correlation of 0.84 in using MoRF and 0.94 in LeRF. This leads us to the conclusion that "No-Retrain" and "Retrain" end up with a highly similar ranking when using Noisy Linear Imputation. Thus, we conclude that the retraining step is not longer justified and can be skipped without significant distortion of the results. Qualitative results are shown in Appendix C.3, cf. Figure 17 (CIFAR-10) and Figure 23 (Food-101).

These results allow us to introduce a novel evaluation framework. We refer to the removal with Noisy Linear Imputation and no retraining as ROAD – Remove and Debias. We showed that ROAD is highly consistent with the compensated results of the ROAR, but comes at an enormous advantage: The retraining step is no longer required. This permits to save a vast amount of computation time. In our experiments, evaluation using the ROAD took only $0.7\,\%$ of the resources required for ROAR, as given by the runtimes in Table 4 obtained on the same hardware (single Nvidia GTX 2080Ti and 8 Cores).

In the end, we illustrate the evaluation results using ROAD among all eight attribution methods in MoRF and LeRF in Figure 8. In MoRF, the best ones are IG-SG, GB-SQ, GB-Var and IG, which have lower accuracies in the beginning, whereas they have higher accuracies in LeRF.

Strategy	Retrain		No-Retrain	
	fixed†	lin	fixed	lin⋆
Time	3903±117 s	4686±2 s	18.0±0.1 s	33.3±0.1 s
Relative	100 %	120 %	0.5 %	0.9 %

Table 4. Mean runtime (5 runs) for evaluating a single explanation method (IG). † refers to ROAR, and ⋆ to our ROAD.



(a) MoRF (b) LeRF

Figure 8. Evaluation results in MoRF (a) and LeRF (b) using our ROAD framework.

GB and GB-Var both perform badly in MoRF and LeRF. We see that some inconsistencies still remain, which cannot be compensated by the current imputation. However, the evaluation strategies might also consider different characteristics of an attribution method (e.g., one might be particularly good at identifying irrelevant pixels), which is why perfect agreement might not even be desirable.

7. Conclusion and Outlook

We introduced ROAD, an evaluation approach for measuring global fidelity among attribution explanations. ROAD comes with two key advantages over existing methods: (1) it is highly efficient, e.g., permitting a 99% runtime reduction w.r.t. ROAR, and (2) it circumvents the Class Information Leakage issue, which was thoroughly analyzed in this work. We believe the ROAD framework will be beneficial to the research community because it unifies several methods and is more consistent under varying removal orders. Moreover, it is broadly accessible due to its low resource requirements. ROAD is open-source[4], and can be readily implemented in practical use-cases. Going forward, we plan to investigate more sophisticated imputation models in ROAD as well as other evaluation metrics besides fidelity.

Acknowledgements

We acknowledge the support by the Cluster of Excellence - Machine Learning: New Perspectives for Science, EXC number 2064/1 - Project number 390727645, and the support of the Training Center for Machine Learning (TCML) Tübingen, funded by the German Federal Ministry of Education and Research (BMBF) with grant number 01IS17054, which provided substantial resources for running our large-scale Food-101 experiment.

[4]An official implementation is also included in the Quantus framework (Hedström et al., 2022)

References

Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

Afchar, D. and Hennequin, R. Making neural networks interpretable with attribution: application to implicit signals prediction. In *Fourteenth ACM Conference on Recommender Systems*, pp. 220–229, 2020.

Afchar, D., Guigue, V., and Hennequin, R. Towards rigorous interpretations: a formalisation of feature attribution. In *International Conference on Machine Learning*, pp. 76–86. PMLR, 2021.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3016 –3022, 2020a.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020b.

Bossard, L., Guillaumin, M., and Gool, L. V. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-shapley and c-shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2018.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley and Sons, 2006. doi: 10.1002/047174882X.

Covert, I., Lundberg, S., and Lee, S.-I. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Eitel, F., Ritter, K., Alzheimer's Disease Neuroimaging Initiative (ADNI), et al. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2019.

Fel, T., Vigouroux, D., Cadène, R., and Serre, T. How good is your explanation? algorithmic stability measures to assess the qualityof explanations for deep neural networks. 2021.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Hartley, T., Sidorov, K., Willis, C., and Marshall, D. Explaining failure: Investigation of surprise and expectation in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 12–13, 2020.

Hase, P. and Bansal, M. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, 2020.

Haug, J., Zürn, S., El-Jiz, P., and Kasneci, G. On baselines for local feature attributions. *AAAI Workshop on Explainable Agency in AI Workshop*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M.-C. Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations. *arXiv preprint arXiv:2202.06861*, 2022.

Hellman, M. E. and Raviv, J. Probability of Error, Equivocation, and the Chernoff Bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

Hooker, S., Erhan, D., Kindermans, P. J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Izzo, C., Lipani, A., Okhrati, R., and Medda, F. A baseline for shapely values in mlps: from missingness to neutrality. *arXiv preprint arXiv:2006.04896*, 2020.

Jethani, N., Sudarshan, M., Aphinyanaphongs, Y., and Ranganath, R. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467. PMLR, 2021.

Kachuee, M., Karkkainen, K., Goldstein, O., Darabi, S., and Sarrafzadeh, M. Generative imputation and stochastic prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Kasneci, G. and Gottron, T. Licon: A linear weighting scheme for the contribution ofinput variables in deep artificial neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 45–54, 2016.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

Meng, C., Trinh, L., Xu, N., and Liu, Y. Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *arXiv preprint arXiv:2102.06761*, 2021.

Meyen, S. Relation between classification accuracy and mutual information in equally weighted classification task. Master's thesis, University of Hamburg, 2016. URL https://osf.io/zru7b/.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation

methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

Nguyen, A. P. and Martínez, M. R. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shah, H., Jain, P., and Netrapalli, P. Do input gradients highlight discriminative features?, 2021.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, ICML*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sturmfels, P., Lundberg, S., and Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Tjoa, E. and Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6021–6029, 2020.

Vergara, J. R. and Estévez, P. A. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 2014.

Xu, S., Venugopalan, S., and Sundararajan, M. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9680–9689, 2020.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32:10967–10978, 2019.

Yoon, J., Jordon, J., and Schaar, M. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pp. 5689–5698. PMLR, 2018.

A. Additional Theory

A.1. Formulation of the MI Bounds for the Binary Case

As we discussed in our main paper, the relationship between Mutual Information (MI) and accuracy is not a function, but comes in form of upper and lower bounds of the obtainable accuracy. If, for example, the binary classification case with equal class priors $p(C=0) = p(C=1) = \frac{1}{2}$ is considered, the following bounds can be derived (Hellman & Raviv, 1970; Meyen, 2016):

$$\frac{I(\boldsymbol{x}; C) + 1}{2} \leq \mathrm{Acc}(C|\boldsymbol{x}) \leq H_2^{-1}(1 - I(\boldsymbol{x}; C)), \tag{5}$$

where $H_2^{-1} : [0,1] \to \left[\frac{1}{2}, 1\right]$ is the inverse of the binary entropy with support $\left[\frac{1}{2}, 1\right]$. For completeness, we restate the proof of this upper bound in Appendix A.2.

A.2. Reproduction of the proof of the relation between mutual and accuracy in the binary case

In this section, we reproduce the proofs for the upper and lower bounds of bayesian classifier accuracy given a certain amount of mutual information from the master's thesis by (Meyen, 2016) for completeness. The upper bound given there is tighter than the bounds present in the literature.
We consider the following setting (C, $\boldsymbol{x}$ are random variables):

- binary classification problem, $C \in \Omega_C = \{0, 1\}$
- equal class priors $P(C=0) = \frac{1}{2}, P(C=1) = \frac{1}{2}$
- discrete features $\boldsymbol{x}$ (which can be the product of multiple random variables)
- support set $\Omega_x = \mathrm{supp}\,\{\boldsymbol{x}\}$ of *countable* size

We first prove the following Lemma:

Lemma A.1. *Let the assumptions stated above be true. Then, the mutual information is the weighted mean of a function of the conditional accuracies* $\mathrm{Acc}(C|s)$*, where* $s \in \Omega_x$*:*

$$I(C; \boldsymbol{x}) = \sum_{s \in \Omega_S} p(s) \left(1 - H_2\left[\mathrm{Acc}(C|s)\right]\right)$$

In this formulation, $p(s)$ is a shorthand for $P(\boldsymbol{x} = s)$ and $H_2(p) := -p \log p - (1-p)\log(1-p)$ is the entropy for a binary random variable.
Proof.

$$\begin{aligned}
I(C; \boldsymbol{x}) &= H(C) - H(C|\boldsymbol{x}) && (6)\\
&= \sum_{c \in \Omega_C} p(c) \log \frac{1}{p(c)} - \sum_{s \in \Omega_S} p(s) \sum_{c \in \Omega_C} p(c|s) \log \frac{1}{p(c|s)} && (7)\\
&= \sum_{s \in \Omega_x} p(s) \left[\sum_{c \in \Omega_C} p(c) \log \frac{1}{p(c)} - \sum_{c \in \Omega_C} p(c|s) \log \frac{1}{p(c|s)} \right] && (8)\\
&= \sum_{s \in \Omega_x} p(s) \left[H(C) - H(C|s)\right] && (9)
\end{aligned}$$

In our consideration, $\Omega_C = \{0, 1\}$ and $P(C=0) = \frac{1}{2}, P(C=1) = \frac{1}{2}$, so $H(C) = 1$. Additionally, the bayesian classifier rule yields

$$acc(C|s) = \begin{cases} P(C=0|s), & \text{for } P(C=1|s) \leq 0.5 \\ P(C=1|s), & \text{for } P(C=1|s) > 0.5 \end{cases} \tag{10}$$

and

$$H(C|s) = -P(C=0|s)\log P(C=0|s) - P(C=1|s)\log P(C=1|s) \tag{11}$$
$$= H_2(P(C=0|s)) = H_2(P(C=1|s)) \tag{12}$$
$$= H_2(acc(C|s)) \tag{13}$$

Plugging in the results $H(C) = 1$ and $H(C|s) = H_2(\mathrm{Acc}(C|s))$, we obtain the proposed lemma. □

For the derivation of upper and lower bounds, Jenssen's inequality is used. $1 - H_2(\cdot)$ is a convex function and the $\{p(s)\}_{s\in\Omega_x}$ are convex multipliers, i.e., they are non-negative and sum up to one. Then,

$$1 - H_2\left(\mathrm{Acc}(C|\boldsymbol{x})\right) = 1 - H_2\left(\sum_{s\in\Omega_x} p(s)\,\mathrm{Acc}(C|s)\right) \tag{14}$$
$$\leq \sum_{s\in\Omega_x} p(s)\left[1 - H_2\left(\mathrm{Acc}(C|s)\right)\right] = I(\boldsymbol{x};C) \tag{15}$$

We can restate this equation in terms of accuracy.

$$H_2\left(\mathrm{Acc}(C|\boldsymbol{x})\right) \geq 1 - I(C;\boldsymbol{x}) \tag{16}$$

Using that $H_2(\cdot)$ is decreasing monotonically on the interval $\left[\frac{1}{2}, 1\right]$, so its inverse H_2^{-1} exists, and that $\mathrm{Acc}(C|s) \geq 0.5$:

$$\mathrm{Acc}(C|\boldsymbol{x}) \leq H_2^{-1}\left(1 - I(C;\boldsymbol{x})\right). \tag{17}$$

The inequality sign is flipped again, due to the inverse being monotonically decreasing. Note that the bounds derived for the special case are much tighter than the general ones provided by Vergara & Estévez (2014) and Cover & Thomas (2006, Chapter 2.10), that are not of any use, because they are even less strict than the trivial bound $\mathrm{Acc}(C|\boldsymbol{x}) \leq 1$, for the simple case considered here.

For the lower bound, we refer the reader to Hellman & Raviv (1970, eqn. 18), where the term I corresponds to $H(C|\boldsymbol{x}) = H(C) - I(C;\boldsymbol{x})$ in our notation. Rewriting the result from Hellman & Raviv (1970) in our notation, we obtain

$$1 - \mathrm{Acc}(C|\boldsymbol{x}) \leq \frac{H(C) - I(C;\boldsymbol{x})}{2}. \tag{18}$$

Using $H(C) = 1$ and rearranging yields

$$1 - \mathrm{Acc}(C|\boldsymbol{x}) \leq \frac{1 - I(C;\boldsymbol{x})}{2} \tag{19}$$

and

$$\mathrm{Acc}(C|\boldsymbol{x}) \geq \frac{I(C;\boldsymbol{x}) + 1}{2}. \tag{20}$$

□

A.3. Analysis of the LeRF Ordering

In this section, we analyze the masking impact for the case of the Least Relevant First (LeRF) ordering. We first provide a definition for the operators involved as we did for the Most Relevant First (MoRF) case. In the LeRF setting, the k least important important features per instance are removed. We model the explanation as a choice of features via a binary mask $\boldsymbol{M} = \boldsymbol{e}(f, \boldsymbol{x}) \in \{0,1\}^d$, with the corresponding value set to one, if the corresponding feature is among the top-k, and to zero otherwise. Furthermore, suppose $\mathcal{M}_h : \{0,1\}^d \times \mathbb{R}^d \to \mathbb{R}^k$ to be the selection operator for the **h**ighly important dimensions indicated in the mask and $\boldsymbol{x}_h = \mathcal{M}_h(\boldsymbol{M}, \boldsymbol{x})$ to be a vector containing only the remaining, highly important features as shown in Figure 9. We suppose that the features preserve their internal order in $\boldsymbol{x}_h$, i.e., features are ordered ascendingly by their original input indices.

The LeRF approach with retraining (also called "Keep and Retrain", KAR, by Hooker et al. (2019)) measures the accuracy of a newly trained classifier f' on modified samples $\boldsymbol{x}'_h := \mathcal{I}_h(\boldsymbol{M}, \boldsymbol{x_h})$, where $\mathcal{I}_h : \{0,1\}^d \times \mathbb{R}^k \to \mathbb{R}^d$ is an imputation

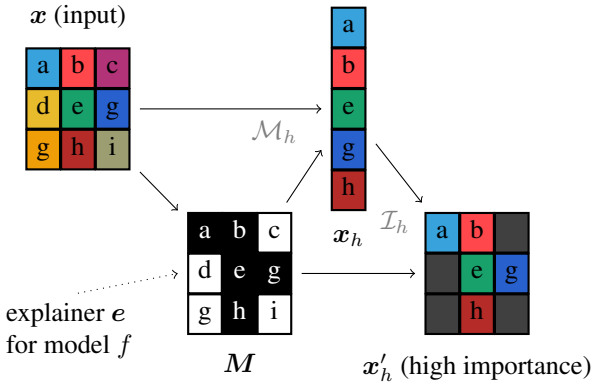


Figure 9. Analogous analytical model of feature removal in the opposite order (LeRF): The input image $\boldsymbol{x}$ is explained by an explanation method that returns a mask $\boldsymbol{M}$ indicating important pixels. The remaining, highly important pixels can be extracted from the image using the masking operator $\mathcal{M}_h$ and transformed to a modified variant of the input $\boldsymbol{x}'_h$ via the imputation operator $\mathcal{I}_h$.

operator that redistributes all inputs in the vector $\boldsymbol{x}_h$ to their original positions and sets the remainder to some filling value. This means only the top-k features are kept. For a better evaluation result, the accuracy should increase quickly with increasing k, indicating the most influential features are present. Accuracy should not increase much for the high values of k, because inserting the low importance features should not have a large effect (equivalently, this means it should not drop much when the least important features are removed). Overall, higher accuracies indicate better attributions in the LeRF setting.

For the LeRF benchmark, the quantity of interest in our analysis will be $I(\boldsymbol{x}'_h; C)$, the class information contained in the filled-in version of the selected high important features. We want to maximize $I(\boldsymbol{x}'_h; C)$ to obtain a good score,

$$\uparrow I(\boldsymbol{x}'_h; C) \;\; \Rightarrow \uparrow \text{LeRF benchmark.}$$

As before, we can apply the following, general identity:

$$\underbrace{I(\boldsymbol{x}'_h; C)}_{\text{Evaluation Outcome}} = \underbrace{I(C; \boldsymbol{x}'_h|\boldsymbol{M})}_{\text{Feature Info.}} + \underbrace{I(C; \boldsymbol{M})}_{\text{Mask Info.}} - \underbrace{I(C; \boldsymbol{M}|\boldsymbol{x}'_h)}_{\text{Mitigator}}. \tag{21}$$

The interpretation of the terms is analogous to that in our main paper.

Class-Leaking Explanation Map For the case of the class-leaking map, we again require the imputation operator to be invertible:

Example A.2. *Invertible Imputation. Let $\mathcal{I}_h : \{0,1\}^d \times \mathbb{R}^k \to \mathbb{R}^d$ be the imputation operator that takes the highly important features as an input. We suppose that there are inverse functions $\mathcal{I}^{-1}_{h,M}$ and $\mathcal{I}^{-1}_{h,x}$, such that*

$$\boldsymbol{x}'_h = \mathcal{I}_h\left(\boldsymbol{M}, \boldsymbol{x}_h\right) \Leftrightarrow \boldsymbol{M} = \mathcal{I}^{-1}_{h,M}(\boldsymbol{x}'_h) \wedge \boldsymbol{x}_h = \mathcal{I}^{-1}_{h,x}(\boldsymbol{x}'_h).$$

If, for instance, the pixels removed are set to some reserved value indicating their absence, the infilling operator is invertible. In this case, also the Mitigator $I\left(C; \boldsymbol{M}|\boldsymbol{x}'_h\right) = 0$ (see Section 4.3 for details). The "Feature Info" term is constrained to be positive. Thus, the Mask Information has a non-negligible impact on the Evaluation Goal, because a higher Mask term will always increase it.

We can create a another example of a spurious explanation map that shows how evaluation scores are influenced even worse for LeRF: Suppose an explanation map that starts masking out pixels at the top for class zero and at the bottom for class one. Thus, a retrained model will be able to infer the category just from the shape of the masked pixels and obtain the best possible accuracy and thus score in the LeRF setting. However, it does not provide a reasonable attribution for the importance of the features.

B. GAN Imputation

We also use Generative Adversarial Imputation Nets (GAIN) proposed by Yoon et al. (2018) as an imputation operator. We first train a GAIN model on CIFAR-10. To find the best-performing setup, we run a hyperparameter selection for the GAIN model. We keep all the default parameters identified by Kachuee et al. (2020), but search for the value of `alpha` (α), which can be seen as a weight factor for the reconstruction loss of the non-imputed pixels in the GAN, and the `hint_rate` (hr) parameter, which provides the Discriminator with hints to balance the difficulty of the tasks. We train the models for 100 epochs which resulted in converged MSEs and Frechet Inception Distances (FIDs). We use MSE to the original pixels to assess the generative quality of the model. Kachuee et al. (2020) reported low values for both these parameters to perform well, but did not provide the exact values. We extended their value ranges to $\alpha = 100$ and performed and exhaustive search. The results for the GAIN models on CIFAR-10 can be seen in Table 5. For the experiments we used the best setup with $\alpha = 100$ and $hr = 0.01$.

	α=0.1	α=1	α=10	α=100
hr=0.01	0.0131	0.0164	0.0090	**0.0085**
hr=0.1	0.0113	0.0133	0.0131	0.0101
hr=0.3	0.0172	0.0183	0.0151	0.0127
hr=0.9	0.0303	0.0484	0.0379	0.0088

Table 5. Mean-Squared-Errors for GAIN on CIFAR-10 using different hyperparameter choices.

In Figure 10, we demonstrate imputation results using three operators for one image (a) from CIFAR-10. Compared to the fixed value imputation (b) and noisy linear imputation (c), GAN imputation (d) yields most natural imputed image. Although it cannot perfectly reconstruct the original image, for example the background is noisy and the body color is different from the original one, it is not easy to deduce the mask from (d). A trained imputation predictor also verifies that GAN imputation is closest to the optimal condition, Minimally Revealing Imputation.

However, there are drawbacks of the GAN imputation. It may introduce some new "features" that do not exist in the original sample. For instance the dog in (d) has new patterns on its body. Moreover, it does not give very good results when too many pixels are removed (cf. Figure 12). The GAIN training again requires tuning hyperparameter settings and is highly expensive. Therefore, this model does not allow for the desired improvements (few hyperparameters, efficiency). Compared to GAN, our Noisy Linear imputation does not have these drawbacks. Considering all these factors, we recommend to use Noisy Linear Imputation in the evaluation framework.

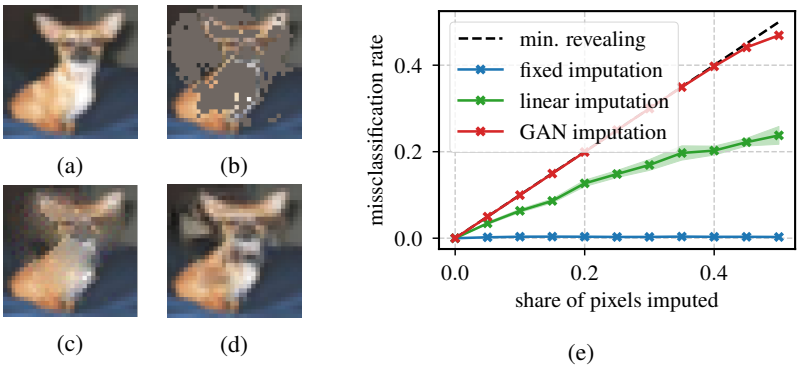


Figure 10. The considered imputation operators. When 30 % of the original image (a) are removed, they can either be completed by a fixed value (b) or by our proposed Noisy Linear imputation (c) or GAN imputation (d). Training of an imputation predictor (e) shows that it is much harder to tell which pixels are original and which were imputed when using our proposed imputation models, which is closer to the theoretical optimum (black). Hence, Class Information Leakage is reduced by our imputation methods.

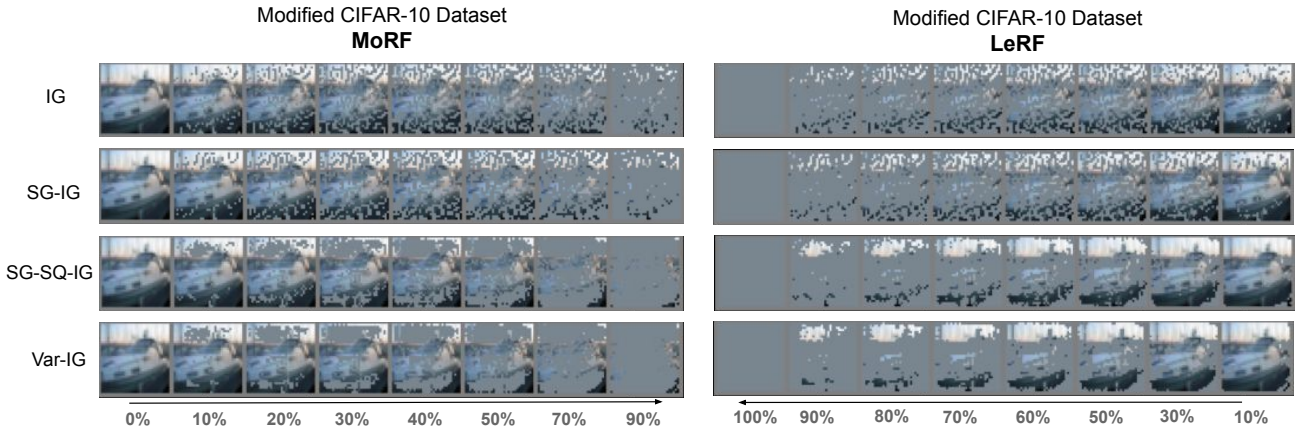


Figure 11. Illustration of modified data set in MoRF/LeRF and fixed value imputation settings. **Left**: Modifications in the MoRF framework. **Right**: Modifications in the LeRF framework. **Top to Bottom**: Modifications using Integrated Gradient (IG) (Sundararajan et al., 2017) and three ensemble variants of IG: SmoothGrad (SG-IG) (Smilkov et al., 2017), SmoothGrad^2 (SG-SQ-IG) (Hooker et al., 2019), and VarGrad (Var-IG) (Adebayo et al., 2018). The percentage of pixels that are removed or kept is given at the bottom.

C. Additional Experiments on CIFAR-10

C.1. Implementation Details

In this section, we report implementation details on CIFAR-10 as well as additional results for comparison between fixed value imputation and our *Noisy Linear Imputation*. We also include GAN imputation results. In Figure 12, an overview of using three different imputations with different perturbation percentages are illustrated.

We train a vanilla ResNet-18 (He et al., 2016) on CIFAR-10 and compute different explanations using the trained model. The model is trained with the initial learning rate of 0.01 and the SGD optimizer (Sutskever et al., 2013). We decrease the learning rate by factor 0.1 after 25 and train the model for 40 epochs on one GPU. The trained model achieves a test set accuracy of $84.5\,\%$ (comparable to the model in (Tomsett et al., 2020)). For attributions, we use the same settings as in (Hooker et al., 2019): As base explanations we implement Integrated Gradient (IG) (Sundararajan et al., 2017) and Guided Backprop (GB) (Springenberg et al., 2015). Additionally, we use three ensembling strategies for each: SmoothGrad (SG) (Smilkov et al., 2017), SmoothGrad^2 (SG-SQ) (Hooker et al., 2019) and VarGrad (Var) (Adebayo et al., 2018). For each explanation method, we modify the data set using the fraction of pixels $\eta = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9]$. Figure 11 illustrates the modified images by using four different explanations in the GB-family within MoRF and LeRF orders (fixed mean value imputation is used).

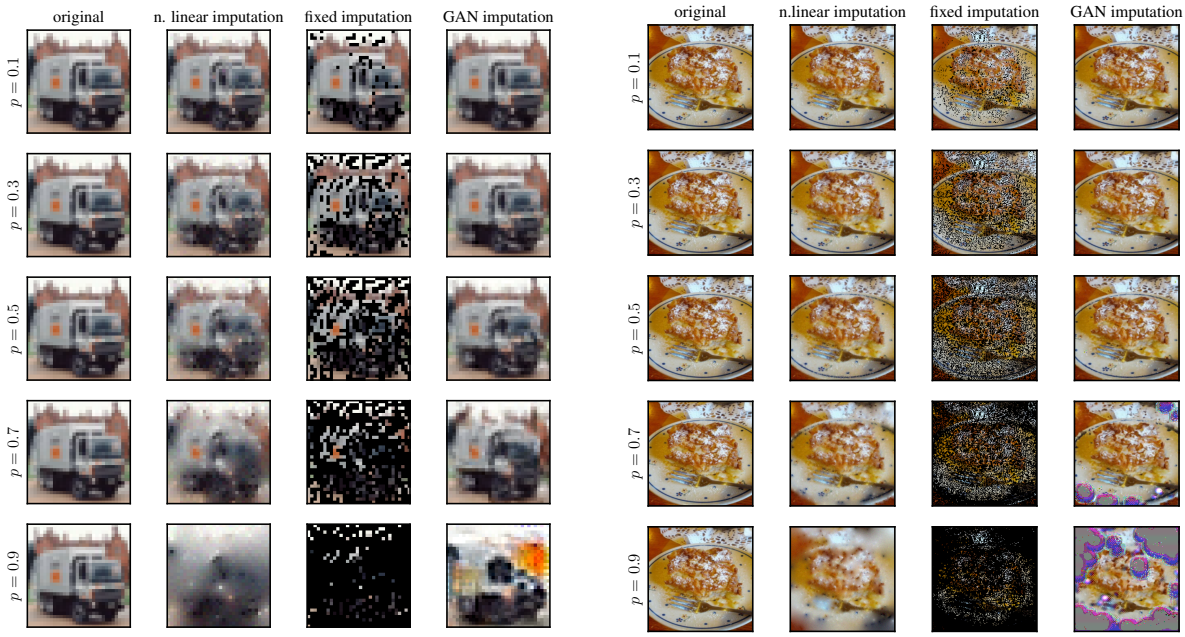
We use $N = 5$ **runs** and report averaged results for all CIFAR-10 experiments in our paper and indicate the standard errors (which are very small) as an area behind our plots. In Table 6 and Table 7, we show the mean accuracy and its standard deviation at each the fraction of pixels η for IG-SG and GB-SG explanations. For other explanations we used, the standard deviation at each η in the magnitude of below one percent as well. Mean runtimes (average over 5 runs) for evaluating one explanation method (IG) using all three imputation methods are listed in Table 8.

C.2. Correlation Analysis

In Table 9, we show a full view of the Spearman Correlation of rankings between all twelve different evaluation strategies ("Retrain"/"No-Retrain", MoRF/LeRF, and fixed value/Noisy Linear/GAN imputation) used in this paper. In this work, our primary focus was on consistency between the respective Retraining/No-Retraining Methods and the consistency between MoRF/LeRF and we mark the results used in the main paper in bold.

C.3. Extended Figures

In this section, we include full qualitative results of using four variants in evaluation strategies ("Retrain"/"No-Retrain", MoRF/LeRF) for three different imputation operators (fixed value/Noisy Linear/GAN imputation). In Figure 13, the full



(a) Sample from CIFAR-10

(b) Sample from Food-101

Figure 12. Sample images from CIFAR-10 and Food-101 imputed with the three methods considered in this work for different percentages. The missing pixels are determined by the IG attribution method (in MoRF order). While the GAN leads to sharper images for the early percentage values, where the linearly imputed samples become more blurry. Artefacts are introduced for high missingness percentages (0.9) in GAN imputation, which may distort the results of the evaluation once again. Therefore, we decide to stick to the Noisy Linear Imputation that operates more stably.

plots of IG-family attribution methods using fixed value imputation are shown, while Figure 16 illustrates for the GB-based attribution methods. Figure 14 and Figure 17 show the evaluation results when using our Noisy Linear Imputation for IG- and GB-family attribution methods, respectively. From results, we see that using our Noisy Linear Imputation, the consistency between the evaluation rankings conducted in MoRF and LeRF with and without retraining increases, for instance in Figure 14 compared to Figure 13.

D. Additional Experiments on Food-101

D.1. Implementation Details

We trained a vanilla ResNet-50 (He et al., 2016) on Food-101 (Bossard et al., 2014). Concretely, we trained the model using the SGD optimizer. Additionally the model was trained with the initial learning rate of 0.01. The learning rate was reduced by factor of 0.1 after every 10 epochs. In total, we trained 40 epochs with a batch size of 32 and the model achieved the accuracy of 81.67% on the test set. To run the GAN imputation operator, we first trained a GAIN model on Food-101 as introduced in Appendix B. We used the hyper-parameters $\alpha = 100$ and $hr = 0.1$ and trained the GAIN model with the batch size of 32 for 100 epochs. We computed the eight explanations and run ROAD and ROAR evaluation using the same settings as introduced in Appendix C.1 for CIFAR-10.

D.2. Correlation Analysis

In Table 10, we show a full view of Spearman Correlation of rankings given by eight different evaluation strategies ("Retrain"/"No-Retrain", MoRF/LeRF, and fixed/Noisy Linear/GAN imputation) on Food-101. In the table, results marked in bold indicate the consistency of using three imputation operators. We observe that the consistency between the respective Retrain and No-Retrain methods is still very high, which confirms that the efficiency gains reported in the main paper can be realized for larger data sets. Consistency between MoRf/LeRF is improved (over fixed imputation) when using retraining, but decreases slightly when the No-Retraining approach is used. Because the curves are often very close on this dataset

		10	20	30	40	50	70	90
Retrain MoRF	fixed	74.94±0.57	75.42±0.45	75.62±0.24	75.16±0.50	74.95±0.45	73.73±0.48	65.18±0.85
	lin	69.72±0.49	68.10±0.34	67.28±0.34	67.32±0.22	67.52±0.22	66.46±0.54	60.37±0.51
	gan	74.78±0.31	73.16±0.22	72.02±0.03	71.40±0.23	70.72±0.30	68.44±0.43	59.37±0.44
No-Retrain MoRF	fixed	44.06±0.04	29.81±0.03	21.99±0.03	17.35±0.02	14.67±0.01	11.50±0.04	10.90±0.03
	lin	67.66±0.02	59.94±0.03	54.05±0.05	49.46±0.04	45.63±0.06	36.87±0.05	24.55±0.04
	gan	74.53±0.04	71.41±0.04	69.10±0.06	67.55±0.09	66.55±0.07	60.73±0.12	25.46±0.10
Retrain LeRF	fixed	80.88±0.14	81.34±0.15	81.41±0.01	81.36±0.14	81.34±0.11	80.95±0.01	76.86±0.34
	lin	81.41±0.10	81.67±0.18	81.88±0.16	81.56±0.13	81.31±0.22	79.89±0.23	72.83±0.36
	gan	81.05±0.22	80.99±0.15	80.14±0.16	79.25±0.18	78.24±0.22	74.92±0.15	68.69±0.21
No-Retrain LeRF	fixed	74.34±0.02	69.04±0.03	64.06±0.04	59.86±0.03	57.59±0.03	53.81±0.06	46.74±0.02
	lin	82.20±0.04	82.04±0.03	81.76±0.08	81.34±0.06	80.97±0.03	77.89±0.07	56.74±0.13
	gan	80.80±0.02	80.38±0.03	79.90±0.02	78.85±0.07	77.47±0.08	71.14±0.10	32.96±0.17

Table 6. Mean accuracy at each η by using IG-SG in all methods with standard deviations of five individual runs. For LeRF, the accuracy is at (1-η).

		10	20	30	40	50	70	90
Retrain MoRF	fixed	76.30±0.43	75.60±0.27	74.89±0.29	74.27±0.29	73.37±0.28	72.15±0.09	67.99±0.24
	lin	72.83±0.37	71.87±0.41	71.58±0.19	70.98±0.15	70.47±0.20	67.81±0.45	59.38±0.46
	gan	76.64±0.13	75.44±0.13	74.73±0.28	73.69±0.30	72.85±0.34	68.97±0.08	56.81±0.30
No-Retrain MoRF	fix	73.03±0.03	66.72±0.03	58.72±0.07	52.51±0.04	48.52±0.08	48.79±0.06	44.43±0.06
	lin	74.57±0.08	71.18±0.06	68.70±0.08	67.24±0.08	64.82±0.11	57.68±0.06	32.59±0.09
	gan	76.57±0.03	74.70±0.04	72.51±0.09	71.19±0.07	69.64±0.08	60.89±0.15	21.11±0.16
Retrain LeRF	fixed	72.39±0.39	71.76±0.41	71.21±0.30	70.26±0.50	69.83±0.22	68.32±0.45	63.29±0.56
	lin	72.86±0.24	71.63±0.27	70.67±0.42	70.08±0.30	69.82±0.22	68.10±0.18	60.12±0.34
	gan	75.97±0.27	74.73±0.27	73.41±0.24	72.74±0.34	72.20±0.28	69.89±0.26	57.57±0.24
No-Retrain LeRF	fixed	69.61±0.04	64.90±0.02	57.88±0.05	51.67±0.09	46.93±0.06	42.40±0.09	37.10±0.03
	lin	71.84±0.06	66.71±0.08	63.79±0.05	61.46±0.09	59.69±0.09	55.09±0.06	35.72±0.13
	gan	75.13±0.02	72.13±0.05	70.25±0.05	68.56±0.08	67.35±0.08	62.32±0.13	24.61±0.19

Table 7. Mean accuracy at each η by using GB-SG in all methods with standard deviations of five individual runs. For LeRF, the accuracy is at (1-η).

(in particular for the No-Retraining setup), small differences might already lead to a change in the ranking and the results are in general noisier than on CIFAR-10. In summary, we observe similar trends, although the consistency gain between MoRF/LeRF in No-Retrain is not as pronounced. Nevertheless, a perfect agreement between MoRF/LeRF might not be desirable.

D.3. Extended Figures

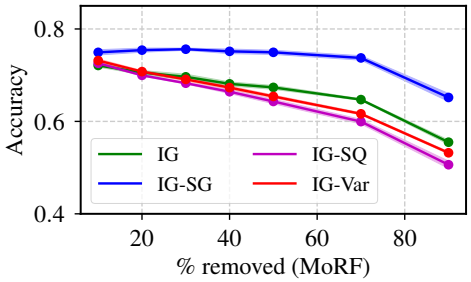
Full qualitative results of using four variants in evaluation strategies ("Retrain"/"No-Retrain", MoRF/LeRF) for three different imputation operators (fixed value/Noisy Linear/GAN imputation) are listed from Figure 19 to Figure 24. Figure 20 and Figure 23 show the evaluation results when using our Noisy Linear Imputation for IG- and GB-family attribution methods, respectively. From results, we see that using our Noisy Linear Imputation, the consistency between the evaluation results using "Retrain" and "No-Retrain" are more consistent compared to using the fixed value imputation. Therefore, retraining can be safely skipped by using our Noisy Linear Imputation.

Strategy	Retrain			No-Retrain		
	fixed†	lin	gan	fixed	lin*	gan
Time	3903±117 s	4686±2 s	6421±74 s	18.0±0.1 s	33.3±0.1 s	35.0±0.1 s
Relative	100 %	120 %	164 %	0.5 %	0.9 %	0.9 %

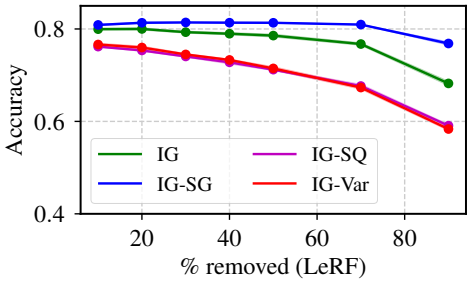
Table 8. Mean runtime (5 runs) for evaluating a single explanation method (IG) on three imputation operators. † refers to ROAR, and $\star$ to our ROAD.

		Retrain MoRF			No-Retrain MoRF			Retrain LeRF			No-Retrain LeRF		
		fixed†	lin	gan	fixed	lin*	gan	fixed	lin	gan	fixed	lin	gan
Retrain MoRF	fixed†	1.00 ±0.00											
	lin	0.68 ±0.02	1.00 ±0.00										
	gan	0.76 ±0.01	0.82 ±0.01	1.00 ±0.00									
No-Retrain MoRF	fixed	**0.15** ±0.01	0.38 ±0.02	0.23 ±0.01	1.00 ±0.00								
	lin*	0.66 ±0.01	**0.84** ±0.01	0.86 ±0.01	0.43 ±0.01	1.00 ±0.00							
	gan	0.65 ±0.01	0.62 ±0.01	0.84 ±0.01	0.14 ±0.01	0.78 ±0.01	1.00 ±0.00						
Retrain LeRF	fixed	**-0.01** ±0.01	0.48 ±0.02	0.28 ±0.02	0.66 ±0.00	0.47 ±0.02	0.13 ±0.01	1.00 ±0.00					
	lin	0.16 ±0.01	**0.61** ±0.01	0.34 ±0.01	0.78 ±0.01	0.50 ±0.01	0.10 ±0.01	0.87 ±0.01	1.00 ±0.01				
	gan	0.15 ±0.01	0.59 ±0.01	0.32 ±0.01	0.74 ±0.00	0.50 ±0.01	0.10 ±0.01	0.90 ±0.01	0.96 ±0.01	1.00 ±0.00			
No-Retrain LeRF	fixed	0.49 ±0.01	0.44 ±0.01	0.69 ±0.01	**0.01** ±0.00	0.60 ±0.00	0.77 ±0.00	**0.09** ±0.01	0.03 ±0.01	-0.03 ±0.00	1.00 ±0.00		
	lin	0.21 ±0.01	0.60 ±0.01	0.38 ±0.01	0.81 ±0.00	**0.58** ±0.01	0.22 ±0.01	0.85 ±0.00	**0.94** ±0.01	0.91 ±0.00	0.10 ±0.00	1.00 ±0.00	
	gan	0.05 ±0.01	0.47 ±0.01	0.17 ±0.01	0.69 ±0.00	0.36 ±0.00	-0.07 ±0.01	0.85 ±0.00	0.86 ±0.01	0.90 ±0.01	-0.14 ±0.00	0.79 ±0.00	1.00 ±0.00

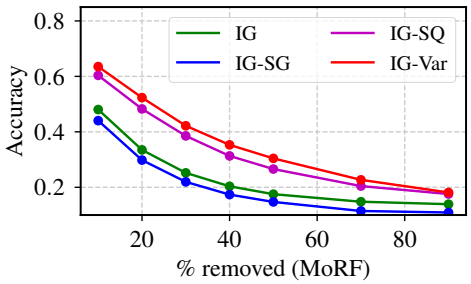
Table 9. **CIFAR-10**: Rank Correlations between all evaluation strategies used with standard deviations computed by considering the rankings obtained through five consecutive runs as independent. Results indicated in bold correspond to those reported in the main paper. The ROAR benchmark is marked by † and our ROAD by *.



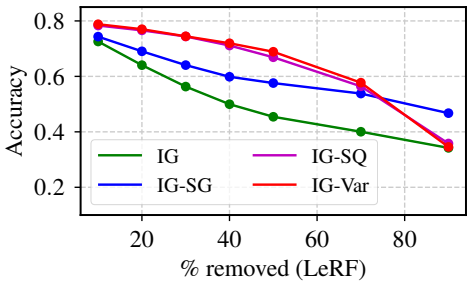
(a) MoRF, Retrain



(b) LeRF, Retrain

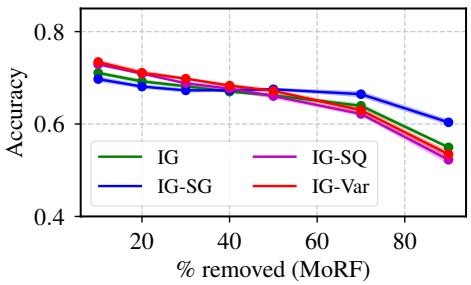


(c) MoRF, No-Retrain

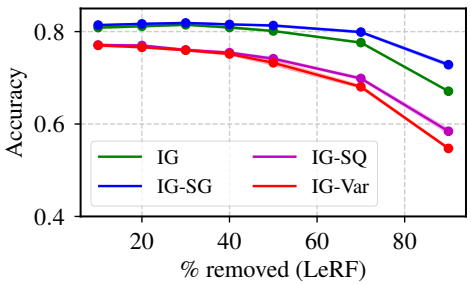


(d) LeRF, No-Retrain

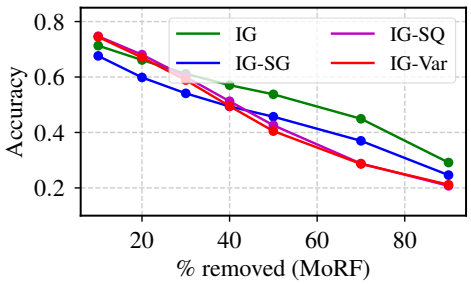
Figure 13. Consistency comparison using **Fixed Value** imputation on **IG**-based methods on CIFAR-10



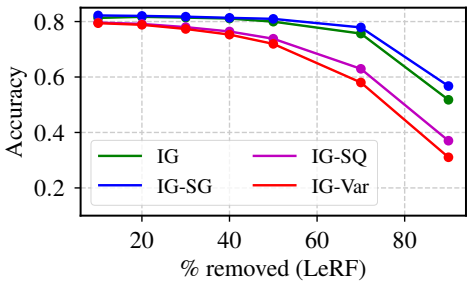
(a) MoRF, Retrain



(b) LeRF, Retrain

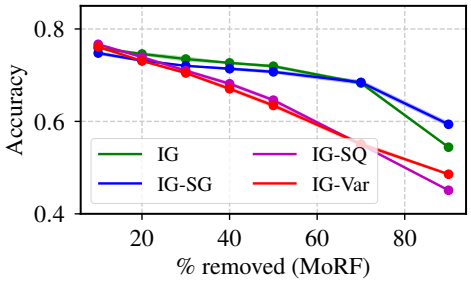


(c) MoRF, No-Retrain

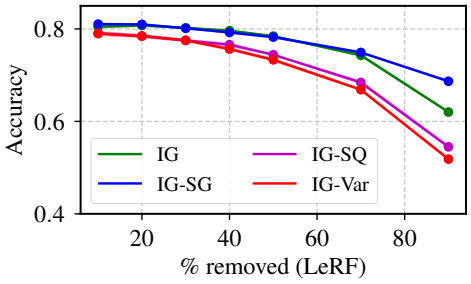


(d) LeRF, No-Retrain

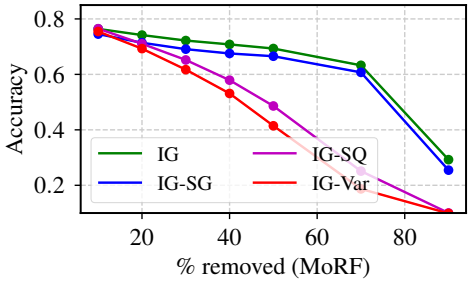
Figure 14. Consistency comparison using **Noisy Linear** imputation on **IG**-based methods on CIFAR-10



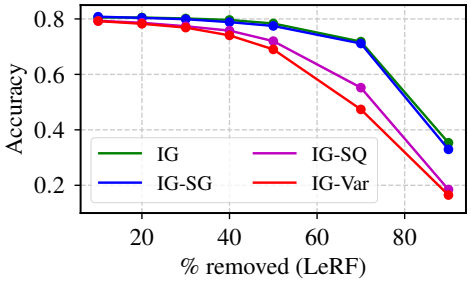
(a) MoRF, Retrain



(b) LeRF, Retrain

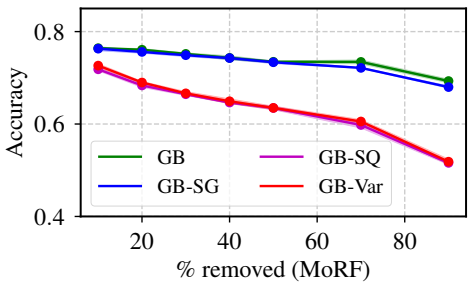


(c) MoRF, No-Retrain

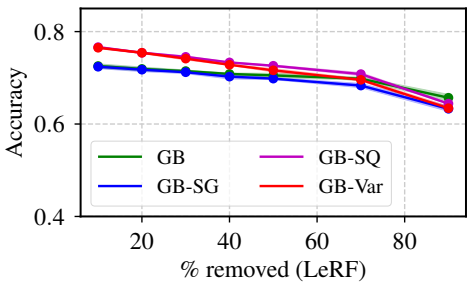


(d) LeRF, No-Retrain

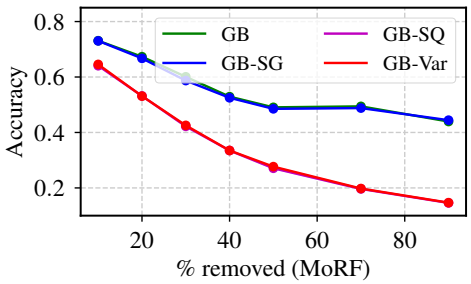
Figure 15. Consistency comparison using **GAN** imputation on **IG**-based methods on CIFAR-10



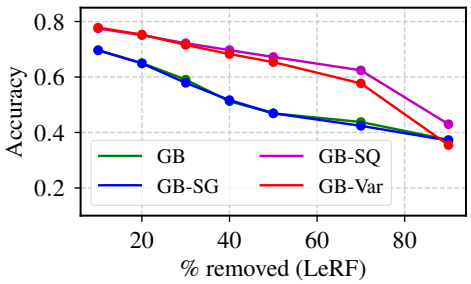
(a) MoRF, Retrain



(b) LeRF, Retrain

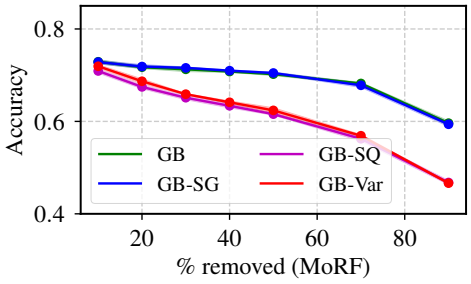


(c) MoRF, No-Retrain

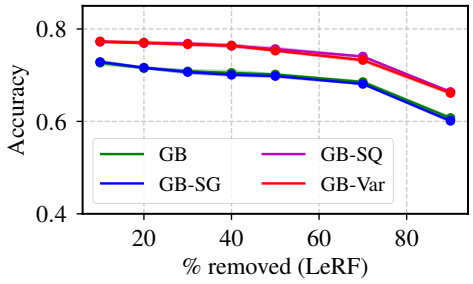


(d) LeRF, No-Retrain

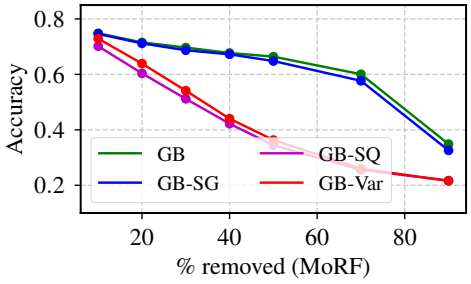
Figure 16. Consistency comparison using **Fixed Value** imputation on **GB**-based methods on CIFAR-10



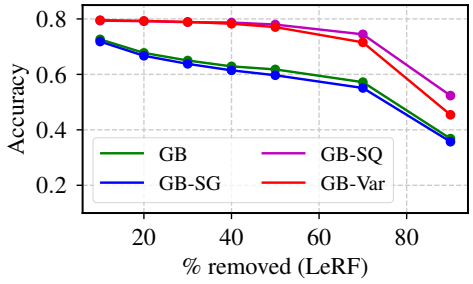
(a) MoRF, Retrain



(b) LeRF, Retrain

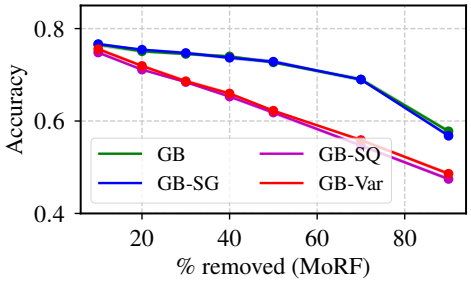


(c) MoRF, No-Retrain

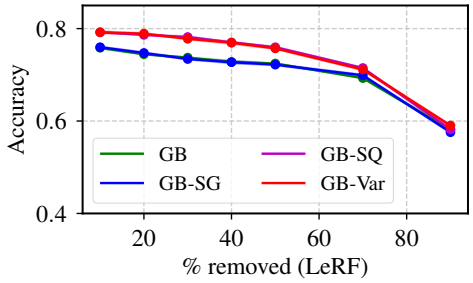


(d) LeRF, No-Retrain

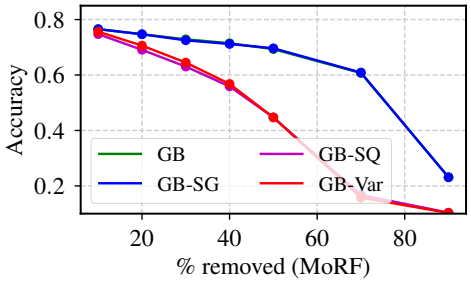
Figure 17. Consistency comparison using **Noisy Linear** imputation on **GB**-based methods on CIFAR-10



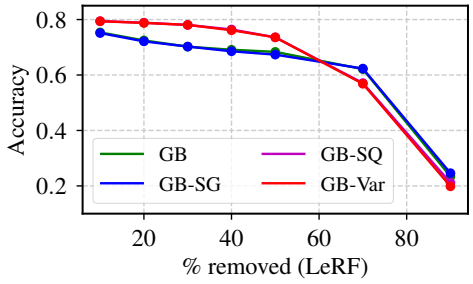
(a) MoRF, Retrain



(b) LeRF, Retrain



(c) MoRF, No-Retrain

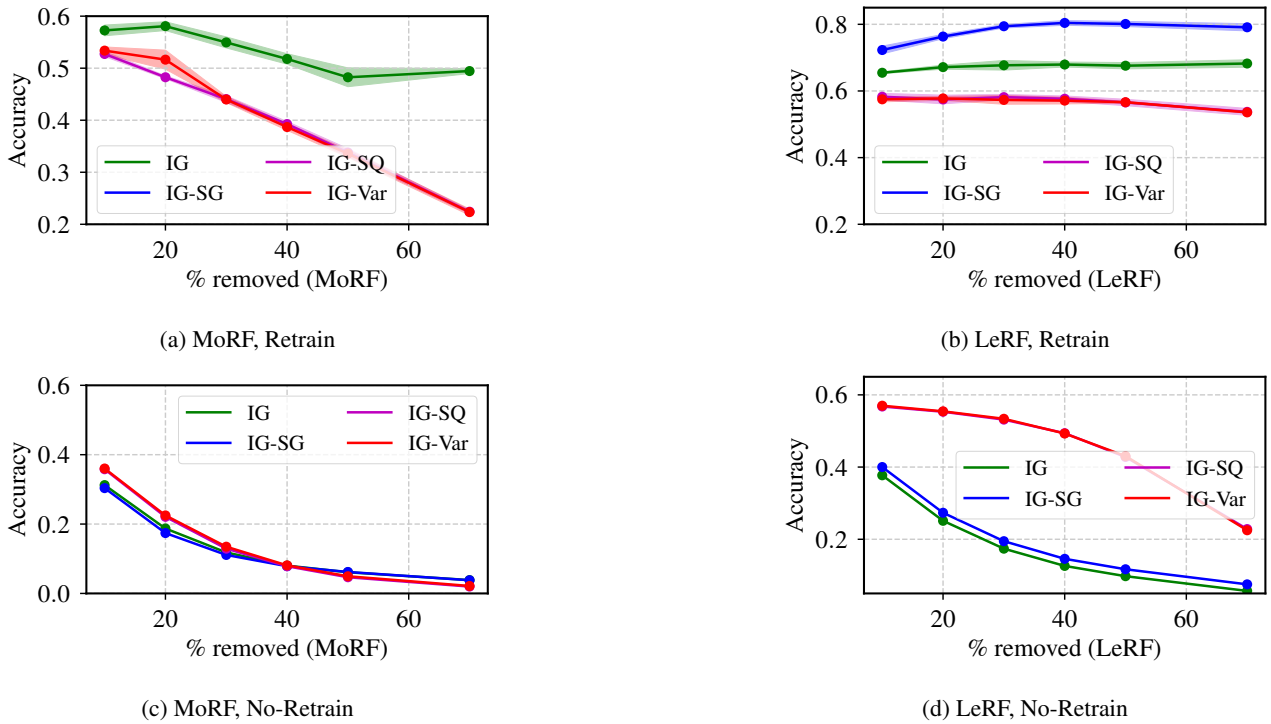


(d) LeRF, No-Retrain

Figure 18. Consistency comparison using **GAN** imputation on **GB**-based methods on CIFAR-10

		Retrain MoRF			No-Retrain MoRF			Retrain LeRF			No-Retrain LeRF		
		fixed†	lin	gan	fixed	lin*	gan	fixed	lin	gan	fixed	lin	gan
Retrain MoRF	fixed†	1.00 ±0.00											
	lin	0.48 ±0.03	1.00 ±0.00										
	gan	0.50 ±0.04	0.79 ±0.03	1.00 ±0.00									
No-Retrain MoRF	fixed	**0.12** ±0.01	0.57 ±0.02	0.50 ±0.01	1.00 ±0.00								
	lin*	0.61 ±0.01	**0.81** ±0.02	0.67 ±0.04	0.31 ±0.01	1.00 ±0.00							
	gan	0.74 ±0.01	0.79 ±0.02	**0.67** ±0.04	0.35 ±0.01	0.86 ±0.00	1.00 ±0.00						
Retrain LeRF	fixed	**-0.26** ±0.02	0.41 ±0.02	0.30 ±0.02	0.53 ±0.01	0.10 ±0.01	0.11 ±0.01	1.00 ±0.00					
	lin	-0.40 ±0.02	**0.26** ±0.04	0.19 ±0.04	0.30 ±0.03	-0.05 ±0.01	0.09 ±0.01	0.83 ±0.01	1.00 ±0.00				
	gan	-0.18 ±0.01	0.46 ±0.04	**0.32** ±0.04	0.50 ±0.03	0.13 ±0.02	0.14 ±0.03	0.89 ±0.02	0.83 ±0.01	1.00 ±0.00			
No-Retrain LeRF	fixed	0.79 ±0.02	0.79 ±0.03	0.63 ±0.05	**0.32** ±0.01	0.85 ±0.00	0.89 ±0.00	**0.02** ±0.01	-0.15 ±0.02	0.10 ±0.03	1.00 ±0.00		
	lin	-0.28 ±0.02	0.35 ±0.02	0.28 ±0.04	0.46 ±0.00	**-0.03** ±0.00	-0.06 ±0.00	0.89 ±0.01	**0.81** ±0.02	0.87 ±0.01	-0.11 ±0.00	1.00 ±0.00	
	gan	-0.45 ±0.02	-0.08 ±0.03	-0.04 ±0.04	0.23 ±0.00	-0.37 ±0.00	**-0.44** ±0.00	0.58 ±0.01	0.61 ±0.01	**0.54** ±0.00	-0.41 ±0.00	0.70 ±0.00	1.00 ±0.00

Table 10. **Food-10**: Rank Correlations between all evaluation strategies used with standard deviations computed by considering the rankings obtained through five consecutive runs as independent. The ROAR benchmark is marked by † and our ROAD by *. Bold results highlight the consistency between Retrain and No-Retrain (still very high) as well as MoRF and LeRF evaluation strategies using different imputation operators (fair increase when using Noisy Linear and GAN imputations instead of fixed imputation in "Retrain", decrease in "No-Retrain").



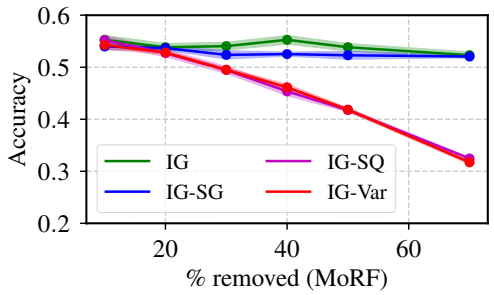
(a) MoRF, Retrain

(b) LeRF, Retrain

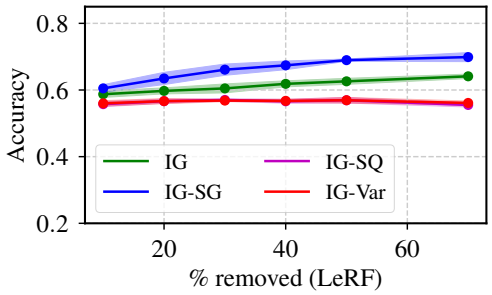
(c) MoRF, No-Retrain

(d) LeRF, No-Retrain

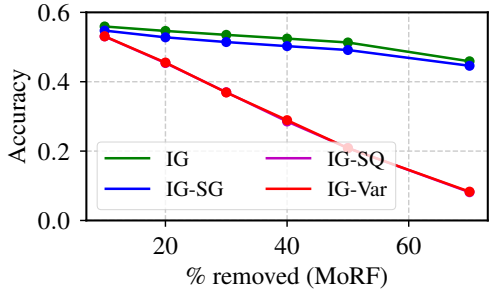
Figure 19. Consistency comparison using **Fixed Value** imputation on **IG**-based methods on Food-101.



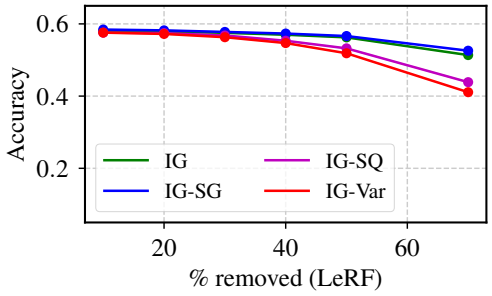
(a) MoRF, Retrain



(b) LeRF, Retrain

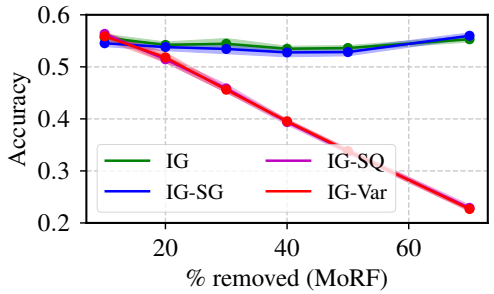


(c) MoRF, No-Retrain

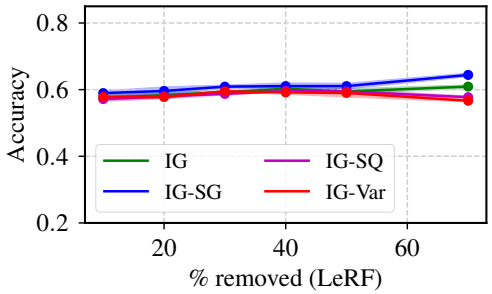


(d) LeRF, No-Retrain

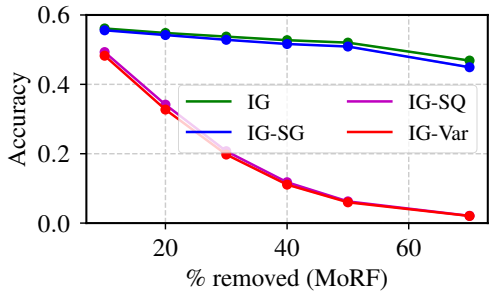
Figure 20. Consistency comparison using **Noisy Linear** imputation on **IG**-based methods on Food-101.



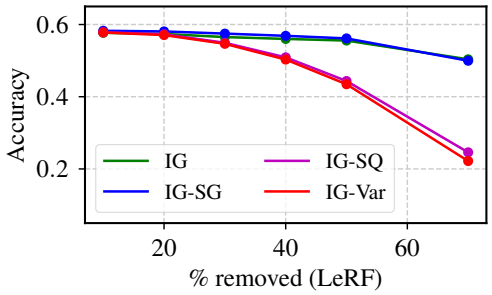
(a) MoRF, Retrain



(b) LeRF, Retrain

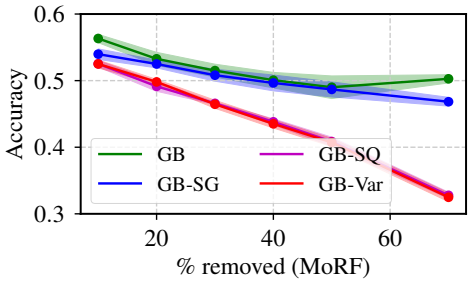


(c) MoRF, No-Retrain

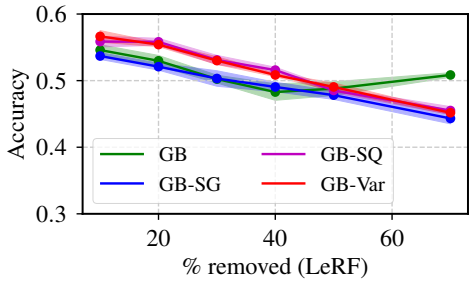


(d) LeRF, No-Retrain

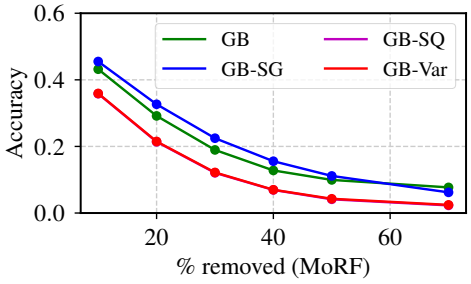
Figure 21. Consistency comparison using **GAN** imputation on **IG**-based methods on Food-101.



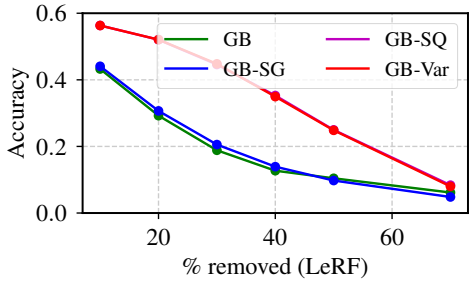
(a) MoRF, Retrain



(b) LeRF, Retrain

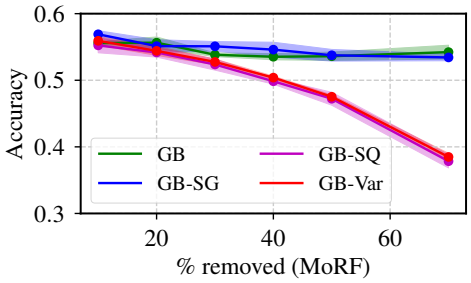


(c) MoRF, No-Retrain

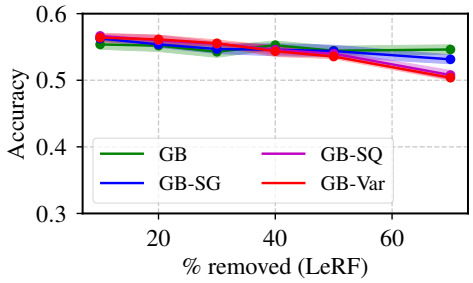


(d) LeRF, No-Retrain

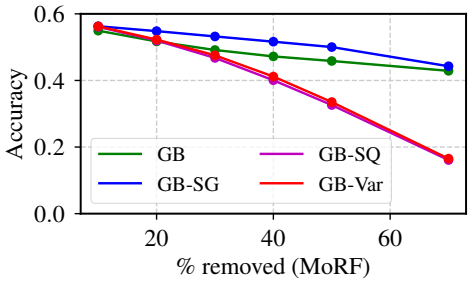
Figure 22. Consistency comparison using **Fixed Value** imputation on **GB**-based methods on Food-101.



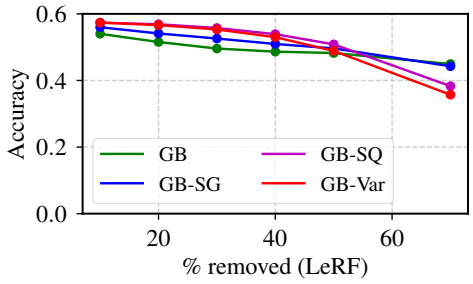
(a) MoRF, Retrain



(b) LeRF, Retrain

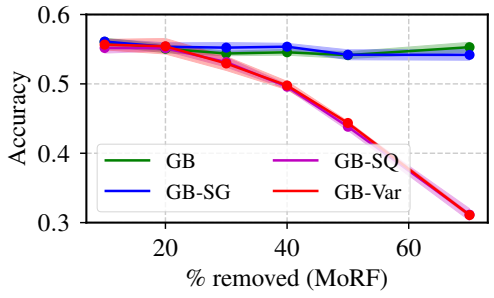


(c) MoRF, No-Retrain

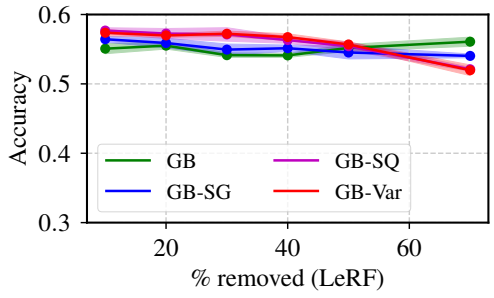


(d) LeRF, No-Retrain

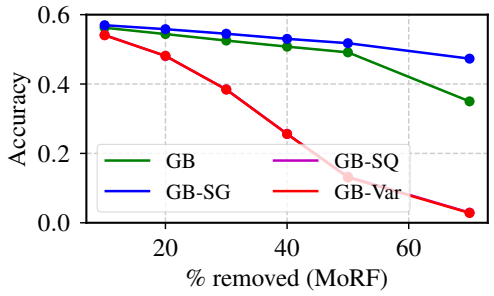
Figure 23. Consistency comparison using **Noisy Linear** imputation on **GB**-based methods on Food-101.



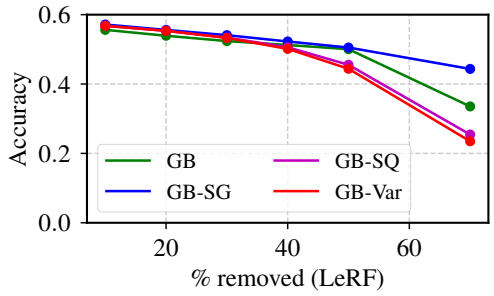
(a) MoRF, Retrain



(b) LeRF, Retrain



(c) MoRF, No-Retrain



(d) LeRF, No-Retrain

Figure 24. Consistency comparison using **GAN** imputation on **GB**-based methods on Food-101.

Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci

(Survey Paper)

***Abstract*—Explainable AI (XAI) is widely viewed as a sine qua non for ever-expanding AI research. A better understanding of the needs of XAI users, as well as human-centered evaluations of explainable models are both a necessity and a challenge. In this paper, we explore how human-computer interaction (HCI) and AI researchers conduct user studies in XAI applications based on a systematic literature review. After identifying and thoroughly analyzing 97 core papers with human-based XAI evaluations over the past five years, we categorize them along the measured characteristics of explanatory methods, namely *trust, understanding, usability*, and *human-AI collaboration performance*. Our research shows that XAI is spreading more rapidly in certain application domains, such as recommender systems than in others, but that user evaluations are still rather sparse and incorporate hardly any insights from cognitive or social sciences. Based on a comprehensive discussion of best practices, i.e., common models, design choices, and measures in user studies, we propose practical guidelines on designing and conducting user studies for XAI researchers and practitioners. Lastly, this survey also highlights several open research directions, particularly linking psychological science and human-centered XAI.**

***Index Terms*—Explainable AI (XAI), human-centered XAI, explainable ML, user study, human-AI interaction.**

I. Introduction

ARTIFICIAL Intelligence (AI) is driving digital transformation and is already an integral part of various everyday technologies. Recent developments in AI are essential to progress in fields such as recommendation systems [97], [98], [99], autonomous driving [100], [101], [102] or robotics [103], [104], [105]. Moreover, AI's success story has not excluded high-stakes decision-making tasks like medical diagnosis [106], [107], [108], credit scoring [109], [110], [111], jurisprudence [112], [113] or recruiting and hiring decisions [114], [115], However, the behavior and decision-making processes of modern AI systems are often not understandable, so they are frequently considered black boxes. Deploying such black-box models presents a serious dilemma in certain safety-critical domains, for instance, public health or finance [116]. This is due to the necessity for a transparent and trustworthy AI system, which is required by both practitioners (to gain better insights into system functioning) and end users (to rely on model decisions).

Methods to increase the interpretability and transparency of an AI system are developed in the research area of Explainable AI (XAI). Specifically, human-centered XAI, which addresses the importance of human stack-holders to the AI systems, has been proposed and discussed since [117], [118]. While a huge number of model explanations are available, the question of how to transparently evaluate their quality is still an open research question, and hence, extensively studied in recent years. A popular taxonomy of evaluation strategies for XAI methods proposes three categories: functionally-grounded evaluation, application-grounded evaluation, and human-grounded evaluation [119]. While functionally-grounded measures do not require human labor, the other two involve human subjects and are more costly to conduct.

Many functionally-grounded measures have been proposed to evaluate XAI algorithms (see [120] for review), however, the difficult comparability between different automatic evaluation measures is a common problem [121], [122]. Another drawback of automated measures is that there is no guarantee that they truly reflect humans' preferences [40], [123]. Consequently, user studies in XAI, especially when moving towards real-world products, are inevitable if one wishes to test more general beliefs of the quality of explanations [16]. However, only a small portion (about 20%) of XAI evaluation projects consider human subjects [120]. There exist efforts in developing taxonomies or introducing the definitions or implications of different human-centric evaluations [124], [125], [126], but the recent generation of user studies and their findings have not been systematically discussed yet. Moreover, Yang et al. [127] point out that XAI is growing separately and treated differently in different communities (e.g., machine learning and HCI). Hence, effective guidance in XAI user study design is crucial to better let both XAI algorithm

Manuscript received 3 February 2023; revised 26 October 2023; accepted 4 November 2023. Date of publication 13 November 2023; date of current version 6 March 2024. Recommended for acceptance by M. Cheng. *(Corresponding author: Yao Rong.)*

Yao Rong, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci are with the Technical University of Munich, 80335 Munich, Germany (e-mail: yao.rong@tum.de; tina.seidel@tum.de; gjergji.kasneci@tum.de; enkelejda.kasneci@tum.de).

Tobias Leemann, Thai-Trang Nguyen, and Lisa Fiedler are with the University of Tübingen, 72076 Tübingen, Germany (e-mail: tobias.leemann@uni-tuebingen.de; thai-trang.nguyen@student.uni-tuebingen.de; lisa.fiedler@student.uni-tuebingen.de).

Peizhu Qian and Vaibhav Unhelkar are with the Rice University, Houston, TX 77005 USA (e-mail: pq3@rice.edu; vaibhav.unhelkar@rice.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3331846, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3331846

TABLE I
OVERVIEW OF THE CORE PAPERS CONTAINING USER STUDIES IN XAI GROUPED BY CATEGORIES OF MEASUREMENTS AS SOME CORE PAPERS ASSESS QUANTITIES BELONGING TO SEVERAL GROUPS, A SINGLE PAPER CAN ALSO BE LISTED AMONG MULTIPLE GROUPS

Trust		[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]
Understanding	subjective	[7, 12, 13, 14, 16, 17, 22, 28, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]
	objective	[12, 13, 22, 32, 35, 39, 40, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]
	explanation model	[21, 46, 49, 61, 62, 63, 64, 65]
Usability	workload	[3, 16, 21, 48, 66]
	helpfulness	[13, 45, 46, 48, 56, 65, 67, 68]
	satisfaction	[1, 6, 7, 16, 18, 19, 29, 47, 69, 70]
	undesired behavior detection	[2, 24, 27, 38, 53, 57, 71, 72, 73, 74, 75, 76, 77, 78, 79]
	ease of use and others	[1, 3, 13, 20, 21, 24, 30, 32, 37, 48, 65, 66, 71, 80, 81, 82, 83, 84, 85, 86, 87]
Human-AI Collaboration Performance		[10, 13, 15, 25, 25, 29, 30, 39, 43, 53, 56, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97]

and application designers recognize the users' real needs. This work aims to bridge this research gap in modern XAI user study design by distilling practical guidelines for user studies through a comprehensive and structured literature review.

Therefore, we reviewed highly relevant papers that include user studies from top-tier HCI and XAI venues. Specifically, we included the recent *five* years of CHI, IUI, UIST, CSCW, FA(cc)T, ICML, ICRL, NeurIPS, and AAAI. As we aim at analyzing human user evaluation of advanced model explanations, we ran search queries involving keywords from the two groups "explainable AI" and "user study", as listed in the Table II. We selected the papers containing at least one keyword from each group, resulting in over one hundred papers. Then, we thoroughly studied these papers and filtered out papers that did not fulfill the criteria: (1) deploying explainable models or techniques and (2) conducting an assessment with human subjects. We identified a total of 97 core papers for this survey (see Table I for an overview of core papers with respect to their measured quantities in user studies). Based on these core papers, we performed a comprehensive analysis to fill the research gap by offering a systematic overview of user studies in XAI. We highlight the main contributions:

1) To offer an overview of the foundational work of user studies in XAI, we investigated references of all 97 core papers in a data-driven manner. Likewise, we analyzed follow-up works building on these core papers (identified through citations of core papers) to reveal the fields impacted by XAI user evaluations (Section III).
2) We present a summary of the design details in XAI user studies with particular focus on the deployed models and explanation techniques, experimental design patterns, participants as well as concrete measures, providing inspiration of how to collect human assessment (Section IV).
3) We discuss the impact of using explanations on different aspects of user experience (Section V), which can serve as an overview of the effectiveness of the current XAI technology and a summary of the state-of-the-art.
4) Based on the examined user study details and their best-practice findings, we synthesize guidelines for designing an effective user study for XAI (Section VI).
5) Beyond the user study design, we discuss potential paradigms of AI systems understanding humans in the context of e.g., theory of minds, as well as other future research directions (Section VII).

Our study highlights under-investigated areas in the context of current user-centered XAI research such as cognitive or psychological sciences through data-driven bibliometric analysis. Together with our proposed guidelines, we believe that this work will benefit XAI practitioners and researchers from various disciplines and will help to approach the overarching goal of human-centered XAI.

II. RELATED WORK

As a vast amount of explanation methods have been proposed, many researchers seek a systematic overview of the ever-growing field of XAI. In [128], [129], [130], [131], [132], [133], the authors aim to cover many facets of XAI technologies ranging from problem definitions, goals, AI/ML model explanations to evaluation measures, while in [134] the authors emphasize the research trends and challenges in Human-Computer-Interaction (HCI) applications. A large body of XAI surveys focuses mainly on the interpretability of a particular family of models and corresponding explanation techniques. For instance, [135], [136], [137] investigate explanations for Deep Neural Networks (DNNs), where models often take images as input [135], [136]. Joshi et al. [137], however, provide an extensive review for DNNs with multimodal input for instance that of joint vision-language tasks. Causal interpretable models are gaining more attention recently and Moraffah et al. [138] provide a literature review for causal explanations. A systematic literature review on explanations for advice-giving systems is conducted in [139]. Among these surveys focusing on general XAI technologies, evaluation measures are only briefly examined.

One challenge in XAI research is to evaluate and compare different explanation methods, due to the multidisciplinary concepts in interpretability/explainability [119], [120], [140]. Evaluation measures can be divided into two groups: human-grounded measures that rely on human subjects and functionally-grounded metrics that can be computed without human subjects [119], [120]. Many researchers seek solutions to evaluate explanations automatically. A comprehensive literature review with a focus on these functionally-grounded evaluation methods (without human subjects) can be found in [120]. Explainability is an inherently human-centric property, therefore, the research community should and has started to recognize the need for human-centered evaluations when working on XAI [119], [141].

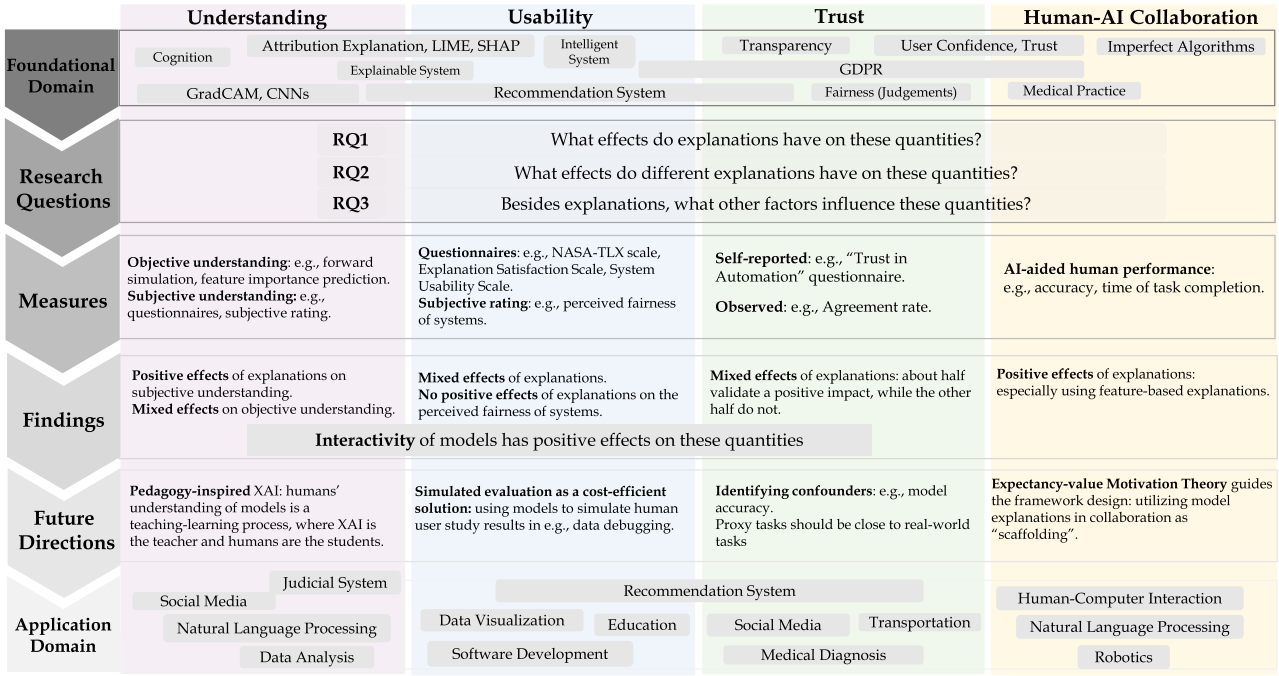


Fig. 1. Roadmap of our literature analysis. We find out the foundational works of core papers and their application domains using a data-driven method introduced in Section III. Three main research questions in user studies are distilled from core papers. Methods related to measures of each category are discussed in Section IV, and findings of the research questions are summarized in Section V. Based on the findings, we propose future directions to further promote human-centered XAI in Section VII. We distill important messages in this figure, but refer to the discussion in the corresponding sections for more details.

For instance, Chromik and Schuessler [125] propose a taxonomy on XAI evaluations involving humans. Mohseni et al. [126] summarize four groups of human-related evaluation metrics: mental model (e.g., user's understanding of the model), user trust, human-AI task performance and explanation usefulness and satisfaction (i.e., user experience). Hoffman [124] places more focus on psychometric evaluations by proposing a conceptual model of the XAI process and specifying four key components that should be evaluated: explanation goodness and satisfaction, (user's) mental models, curiosity, trust and performance. Beyond assessing evaluation methods, XAI applications are designed to eventually support decision-making and benefit end users. A recent review by Lai et al. [142] considers studies on collaborative Human-AI decision-making, which may include AI agents providing explanations. Success in human-AI decision-making tasks can be seen as one amongst many other ways to evaluate the effect of explanations. Ferreira and Monteiro [143] present a review of the user experience of XAI applications to answer who uses XAI, why, and in which context (what + when) the explanation is presented.

Closer to our focus on user studies concerning XAI, Liao et al. [141] study user experiences with XAI to reveal pitfalls of existing XAI methods, underscoring the important role of humans in XAI development. As suggested by Doshi-Velez and Kim [119], a human-subject experiment needs to be designed sophisticatedly to reduce confounding factors. In contrast to previous surveys on XAI, we aim to provide XAI researchers and practitioners with a comprehensive overview of the research questions explored in user studies, along with thorough information on experimental design. To this end, we present a practical guideline in user study design, which can be used as a starting point for future exploration of human-centric XAI applications.

III. Methodology

To analyze the collected papers related to user studies on XAI, we first categorize them into four groups based on their objectives. From these studies, we distill three main research questions concerning the effects of model explanations on each objective. We then summarize the methods used in these studies to quantify these objectives. Important findings from the papers are discussed, and we propose future directions based on these findings. Additionally, we examine the foundational works upon which these user studies are based (i.e., their references) and the follow-up papers that cite them, shedding light on the foundational works and emerging trends in human-centered XAI studies. Fig. 1 presents a roadmap of our analysis.

In this section, we first describe the criteria used for their categorization. We then discuss the foundational and application domains of these papers, providing a broader view before diving into their detailed analysis.

A. Categorization of User-Study Objectives

Since the core papers cover various factors of model explanations, we decided to categorize the core papers into different clusters to better study their commonalities and differences. In [119], *interpretability* in the context of ML systems is defined

as the ability to explain or present model predictions in understandable terms to a human. Beyond fostering comprehension, the authors argue that interpretability can assist in qualitatively ascertaining whether other desiderata, such as *usability* and *trust* are met. During a profound study of the relevant literature that was previously selected, we identified four sensible categories, that are derived from the considered dependent variables in user studies (desiderata of interpretability). These four categories are *trust, understanding, usability*, and *human-AI collaboration performance*. In Table I, the studied papers are categorized according to the measured quantities. As each measure can usually be assigned to only one of these categories, we found this distinction to be intuitive.

These categories reflect different functionalities (goals) of XAI. As interpretability is defined as "*the ability to explain or to present in understandable terms to a human.*", humans' "understanding" is the direct goal of XAI. To be concrete, understanding in the context of interacting with an ML model refers to a user's grasp or "mental model" of how the model operates, and this knowledge grows from using the system and from clear explanations about it [141]. "Usability" is commonly studied in human-computer interaction [144], which is one of the desiderata of XAI [119]. According to [145], usability is the extent to which users can utilize a product to successfully, efficiently, and satisfactorily accomplish their intended objectives. Thus, this category encompasses user studies that employ model explanations to support users in achieving specific tasks. In usability, different aspects are measured, for instance, whether the system is easy to use or how much cognitive load it requires. The aspect "undesired behavior detection" relates to use cases where explanations uncover model discriminatory behaviors, such as the utilization of undesired features. "Trust" in AI is summarized as a combination of the user's confidence in a model's accuracy, a personal comfort level with understanding and using it, and the willingness to let the model make decisions [140]. It encompasses more requirements. Human-AI collaboration performance is related to scenarios where the AI system provides its predictions, but humans retain the final decisions [89]. In this case, model explanations are deployed to reach a performance superior to that of the AI system or the human decision-maker alone. These categories cover different dependent variables of interest in the reviewed user studies, primarily related to how XAI methods function. These functions mainly tie to the models' reasoning and knowledge representation. A wider perspective on XAI, which assesses generalization or robustness, remains an important field for future exploration through user studies.

B. Foundations of User Studies

Based on a data-driven bibliometric analysis of the references in core papers, we highlight significant research topics within the "Foundational Domain" in Fig. 1. It is evident that model explanations and interpretability are pivotal components. This includes papers that introduce explanation methods such as LIME [146], SHAP [147], and other attribution methods. These are a frequent subject of study in works measuring understanding and usability. Additionally, convolutional networks, which are commonly employed in experiments, use tools like GradCAM [148] and various saliency maps to generate model explanations. Notably, many research papers appear within the domain of recommender systems, because many XAI user studies are conducted in the context of recommendation solutions. he EU's General Data Protection Regulation (GDPR) [149] is frequently mentioned in core papers due to the ongoing debate on the right to explanation" [150]. This debate has significantly influenced the shift in modern AI systems towards explainability. While the ultimate consumers of model explanations are humans, well-established research domains that focus on human understanding are underrepresented. For instance, only a few papers related to "Cognition" are cited compared to those on other algorithmic topics. Millecamp et al. [18] suggest enhancing XAI theory with insights from social sciences, including cognitive science and psychology. Given the scant references to psychology, it appears that only a handful of XAI user studies delve into evaluating XAI from a psychological standpoint. We highlight a nascent research domain of XAI frameworks based on human cognition and behavior theories [141]. This theoretical guidance can also offer conceptual tools for better evaluating XAI from user perspectives. More details about common references can be found in Appendix A.1, available online.

C. Impact of User Studies

Fig. 1 presents applications that make use (and thus are the consumers) of the findings from core papers. We noticed that studies on user understanding and trust span a wide range of applications. For example, trust is frequently addressed in the contexts of medical diagnosis and transportation, indicating its significance in high-risk scenarios. Recommendation systems emerge as a primary focus in follow-up works. Papers on usability have a significant impact on fields like data visualization, software development, and education. In these areas, models frequently serve as tools to ease the burden on end users. Human-AI collaboration measures particularly promote the further development of robotics and or natural language processing. The prominence of recommendation systems in both foundational works and their impact implies that XAI is an integral component of contemporary recommendation systems. A comprehensive overview of the fundamental works and application domains can be found in Appendix A.1, available online.

IV. Comprehensive User Study Analysis

In this section, we present details of the covered XAI user studies. We first introduce some commonly used AI models and explanation techniques (Section IV-A), followed by a discussion of application domains and measures with respect to the four measured quantities. The experimental designs, as well as analysis tools are presented in Section IV-C.

TABLE II
KEYWORDS FOR OUR PAPER SEARCH QUERY

	Explainable AI	User Study
Keywords	XAI, explainable AI, explanation, explainable, explanatory, interpretable, intelligible, black-box, machine learning, explainability, interpretability, intelligibility, explain attribution, feature	user study, participant , human subject, empirical study, lab study, user evaluation, human evaluation

Two groups of keywords were used.

TABLE III
MODELS AND EXPLANATIONS IN CORE PAPERS

		White-box	Black-box	Other
Feature-based	**local**	[21, 48, 153] [12, 22, 39] [6, 50]	[21, 45, 49, 55] [29, 34, 72, 92] [35, 39, 40] [42, 47, 65] [54, 57, 58] [50, 56, 71] [40, 41, 89] [25, 59, 90] [43, 60, 95]	
	global	[12, 53, 74] [21, 50]	[50]	
Example-based		[12, 21, 43] [6, 74, 96]	[17, 52, 57] [13, 25, 40]	[32] (generative models)
Counterfactual		[12, 37] [21, 82]	[27, 100] [57, 65]	
Concept-based			[61, 62, 71] [63, 64, 67] [57, 99]	
Other		[11, 88] [7, 10]	[1, 9, 15, 154] [3, 13, 51] [3, 56, 58] [36, 49, 55] [16, 28, 85] [33, 38, 68]* [8, 23, 76] * [69, 70]*	[2] [18, 19, 20] † [20, 26, 84] † [66, 81, 83] † [14, 30, 85] † [5, 91] †

Papers are categorized according to types of explanations (column) and types of models (row). * denotes papers using recommendation systems as models; † denotes papers proposing novel interpretable interfaces as studied models.

A. Models and Explanations

As our selected core papers comprise a large spectrum of AI models, data modalities, and explanation approaches, we initially list the models and explanation techniques deployed along with the corresponding core paper references in Table III. It presents the utilization of explanation types in columns and model types in rows. The explanation methods used is organized according the the taxonomy by Molnar [151]. First, there are intrinsically interpretable models, also known as *white-box models*. For instance, white-box models include decision trees and linear models. Second, there are *black-box models* that provide no parameter access or are too complex to be explained in a human-understandable way [152]. These include ensembling techniques such as Random Forests or neural models.

As for explanation techniques, we identified five key types in the scope of the surveyed papers (rows of Table III). Most frequently used are feature-based (attribution) explanations, for instance, SHAP (Shapley additive explanations [147]) and LIME (Local Interpretable Model-Agnostic Explanations [146]). There is a clear differentiation between local, instance-wise, explanations and global explanations that apply to the model in its entirety. For instance, the weights of a linear model have a global scope. This differentiation is common among these feature-based explanations, where most of the papers using local explanations. Other popular explanation types are example-based explanations, counterfactual explanations, which aim at providing actionable suggestions for attaining a user-preferred prediction by changing certain input features, and concept-based explanations, which use meaningful high-level concepts such as objects or shapes to explain a prediction.

Besides these four main types of explanations, there are other explanations such as rules [11], [88] or game strategies [7], [10] when AI plays games. More details about concrete models and explanations can be found in Appendix B, available online.

B. Measurements

The effectiveness of explanations can be characterized from several angles. We specifically identified the categories of trust, understanding, usability, and human-AI collaboration performance. In this section, we give an overview of the contexts in which each of these variables is studied and the measures used to quantify them.

1) Trust: User trust is studied in decision-making applications such as image classification [13], [17], (review) deception detection [25] or loan approval [27]. Besides decision making, [5], [8], [16], [18], [19], [23] study user trust in the domain of recommendation systems. Whether explainable ML models can increase user trust in the medical domain is studied in [1], [6], [9]. Moreover, Colley et al. [3] measure user trust in an autonomous driving application with and without explanations.

Trust measures used in much of the existing research can be divided into two groups: *self-reported* and *observed* trust [155]. Self-reported trust is commonly measured by asking users to fill out questionnaires whereas observed trust is quantified by humans' agreement with the model's decisions. In Table III in Appendix, available online, trust measures in these two groups are listed. The agreement rate of users with the model decisions is commonly used [9], [11], [12], [25] as a measure of observed trust. Parallel to observed trust measurement, van der Waa et al. [156] ascribe the user's alignment behaviors to the *persuasive power* of model explanations, i.e., the capacity to convince users to follow model decisions despite the correctness. As an extension, trust calibration is defined based on this measure. For example, a high agreement rate to wrongly made decisions represents *overtrust*, while a low agreement rate to correct decisions means *undertrust* [12]. In self-reported measurements, researchers either utilize well-developed questionnaires or self-designed ones, with the exception of [4] which conducts a semi-structured interview to explore user opinions. Several works [6], [11], [13], [16], [17], [18], [19], [24], [27] propose their own questionnaires. Among these, a subgroup [13], [16], [18], [19], [24] simply asks users to rate a single statement such as "I trust the system's recommendation/decision", which is named as one-dimensional trust by [8]. When deploying previously

proposed questionnaires [2], [3], [5], [7], [8], [10], [21], [22], [23], [157], Trust in Automation [158] is the most commonly used one, in which the underlying constructs of trust between human and computerized systems are explored.

2) Understanding: An important goal of explanation techniques is to foster users' understanding of complex ML systems. An important separation has to be made between users' perceived understanding and their actual comprehension of the underlying model, as the two often do not agree [35], [40]. Cheng et al. [22] explicitly differentiate between *objective* understanding and self-reported understanding, which we term *subjective* understanding in this work. While subjective understanding is usually measured through questionnaires, measuring objective understanding requires a proxy task where the users' understanding is put to a test. Additionally, user studies can be run to assess how well users can understand the explanation itself (and not the underlying model). This can be an important sanity check and is particularly used in the domain of conceptual explanations [62], [159], where the intelligibility of concepts needs to be verified. We refer to the third category as *understanding of explanations* but defer its detailed findings to Appendix C.3, available online.

Objective Understanding: Works in the subdomain of objective understanding deploy proxy tasks to verify users' understanding of a model's inner workings. The most commonly considered domain in works on understanding is finance [35], [39], [40], [47], [48], [49], [53] followed by image classification [13], [21], [52]. One of the most critical design choices when assessing objective understanding is the selection of a suitable proxy task. Doshi-Velez and Kim [119] argue that the task should *"maintain the essence of the target application"* that is anticipated. One of the most prominent tasks is forward simulation [119], [140]. This task demands subjects that are given an input to simulate, i.e., predict, the model's output. The extent to which participants can successfully provide the model's output is also referred to as *simulatability* [140]. However, scholars have designed many more tasks to quantify understanding and applied them across a variety of data modalities (cf. Table 2 in Appendix, available online for an exhaustive listing).

We briefly describe other common tasks below. A special variant of forward simulation is called *relative simulation*. In this task, users predict which example out of a predefined choice will have the highest prediction score (or class probability). *A manipulation or counterfactual simulation task* [119] asks users to manipulate the input features in such a way that a certain model outcome (counterfactual) is reached. Users' performance on this task can be used as a proxy for their understanding. Lipton [140] pointed out that simulatability can only be a reasonable measure, if the model is simple enough to be captured by humans and that simpler tasks are required otherwise. An example could be a *feature importance* query, where users have to tell which features are actually used by the model. A directed and more local version of this task is *marginal effects queries*, where the subjects predict how changes in a given input feature will affect the prediction (e.g., *"Does increasing feature X lead to a higher prediction of Y being class 1?"*). Because explanations should allow the identification of weaknesses in models, the task of *failure prediction* measures the accuracy of users' prediction when the model prediction is wrong.

Subjective Understanding: Besides the objective understanding which is supported by performance indicators, understanding of a model may be subjective, i.e., it may depend on a user's own perception. The most commonly used applications that measure subjective understanding are various recommendation system setups [16], [33], [34], [38].

Most of the works assess the subjective understanding of a user with a post-task questionnaire. Guo et al. [7] adapted a popular questionnaire designed for recommendation systems by Knijnenburg et al. [160], while Bell et al. [39] accommodated the questionnaire which originally intended to measure the intelligibility of differenet explanations by Lim and Dey [161]. On the other hand, agreement to simple subjective statements such as *"I understand this decision algorithm"* [22], *"I understand how the AI..."* [13], [17] or *"The explanation(s) help me to understand..."* [33] can be collected to assess subjective understanding.

3) Usability: Usability is a key concern of every HCI system and thus applies to almost all domains. This is reflected in the surveyed papers, where usability is studied in a wide range of setups and contexts. We also include application-specific performance measures in this category.

Based on the measurements in the user studies, we refined usability into measures of helpfulness, workload (cognitive load), satisfaction, ease of use and detecting undesired behaviors of the system, as shown in Table I. To assess workload (cognitive load), NASA-TLX scale [162] is used in [3], [6], [16], [21], [66], while Abdul et al. [48] measure cognitive load by capturing the log-reading time of memorizing the explanation. Most of the works use self-designed questionnaires or statements to measure satisfaction [6], [16], [18], [19], [29], [30], [69], [70], however, the Explanation Satisfaction Scale [163] can be deployed as an established alternative [1], [47]. Helpfulness can be assessed by simply asking for subjective ratings of the explanations for accomplishing a specific task [13], [46], [56], [65], [67], [68]. Colley et al. [3] use an adapted version of the System Usability Scale proposed in [164].

Using model explanations to audit models is one purpose of explainability [129]. Some of the surveyed works study how model explanations can assist users in detecting undesired behaviors of models. These issues mainly include (perceived) unfairness in the model decision-making [38], [74], [78], [79], biases in models [72] or features [57], and wrong decisions (failures) [24] in the studied papers. A detailed summary of types of undesired behaviors is listed in Table VI. In the undesired behavior detection, the effectiveness of explanations is evaluated by objective performance measures, such as the number of bugs identified [71], the share of participants that identify a certain bias [57, First Experiment] or by the deviations between model predictions and human predictions for unusual samples [53]. The perception of users regarding fair treatment by a system has primarily been researched in high-stakes applications such as granting loans [27] or granting bail for criminal offenders [73], [74], [75]. For example, [73], [74], [75] investigate the fairness of COMPAS, a commercial criminal risk estimation tool that was

TABLE IV
EXPERIMENTAL DESIGNS IN CORE PAPERS

	Experimental Design		
	Between-Subjects	Within-Subjects	Mixed
Papers	[5, 7, 8, 12, 15, 27, 59]		
	[17, 21, 22, 23, 25, 72]	[1, 3, 4, 9, 19]	
	[11, 28, 32, 40, 46, 95]	[10, 18, 21, 24, 70]	[2, 13, 16, 52, 66]
	[47, 49, 50, 51, 53, 84]	[13, 26, 35, 52, 91]	[10, 28, 34, 64, 74]
	[36, 37, 38, 39, 43, 56]	[57, 71, 78, 81, 93]	[12, 33, 65, 83, 146]
	[29, 30, 54, 90, 92, 96]	[6, 62, 63, 67, 69]	[61]
	[12, 38, 57, 58, 73]	[14, 41, 45, 60, 68]	
	[75, 76, 77, 78, 82]		

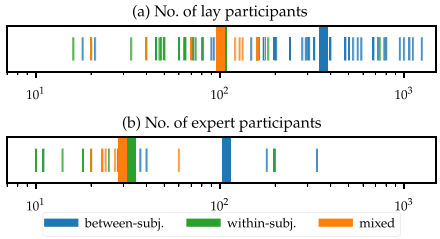


Fig. 2. Distribution of participant numbers in the surveyed user studies by design and participant type (each bar represents one study). Per-design means are indicated in bold.

used in the US to help make judicial bail decisions. It is also considered in everyday use-cases such as news [38] and music [77] recommendations, or possible career suggestions [76], where a bias in the underlying system can be to the detriment of the user. As the assessment of fairness is a very subjective matter, questions regarding perceived fairness are prevalent, e.g., "how the software made the prediction was fair" [74], which can be answered on 5- or 7-point Likert scales [2], [27], [38], [73], [74], [75]. Among these works, an effective explanation is the one that can either increase or decrease the fairness perceptions, since the aim of explanations is to show fairness or unfairness. An exhaustive overview of measures for usability is given in Table IV of the Appendix, available online.

4) Human-AI Collaboration Performance: The goal of human-AI teaming is to improve the performance in AI-supported decision-making above the bar set by humans or an AI alone [89]. Improving human performance with the help of AI has been considered in games [10], [88], question answering tasks [89], [91], deception detection [25], [90] and topic modeling [29], [30].

The most common assessment is to rate AI-aided human performance by the percentage of correctly predicted instances in the decision-making process [25], [89], [90]. Paleja et al. [10], however, define the performance as the time to complete the task. In [88], performance is measured in a game-based application, chess, using a winning percentage (which is commonly used in sports) as well as a percentile rank of player moves.

C. Experimental Design and Analysis

There are three common experimental settings when conducting user evaluation: between-subjects (or between-groups) designs, within-subjects designs, and mixed designs that combine elements of both. An overview of the designs found in the core papers and their participant numbers is presented in Table IV and Fig. 2, respectively.

1) Between-Subjects: With slightly above $55\,\%$ of the user studies conducted in a between-subjects manner, i.e., one subject is only exposed to one condition, this design choice is most common in the XAI literature. The number of participants in the between-subjects manner usually starts at around 30 participants, while it may go up to 1070 in total for 3 conditions as in [17] and to 1250 for 5 conditions in [53]. However, the number of participants can be limited when the studied application is designed for specific groups of lay persons, which cannot be easily recruited from the Internet platforms such as Amazon Mechanical Turk. For instance, Ooge et al. [8] use 12 school students per condition. Some authors place particular emphasis on participants being similar to the average demographic [73], [75].

The conditions usually include the different explanation techniques in combination with other parameters such as the model, data set, data modality, or a number of features used as independent variables. Note that a full grid design with many independent variables may quickly result in a very high number of conditions, which in turn requires many participants. The outcome variable of interest is commonly measured on a numerical or ordinal scale right away, however, in the fairness domain, qualitative analyses are sometimes obtained through conducted interviews or written responses [2], [27], [73].

The statistical analysis directly follows from this design. If one is interested in identifying significant differences between the groups, common statistical hypotheses tests are used. For overall comparison, one or two-way ANOVA tests are the most commonly used statistical tool. Interesting post-hoc comparisons between two groups can be made with a standard T-test, if the data is normally distributed with equal variance, or by using non-parametric tests such as the Wilcoxon rank-sum test (also known as Mann-Whitney U-test) for comparison of two populations (e.g, [57]) or the Tukey HSD test (e.g., [49]) for multiple populations. When running multiple post-hoc tests, some works make use of the Bonferroni correction (e.g, [57]).

2) Within-Subjects: Around $30\,\%$ of the papers use the within-subjects design, where each participant sequentially passes through all conditions and provides feedback. Fewer participants are recruited in within-subjects experiments compared to the between-subjects ones. Hence, they are particularly popular when participants with restrictive characteristics, such as domain-specific professional expertise, are required. For example, Suresh et al. [9] and Rong et al. [26] recruit fourteen medical professionals and five radiologists in their user studies, respectively. The small number of medical experts contributing to the user study is a limitation [26], however, it is often the case in expert user research. Gegenfurtner et al. [165] evaluate 73 sources and point out that the majority of these studies include only five, maybe ten experts. Besides the medical domain, other works [3], [4], [19], [21] also invite subjects with particular professions such as engineers in a technology company. When no specific knowledge is required, however, participant

TABLE V
USER STUDY FINDINGS WHEN USING MODEL **EXPLANATIONS** AS EVALUATION DIMENSIONS

		Evaluation Dimension: Explanations Effect of explanations compared to **no explanations**	
		Positive	Non-positive / Mixed
Trust		[13]: example-based, rule-based explanations [16]: example-based explanations for recommendations [27]: feature importance [10]: decision-tree explanation for policy [28]: explanation corpus given by researchers [25]: feature-based (saliency map), example-based explanations [8]: explanations for recommendations [6]: rationale-based, example-based and feature-based (best) explanations for online symptom checkers [15]: confidence scores	[3]: positive in simulation but no improvement in real-word [1]: explanations for medical suggestions (Doctor XAI [155]) pos.for observed trust but insignificant for reported trust [12]: feature-based explanations increase appropriate trust slightly but counterfactual explanations inconclusively [21, 22, 24]: feature-based explanation, insignificant [11]: rule-based explanation, insignificant [15]: Shapley values, insignificant [29]: feature-based explanation, negative
Understanding	Obj.	[22, 53] white-box model [40] feature importance, LIME (tabular) [46] counterfactuals+cues (audio) [50] manipulatability improved by white-box log. reg. [54] saliency maps (image) [59] saliency maps for bias detection and strategy identification [12] counterfactuals+feature importance	[39]: SHAP, negative for black-box model (education domain) [39]: Insignificant difference btw. black-box and white-box models [40]: Prototypes, Anchors, LIME on textual data insignificant [46]: Counterfactuals and Concepts insignificant (audio data) [50]: Simulatability results insignificant for LIME, IG, surrogate model on BERT and Logistic Regression Model, Manipulatability insignificant for BERT [58]: Insignificant results for GRAD-CAM, Saliency Map, uncertainty scores in VQA [59] saliency maps for failure prediction (image) [60]: saliency maps, negative for a mix of three interpretation techniques in simulation task
	Sub.	[13]: example-based, rule-based explanations [28]: explanation corpus given by researchers [12]: feature-, example- and counterfactual-based [38]: explanations provided by [167] for Facebook News Feed [16]: example- and feature-based explanations [17]: example-based explanations [34]: feature importance, SHAP and LIME [35]: feature importance, SHAP	[22]: white-box model, insignificant [39]: white-box < black-box, both insignificant [31]: feature importance explanation (transparent system) can be distracting
Usability		[81]: counterfactuals, pos. for usability [16, 47]: example-based explanations, pos. for satisfaction [67]: CAM-related explanations, pos. for helpfulness [6]: rational-, feature-, example-based explanations, pos. for satisfaction [70]: content-based explanations, pos. for satisfaction [83]: explanations regarding driving information, pos. for ease of use [13]: example-based and rule-based explanations, pos. for helpfulness [71]: local, global, visual (saliency map) explanations, pos. for bug identification [65]: attribution methods and conceptual explanations, pos. for usefulness [84]: feature-based, pos. for reliability [24]: (proposed) template-based expl. pos. for debugging and usefulness [27]: feature importance, counterfactual explanations pos. for perceived fairness	[82]: counterfactuals, significant for helpfulness/usability but insignificant for usefulness [1]: ontology-based explanation, insignificant for satisfaction [65]: attribution methods and conceptual explanations, insignificant for ease of use [24]: visual explanations increases usefulness, but improvement is insignificant [3]: pos. for cognitive load/usability (simulation), but insignificant in real-world [29]: feature-based explanations, negative for satisfaction [38]: informing users about the algorithmic decisions, negative ranking scores of recommendations, insignificant for perceived fairness [27]: highlight features only, insignificant for perceived fairness [78]: insignificant in between-subjects but significant in within-subjects for perceived fairness
Human-AI Collaboration Performance		[88]: textual explanations with domain knowledge (in chess) [25, 90]: feature-based explanations [91]: exampled-based for experts, feature-based for novices [93]: contrastive explanations [13]: example-based and rule-based explanations [95, 96]: example-based explanations, attributions (AI correctness prediction) [96]: important parts in images as explanations	[25]: exampled-based, insignificant [15, 89]: feature-based explanations, insignificant

Effects of explanations compared to the baseline (control group) of "no explanations" on measured quantities. Effects are divided into "positive" where explanation information is given, and "non-positive / mixed" where negative impact is marked with underlines.

numbers reach up to 740 also for within-subjects designs [93]. For within-groups designs, the Wilcoxon signed-rank test (e.g., used by [35], [52]) is the most common method to compare paired samples for significant differences. Repeated-measures ANOVA is a common analysis tool, when multiple comparisons are required (see, e.g., [35]).

3) Mixed: The smallest group of studies, about 15%, use a mixture of between- and within-subjects settings. In these works, subjects are first assigned randomly to one group, where they are exposed to multiple conditions. Anik and Bunt [2] use knowledge background in machine learning as a between-subjects factor to divide the participants into three groups (expert, intermediate and beginner), while inside each group participants interact with explanations in the context of four different scenarios (e.g., facial expression recognition or automated speech recognition). Dominguez et al. [16] make the presence of explanations a between-subjects condition and different types of explanations a within-subjects factor in the group with model explanations. A particular challenge for such a study design is that statistical tools from both the independent-samples and dependent-samples categories need to be combined.

V. FINDINGS OF USER STUDIES

In this section, we summarize the primary findings from the core papers. Table V lists findings with respect to four measured quantities. To build an overview of the findings, we divide papers according to their evaluation dimensions, i.e., the independent variables in the user studies. When using the presence of explanations as the evaluation aspect, the findings are summarized in Table V. The listed impacts using explanations are to be seen in comparison with a control group without explanations. Effects are divided into two groups: (1) Positive effects, for example, increasing user trust or understanding; (2) Non-positive effects: the effect can be negative, or not significantly positive (neural), or a mixture of different effects (e.g., feature-based explanations have positive effects but counterfactual explanations do not). Beyond the explanations themselves, other possible evaluation dimensions such as that might have an impact on the perception of XAI, for instance, AI technology literacy, model performance, or the dimensionality of the data. Instead of using the mere presence of explanations, many works compare different explanation techniques with each other (see Appendix D, available online for more details).

TABLE VI
OVERVIEW OF RESULTS FOR UNDESIRED BEHAVIOR DETECTION USING MODEL EXPLANATIONS

Type	Paper	Detection Result
Wrong decisions (failures)	[24, 53, 71]	High detection rate in [24]; Moderate detection rate (50%) in [71]; Lower detection rate in [53]
Biases in features used by models	[57, 71]	Moderate detection rate (50%) in [71]; Moderately high detection rate (>50%)
Discrimination/Biases in decisions	[72]	Humans perform well in bias detection (accuracy=88.9%) and bias description (66.7%)
Unfairness (perceived) in models	[2, 27, 38, 74] [73, 78, 79]	Succeed to judge [27, 74, 78]; Not succeed to judge [2]; Not always (no consensus) [38, 73]

As various research questions and findings are addressed in 97 core papers, many papers compare explanation types in order to choose a preferable one, it is not possible to cover all results in one table. Based on them, we outline some interesting trends in the effectiveness of explanations on user experience: (1) Explanations are effective in improving users' subjective understanding; (2) The effectiveness of explanations in increasing user trust and usability of models is not clear; (3) Explanations are not good at convincing users that models are fair; (4) Interactivity of the model has positive impact on user trust, understanding and model usability. The first three statements can validated through the number of papers obtaining positive or non-positive effects in each category, while the last finding is extracted from Table V in the Appendix, available online, which details findings with on other independent variables. We encourage the reader to consider the short summary of *primary* findings in the tables and check for further details according to their specific interests. In the following section, we highlight some findings for each category of measurement.

Trust: Among the papers comparing the effect of using explanations to using no explanations, or placebo (randomly generated) explanations [8], [25], about half of the papers validate that explanations have a positive impact on user trust [1], [8], [10], [13], [16], [25], [27], [28], while the other half cannot verify this hypothesis [3], [11], [12], [21], [22], [24]. For instance, Colley et al. [3] investigated the explanations in an autonomous driving task and discover that the trust is improved in simulation but not with the real-world footage. Another example of the mixed effect of using explanations is found in [12], where (minimal) evidence is found that feature-based explanations help increase appropriate trust, but counterfactual explanations do not.

Apart from using explanations as independent variables, the user personalities or expertise may also affect their perceptions [2], [17], [18], [22], [23], [30]. Millecamp et al. [18] captured personal characteristics in the aspects such as the Locus of Control defined by Fourier ("the extent to which people believe they have power over events in their lives"), Need for Cognition ("a measure of the tendency for an individual to engage in effortful cognitive activities") or Tech-Savviness ("the confidence in trying out new technology"). However, no significant interaction effect could be found between the personal characteristics and the trust. Liao and Sundar [5] studied a recommendation system asking users' personal data with different explanations. They hypothesized that explanations in a "help-seeker" style and using the pronoun "I" would gain more trust of users than the explanations formalized in a "help-provider" style. Nevertheless, However, the opposite result is found and using self-referential expression resulted in lower affective trust. Model performance together with model explanation was studied in [17] for an image recognition task. The authors found out when images were recognized (high model performance), users feel the system more capable ("capability" is defined as a belief of trust).

Understanding: The fundamental question in this subdomain is to find out which explanation technique is most beneficial for increasing the user's understanding of a machine learning model. As pointed out earlier, understanding can be measured both in a subjective and objective manner.

We first discuss results on objective understanding. The goal of increasing objective understanding was explicitly posed by Alqaraawi et al. [54] who reported that saliency maps have a positive effect on understanding. Wang and Yin [12] show that counterfactual explanations and feature importance increase users objective understanding. On the contrary, Sixt et al. [57] find none of their examined explanation techniques (counterfactuals, conceptual explanations) superior to a baseline technique consisting of example images for each class and the work by Hase and Bansal [40] reveals that many explanations (including anchors, prototypes) have no effect in increasing objective understanding, which LIME on tabular data being the only exception. Apart from the explanation, several other factors have been identified to have an effect on objective understanding. Hase and Bansal [40] suggest that the *data modality* may have a non-negligible impact on how different explanation techniques increase understanding. Some results highlight that the *choice of proxy task* is influential. Arora et al. [50] show that their manipulatablity task revealed differences remained hidden when forward simulation is used. In spite of these findings, Buçinca et al. [13] underline that preferred explanations may be different in a real-world application from a simulated one. Regarding the *type of model*, there is disagreement on whether white or black-box models can lead to increased objective understanding. While black-box models without explanations resulted in higher simulation performance than white-box models with SHAP values in [39], Cheng et al. [22] observe that white-box models increase simulatability and also conclude that *interactivity* is an important factor when it comes to objective understanding.

In comparison with the objective understanding, the research question in the subdomain subjective understanding is to find out how explanations impact user's *perceived* understanding [7], [12], [17], [22], [32], [33], [34], [37], [56]. There exist a trend of using model explanations to improve subjective understanding [13], [16], [17], [28], [34], [38], [167]. However, Chromik et al. [35] challenge the improvement in perceived understanding with the cognitive bias named *illusion of explanatory depth* (IOED) [168], which means that laypeople often have overconfidence bias in their understanding of complex systems. Their results confirm the IOED issue in XAI, i.e., questioning users' understanding by asking them to apply their understanding in practice consistently reduces their subjective understanding. Explanations can have different impacts on subjective and objective understandings [22], where white-box explanations

increase objective understanding but do not have significant impact on subjective understanding. Similar disagreements have been observed in multiple other works [40], [167]. Radensky et al. [33] examine the joint effects of local and global explanations in a recommendation system and their results provide evidence that both are better than either alone.

Usability: Similar to trust, it is not clear whether explanations are effective in improving users' perceptions of helpfulness, satisfaction or other dimensions of usability. For instance, in [16], [30], [47], the explanations have a positive effect on satisfaction, while no significant effects on satisfaction are observed in [18], [19], [29], [69]. Parallel to trust, Smith-Renner et al. [29] provide evidence for the hypothesis that it is harmful to user trust and satisfaction to show explanations by highlighting the important words in a text classification task. A strong correlation between self-reported trust and satisfaction can also be observed in [3], where explanations have a positive impact in a simulated driving environment, but no significant effects when using real-world data. Beyond explanations, Nourani et al. [56] study the order of observing system weakness and strengths, which reveals that encountering weakness first results in a lower rate of usage of system explanations than encountering strength first. Schoeffer et al. [27] find out that showing feature importance scores or counterfactual explanations (or a combination of both) for explaining decisions helps increase the perceived fairness, whereas highlighting important features without scores does not. However, several studies don't show a significant difference between scenarios with and without explanations [27], [38], [78]. Effects of explanations may be dependent on input samples, as shown in [67]. The authors show that both Debiased-CAM and Biased-CAM improve the helpfulness for a weakly blurred image, however, there is no significant improvement for unblurred or strongly blurred images. When used to assist users in detecting undesired behaviors, model explanations are likely to identify various types of problems that exist within models or data, as demonstrated by [57], [71], [72]. However, successful detection is not guaranteed. For example, Poursabzi-Sangdeh et al. [53] show that users with model explanations are less able to identify incorrect predictions. A limitation of current detection methods is that users may have varying assessments, such as perceived unfairness and irrelevance [53], [71], [73], regarding the features used in models for decision-making. Due to this limitation, the effectiveness of methods assessed through self-reported data may face challenges in generalizability as discussed in [73]. Yet, these methods generally offer a *one-size-fits-all* solution, failing to account for variations in individual assessments.

Human-AI Collaboration Performance: A strain of works [25], [88], [90], [91], [95], [96], [96] show that viewing explanations can improve human accuracy in making decisions, especially with feature-based explanations taking text data as input [25], [90], [91]. When using example-based explanations in text classification, there is no improvement in human performance [25]. Likewise, utilizing explanations has no significant impact on human performance in [89], [92], but simply showing model predictions has a positive effect in [92]. Experts and novices perceive explanations differently, for example, Feng and Boyd-Graber [91] conclude that the performance gain of novices and experts comes from different explanation sources. Paleja et al. [10] reveal that explanations can improve novices' performance but decrease experts' performance. Additionally, less complex models with explanations can better convince humans in correct decisions [90].

VI. A Guideline for XAI User Study Design

Learning from the best practices of the previous works, we summarize a handy guideline for XAI user study, which serves as a checklist for XAI practitioners. This guideline contains suggestions to avoid pitfalls that researchers could easily overlook. We introduce our guidelines in the order of before, during and after user studies, which reflects user study design, execution and data analysis, respectively.

Before the User Study: When designing a user study, the first step is to decide what to measure. To define the measured quantities, one can consider two alternatives: using a general definition or an application-based quantity that is specific to the application at hand. The former one refers to a quantity that is borrowed from previous well-established research, such as using "trust in automation" [2], [3], [21] or "general trust in technology" [7], [23]. To further construct "trust" as a quantitative measurement, one needs to examine how existing work has conceptualized "trust" in both social sciences context as well as XAI and technical context [169]. The application-based quantity depends on the application goal, for instance in a chess game [88], the measurement is the human winning percentage with the help of model explanations (Human-AI collaboration).

From Table V, we can see that previous works have frequently struggled to prove the effectiveness of XAI even with respect to a control group that is without explanation. When only different explanation techniques are considered, there will always be one winner explanation, but the overall benefit will remain undisclosed (see examples in Appendix D, available online). Therefore, it is important to compare with a baseline without explanations to rigorously show the strength of XAI. When a comparative design is explicitly desired, baselines such as random explanations [28], [41], [62]).

When deploying a proxy task, its difficulty should be gauged and monitored carefully. In the past, the forward simulation task has been criticized as being unrealistically complex for domains such as computer vision [54]. Thus, other proxy tasks such as feature importance queries [57] or manipulatability checks [32], [50] were proposed. Another important point is to choose a proxy task that is simplified, but features many characteristics of the application in mind [119]. Notably, the proxy task should be designed close to the final anticipated application, as even slight differences in the tasks may void the validity of the findings on the proxy tasks in the real world [13].

The measurement is often dependent on the definition of the measured quantity. For instance, in [58], the objective understanding is measured as failure prediction (the accuracy of user prediction when the model prediction is wrong). For subjective measurements such as subjective understanding or trust, one-dimensional measures (i.e., simply rating one

question such as *"Do you trust the model explanation?"*) have the drawback that they cannot completely reflect different constructs of measured quantities [8]. Moreover, subjective questions and behavioral measurements often appear to be weakly correlated. For example, the users state that they trust model but they do not really follow the model suggestions [11]. Similar findings have been made with respect to objective and subjective understanding [12], [35], [40]. To overcome this limitation, both self-reported and observed measures shall be used in parallel.

Besides the measures introduced in Section IV-B, there are several psychological constructs that can be deployed to evaluate multiple facets of the interaction between humans and XAI. For instance, the *subjective task value* in the expectancy-value framework is often used to analyze subjective motivation to take any actions [170], which is not thoroughly studied in the XAI experience yet. The subjective task value consists of intrinsic value (enjoyment), attainment value (importance for one's self), utility value (usefulness), and cost (the amount of effort or time needed) [170], [171]. A good explanation interface should be positively correlated with the subjective task value, consequently boosting one's interest and motivation to use the model explanation. With regard to the cost of using model explanations, cognitive load is popularly measured in the current literature with conventional Likert scales [162], [172]. Cognitive load researchers study the validity of different visual appearances in rating scales beyond numerical Likert scales, i.e., pictorial scales such as emoticons (faces with different emotions), or embodied pictures of different weights [173]. Their results demonstrate that numerical scales are more proper in complex tasks while pictorial scales are for simple ones.

Pre-registration using online platforms such as AsPredicted[1] has become a common practice in recent years [174]. In this process, researchers submit a document detailing their planned study online before initiating the data collection. Among other details, the pre-registration includes the measured variables and hypotheses, data exclusion criteria, and the number of samples that will be collected. An exhaustive pre-registration can provide evidence against the findings being a result of selective reporting or p-hacking [175] and thus strengthen the credibility of a study. Expert interviews and pre-studies following a think-aloud protocol [176], e.g., in the references [32], [46], are often mentioned as helpful tools to develop the explanation system and the study design and gain first qualitative insights or complement the qualitative analysis [13], [65].

When preparing for a user study, it is important to plan for explicit steps and to have a backup plan for different situations. Before participants arrive, it is helpful to provide them with information such as where the researchers will meet with them, what they need to bring, and how they can prepare for the study. If conducting the experiment in person, send participants a reminder the day before and provide them with your contact in case they cannot find the experiment site or they need to cancel the experiment session. Once participants arrive, make sure the researchers have a plan that covers all stages of the experiment. The protocol should cover small details (e.g., where participants should leave their backpacks, water bottles, and lunch boxes) and plans for unexpected situations (e.g., uncooperative participants and multifunctional systems). How to obtain participants' consent should be an important part of the procedure. Additional procedure is required for obtaining consent when working with vulnerable populations (e.g., children and pregnant women), in which case alternative consent procedures might take place. Another benefit of pre-designing the experiment script is to fine-tune the language to avoid inadvertent cues. Researchers can unintentionally pass on their expectations to participants through verbal and nonverbal behavior, which might result in participants' skewed performance towards the researchers' desire [169]. To ensure a sound experiment procedure and to protect the integrity of the data, it is worthwhile to put in much effort to design a detailed experiment script.

During the User Study: A sufficient number of participants is the prerequisite of a solid user study analysis. To get a rough estimate of common sample sizes, we refer the reader to the participant statistics in Fig. 2 where we analyze the subject numbers in different experimental designs. For instance, around 350 users without any specific expertise are averagely recruited in between-subject experiments. However, we would like to underline that the required number of participants is highly specific to the study design and should be determined individually, for instance by conducting a statistical power analysis [177]. Additionally, recruited participants should have the same knowledge background as the end users that applications are designed for. For instance, when evaluating an interface explaining loan approval decisions to bank customers, it is not proper to include only students whose major is computer science, since they may have prior knowledge of how model explanations work. Note that the design of an AI application requires different audiences across the project cycle, thus model explanations need to evolve as well [178].

To uphold high-quality standards of the collected data, attention or manipulation checks are essential to filter out careless feedback. This particularly applies to long surveys or online surveys with lay users. Kung et al. [179] justify the use of these checks without compromising scale validity. In within-subject experiments, a random order of conditions is necessary to avoid order effect [1]. Participants can learn knowledge of data or examples shown in the previous conditions, and Tsai et al. [6] choose to use a Latin square design to avoid the learning effect.

After the User Study: After the data collection, statistical tests are run to find significant effects. The applicable tests used are determined by experimental designs and the form and distribution of the data. Generally, ANOVA tests and T-test are usually used when comparing distributions between different conditions. Structural Equation Models (SEM) or multi-level models are used for mediation analysis. More details of statistic tools can be found in Section IV-C. Distributional assumption checks should be applied. When Likert-type data is collected as in most of the questionnaires, non-parametric tests such as paired Wilcoxon signed-rank test, or Kruskal-Wallis H test for multiple groups can be used to avoid normality assumptions.

If multiple measures are aggregated into a single instrument, it is important to assess the validity of this aggregation with

[1] [Online]. Available: https://aspredicted.org

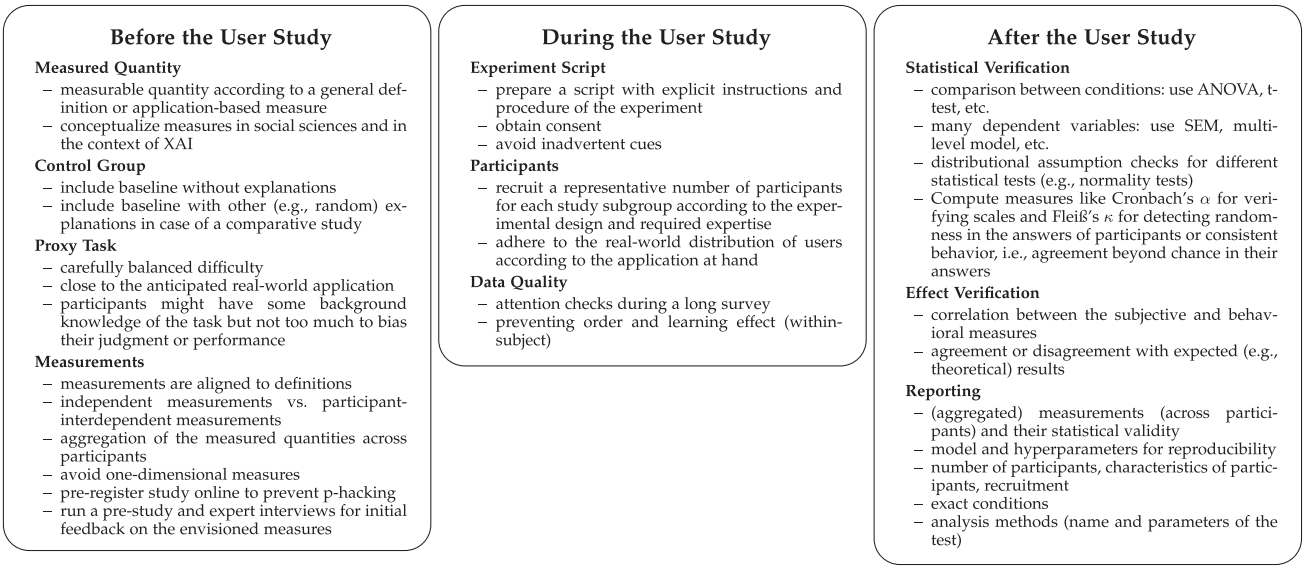


Fig. 3. Summary cards of the guidelines extracted from past XAI user studies.

reliability measures such as the tau-equivalent reliability (also known as Cronbach's α). For example, if objective and subjective measures of a quantity, such as understanding are combined, it is necessary to verify that there is sufficient agreement. If multiple items (e.g., data samples or visualizations) are rated by several subjects, statistics such as Cohan's κ as Fleiß's κ for more than two raters [180] can be used to assess agreement beyond chance between these raters and serve as an indication for the reliability of the ratings.

In the final writing phase, it is essential to report sufficient details that allow readers to estimate the explanatory power of the study. On the level of participants, this should include the total number of participants and how many are assigned to each treatment group, their recruitment, consent and incentivization, and the exact treatment conditions they are subjected to. Furthermore, some descriptive statistics of the collected data can help readers assess the characteristics of the adequacy of the statistical tools used. Regarding the analysis, we found it important to mention how the underlying assumptions of the statistical tests used were checked and to mention the exact variant of the test used (e.g., stating "a two-way ANOVA with the independent variables X and Y" is used instead of just mentioning that ANOVA-test is used).

VII. Future Research Directions

Our survey of recent and ongoing XAI research also helps us identify research gaps and distill a few directions for future investigations. In this section, we highlight these directions and summarize our findings.

A. Towards Increasingly User-Centered XAI

We advocate that user-centered methods should be used not only to assess XAI solutions (e.g., through user studies) but also to design them (e.g., through user-centered design). By explicitly modeling and involving users in the design phase and not just in a post-hoc manner during the evaluation phase, we expect the development of XAI solutions that better respond to user needs. As discussed in [117], there are two aspects of human-centered AI: (1) AI systems that understand humans with a sociocultural background and (2) AI systems that help humans understand them. The former point can guide the design of AI systems. In this section, we discuss XAI research that leverages this insight.

The process of explaining a machine's decisions to human users can be viewed as a teaching-learning process where the XAI system is the teacher and the human users are the students. From a user-centered perspective, the problem of designing effective teaching methods to enhance the student's (i.e., user's) learning outcomes is essential to human-centered XAI algorithms. To leverage the ability of humans and address unique user's needs, it is important to review studies and findings from psychology and education. These studies provide insights into how humans perceive other intelligent agents (humans or artificial agents) and how they utilize limited information to infer and generalize. Understanding how humans think and learn will help XAI developers build and design systems that are not only informative but also user-friendly to people with different backgrounds. In this section, we discuss three pedagogical frameworks, namely (1) the expectancy-value motivation theory, (2) the theory of mind, and (3) hybrid teaching, to shed light on incorporating such methods in computational approaches. Inspired by existing work in pedagogy and XAI, we provide implications for designing future transparent AI systems and human-centered evaluations.

Expectancy-Value Motivation Theory: Human interaction with XAI interfaces can be viewed as an activity where humans learn about the model's inner workings through explanations and then achieve an understanding of the models. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [181]. This research

problem is widely considered in educational psychology through the lens of expectancy-value motivation theory. For instance, Hulleman et al. [171] propose to utilize *interventions* to increase the perception of usefulness (utility value) to subsequently increase motivation and final performance. Intervention here refers to identifying the relevance of model explanations to the user's own situation, which can be a prompt question while working with the interface. Moreover, when utilizing model explanations in human-AI collaboration, explanations can be seen as a type of "scaffolding" (prompt during a task) proposed in a conceptual framework in education.

Theory of Mind: When interacting with XAI systems, humans form mental models of the machine learning algorithms that reflect their belief of how the algorithms work. The formation of these mental models comes from observing explanations or examples given to the human, who often subconsciously applies the observations in a few examples to the broader understanding of the whole machine learning system. This incredible ability to infer, rationalize, and summarize other intelligent agent's decisions is known as the Theory of Mind (ToM) in psychology. Based on this theory, the Bayesian Theory of Mind (BToM) provides a probabilistic framework to predict inferences that people make about mental states underlying other agents' actions. Recent work, at the intersection of XAI and robotics, indicates that humans also attribute ToM to artificial agents that they observe or interact with. Guided by these user-centered results, several works at the intersection of XAI and robotics have utilized BToM to create a simulated user, and then use it to generate helpful explanations.

Hybrid Teaching: Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist. One way of categorizing these strategies is through the following three concepts: (1) direct teaching, (2) indirect teaching, and (3) hybrid teaching. *Direct teaching* utilizes direct instructions that are teacher-centered, involve clear teaching objectives, and are consistent with classroom organizations. In XAI applications, direct teaching methods generate explanations by selecting representative examples of an agent's decisions to convey the patterns in its policy. In contrast, *indirect teaching* is student-centered and encourages independent learning. In the XAI perspective, methods utilizing indirect teaching provide users with tools to actively and independently explore an AI system. Technically, direct teaching focuses on providing guidance (using a computational approach) to assist users in building an understanding of a machine, whereas indirect teaching (often through a user interface) enables users to address individual learning preferences and mitigate individual confusion about the AI. To leverage the advantages of the two teaching strategies, *hybrid teaching* has been widely used in human-to-human teaching with an emphasis on interactivity. Recent work [182] indicates that hybrid teaching reduces the amount of time for a user to understand an agent's policy compared to direct and indirect teaching, and is more subjectively preferred by the participants. Building on this, future XAI systems can consider using hybrid teaching methods that (i) generate direct instructions to provide guidance to user's understanding of an AI system; and (ii) provide methods to allow users to interact with the agent.

Explanations through Large Language Models (LLMs): The recent rise of Large Language Models [183], [184] naturally opens up new research directions. There is a growing interest in leveraging their unprecedented capabilities [185] to offer explanations for model decisions [186], [187]. Through their natural language interface, LLMs offer the possibility to build interactive explainers [188]. Intriguingly, textual explanations can also be used as subsequent inputs to LLMs which may help to solve subsequent problems and result in superior performance [189]. This technique, referred to as chain-of-thought reasoning [190], opens up an interesting research territory combining interpretability and performance considerations.

B. Open Research Problems

1) Automatic versus Human-Subject Evaluations: With automatic evaluations, we refer to evaluation methods that do not require human subjects, which corresponds to the functionally-grounded metrics discussed in [119], [120]. These metrics aim to test desiderata around the "faithfulness"/"fidelity"/ "truthfulness" of model explanations [120], [121], [191]. Faithfulness of explanations is defined as that explanations are indicative of true important features in the input [191]. The automatic evaluations aim at capturing general objectivity which is independent from downstream tasks, while human evaluations are contextualized with specific use cases. Generally speaking, automatic evaluations and human evaluations tackle different research challenges: the former objectively examines how truly explanations reflect models and the latter one measures how humans perceive models through explanations (although there existing algorithms for automated evaluation designed to align with human evaluations, which we will discuss later). All explanations used in human-subject experiments should have satisfying performance in automatic evaluations, i.e., the explanations should be able to faithfully unbox the model. This verification step is essential to guarantee the validity of the empirical user study and to ensure that users are not tricked by unfaithful explanations. However, in most current human-subject experiments, the functional faithfulness of explanations is not thoroughly verified beforehand. Using unfaithful explanations could lead to the problem that only the placebo effect of explanations is measured. Ideally, a good explanation should be faithful to the model as well as understandable by users.

2) Identifying and Handling Confounders: Existing research underscores the vulnerability of model explanation studies to significant confounding effects. For instance, Papenmeier et al. [155] reveal that user trust can be more influenced by model accuracy than the faithfulness of the explanation itself. Similarly, Yin et al. [192] demonstrate that the accuracy score perceived by users and the one shown to users contribute to trust formation.

A different problem is that good explanations also reveal weaknesses of the model. However, when seeing unexpected explanations, users may express their negative feelings about the model through negative ratings of the explanations. Therefore, good model explanations should help users *calibrate* their trust [26], [193], i.e., trust the model's decision when it is correct but distrust it otherwise. There is a disagreement on how to

handle such cases: When evaluating model fairness, several works [2], [27], [38], [73], [75] reckon the increase in perceived fairness as positive, while Dodge et al. [74] define the decrease as positive. Other factors, such as the temporal occurrence of model errors (Nourani et al. [56]), and the dimensions of models (Ross et al. [32], Poursabzi et al. [53]), also come into play.

In summary, these confounding elements suggest that users might be led to put more trust in oversimplified, deceptive, or simply unfaithful explanations. To mitigate this, we recommend meticulous analysis, control and reporting of potential confounders, such as explanation faithfulness and model accuracy, across various test conditions. More advanced measures have been suggested as well. For instance, Schoeffer and Kuehl's [79] propose *appropriate fairness perceptions*, which measures whether people increase or decrease their fairness perceptions depending on the algorithmic fairness of the underlying model. Nevertheless, the thorough investigation of confounding factors remains a challenge. Calibrated measures that are less prone to confounding can be a valuable step forward.

3) Mitigating Personal Biases for XAI: Most XAI techniques and corresponding designed user studies provide *one-size-fits-all* solutions. Individual bias, rooted in a user's mental framework, influences the user's perception of a model. It should be considered in XAI design, development, and evaluation procedures. Several studies that aim to explain reinforcement learning policies utilize cognitive science theories to create a model of the human user [181], [182], [194], [195]. They then generate explanations based on this human model and verify the benefits of tailoring explanations for individual user models. Within the scope of XAI, [196], [197] utilize a Bayesian Teaching framework to capture human perception of model explanations. In user studies, depending on cultural and educational background, participants may likely give different feedback [31]. This kind of personal bias can be mitigated by deploying a large sample size and recruiting participants who are representative of the target audience. We advocate that personal biases should be taken into account in the realm of XAI development.

4) Human-in-the-Loop and Sequential Explanations: In several relevant cases, such as online recommendation systems, users are not only confronted with an explanation once but instead view decisions and potential explanations repeatedly. Recent work in this domain [35] has shown that the order of decisions and explanations may indeed have an effect on user perception and understanding. The AI model may continue to shape the user's mental model over time. The differences between the single-use and the sequential setting still remain to be thoroughly investigated.

5) Proxy Tasks Should Be Close to Real-World Tasks: When using proxy tasks to evaluate models, for instance, to measure subjective understanding, there is a great choice of tasks present in the literature. A good proxy task should have the following features: (1) it has close real-world connections [119]; (2) users or participants have some background knowledge of the task but not too much to affect their judgment or performance during the task; (3) the task is not too complicated to implement or there exists an existing implementation but was used for different purposes (i.e., not used for XAI); and (4) it has connections to existing work. Yet, the link between evaluations through different proxy tasks and real-world applications has not been made very explicit to date. Buçinca et al. [13] show that the outcomes of proxy evaluations can be different from a real-world task. More specifically, the widely accepted proxy tasks, where users are asked to build the mental models of the AI, may not predict the performance in actual decision-making tasks, where users make use of the explanations to assist in making decisions. The results show that users trust different explanations in the proxy task and the actual decision-making task. Therefore, we argue that further research is required to uncover the links between current proxy tasks and on-task performance or to devise new proxy tasks with a verified connection to actual tasks.

6) Simulated Evaluation as a Cost-Efficient Solution: As human-subject experiments are costly to conduct, Chen et al. [198] propose a simulated evaluation framework (`SimEvals`) to select potential explanations for user studies by measuring the predictive information provided by explanations. Concretely, the authors consider three use cases where model explanations are deployed: forward simulation, counterfactual reasoning, and data debugging. Human performance is measured for these three tasks with different explanations. If there is a significant gap in settings of using two types of explanations, the simulated evaluation can also observe such a gap under the same task settings as well. Meanwhile, first attempts to simulate human textual responses in a given context using large language models show that models can provide surprisingly anthropomorphic answers [199]. Undoubtedly and also affirmed by Chen et al. [198], it is not yet realistic to replace human evaluation with the simulated framework as other factors e.g., cognitive biases can affect human decisions. To better simulate human evaluations, more effort should be directed towards modeling human cognitive processes. Concurrently and with appropriate caveats, XAI researchers should also leverage existing and approximate models of human cognition to enable rapid prototyping and assessment of explanations. Section VII-A discusses several candidate human cognition models and highlights recent XAI works [181], [182] that utilize this "Oz-of-Wizard" paradigm.

VIII. Conclusion

In recent years, there has been a proliferation of XAI research in both academia and industry. Explainability is a human-centric property [141] and therefore XAI should be preferably studied by taking humans' feedback into account. In this work, we investigated recent user studies for XAI techniques through a principled literature review. Based on our review, we found out that the effectiveness of XAI in users' interaction with ML models was not consistent across different applications, thus suggesting that there is a strong need for more transparent and comparable human-based evaluations in XAI. Furthermore, relevant disciplines, such as cognitive psychology and social sciences in general, should become an integral part of XAI research.

We comprehensively analyzed the design patterns and findings from previous works. Based on best-practice approaches and measured quantities, we propose a general guideline for

human-centered user studies and several future research directions for XAI researchers and practitioners. Thereby, this work represents a starting point for more transparent and human-centered XAI research.

REFERENCES

[1] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, "Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–9.

[2] A. I. Anik and A. Bunt, "Data-centric explanations: Explaining training data of machine learning systems to promote transparency," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–13.

[3] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, "Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–1.

[4] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–19.

[5] M. Liao and S. S. Sundar, "How should AI systems talk to users when collecting their personal information? effects of role framing and self-referencing on Human-AI interaction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–14.

[6] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll, "Exploring and promoting diagnostic transparency and explainability in online symptom checkers," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–17.

[7] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, "Building trust in interactive machine learning via user contributed interpretable rules," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 537–548.

[8] J. Ooge, S. Kato, and K. Verbert, "Explaining recommendations in E-learning: Effects on adolescents' trust," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 93–105.

[9] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan, "Intuitively assessing ML model reliability through example-based explanations and editing model inputs," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 767–781.

[10] R. Paleja, M. Ghuy, N. Ranawaka Arachchige, R. Jensen, and M. Gombolay, "The utility of explainable AI in ad hoc human-machine teaming," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 610–623.

[11] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your AI: Expertise and explanations," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 240–251.

[12] X. Wang and M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 318–328.

[13] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 454–464.

[14] X. Peng, M. Riedl, and P. Ammanabrolu, "Inherently explainable reinforcement learning in natural language," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 16178–16190.

[15] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 295–305.

[16] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra, "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 408–446.

[17] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 258–262.

[18] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, "To explain or not to explain: The effects of personal characteristics when explaining music recommendations," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 397–407.

[19] C.-H. Tsai and P. Brusilovsky, "Beyond the ranked list: User-driven exploration and diversification of social recommendation," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2018, pp. 239–250.

[20] T. Li, G. Convertino, R. K. Tayi, and S. Kazerooni, "What data should I protect? recommender and planning support for data security analysts," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 286–297.

[21] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.

[22] H.-F. Cheng et al., "Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[23] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[24] D. H. Kim, E. Hoque, and M. Agrawala, "Answering questions about charts and generating visual explanations," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[25] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proc. ACM Conf. Fairness Accountability Transparency*, 2019, pp. 1–13.

[26] Y. Rong, N. Castner, E. Bozkir, and E. Kasneci, "User trust on an explainable ai-based medical diagnosis support system," 2022, *arXiv:2204.12230*.

[27] J. Schoeffer, N. Kuehl, and Y. Machowski, ""there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making," 2022, *arXiv:2205.05758*.

[28] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: A technique for explainable AI and its effects on human perceptions," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 263–274.

[29] A. Smith-Renner et al., "No explainability without accountability: An empirical study of explanations and feedback in interactive ML," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[30] A. Smith-Renner, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater, "Digging into user control: Perceptions of adherence and instability in transparent models," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 519–530.

[31] A. Springer and S. Whittaker, "Progressive disclosure: Empirically motivated approaches to designing effective transparency," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 107–120.

[32] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez, "Evaluating the interpretability of generative models by interactive reconstruction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.

[33] M. Radensky, D. Downey, K. Lo, Z. Popovic, and D. S. Weld, "Exploring the role of local and global explanations in recommender systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–7.

[34] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn, "Improving understandability of feature contributions in model-agnostic explainable AI tools," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–9.

[35] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, "I think I get your point, AI! the illusion of explanatory depth in explainable AI," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 307–317.

[36] J. Rebanal, J. Combitsis, Y. Tang, and X. Chen, "XAlgo: A design probe of explaining algorithms' internal states via question-answering," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 329–339.

[37] U. Kuhl, A. Artelt, and B. Hammer, "Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting," 2022, *arXiv:2205.05515*.

[38] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.

[39] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *Proc. ACM Conf. Fairness Accountability Transparency*, 2022, pp. 248–266.

[40] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5540–5552.

[41] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, and N. T. Vu, "Human interpretation of saliency-based explanation over text," 2022, *arXiv:2201.11569*, .

[42] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11396–11404.
[43] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky, "HIVE: Evaluating the human interpretability of visual explanations," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–298.
[44] M. Szymanski, M. Millecamp, and K. Verbert, "Visual, textual or hybrid: The effect of user expertise on different explanations," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 109–119.
[45] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 10526–10536.
[46] W. Zhang and B. Y. Lim, "Towards relatable explainable ai with the perceptual process," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–24.
[47] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki, "Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 807–819.
[48] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, "COGAM: Measuring and moderating cognitive load in machine learning model explanations," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.
[49] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5968–5979.
[50] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig, "Explain, edit, and understand: Rethinking user study design for evaluating model explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5277–5285.
[51] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hern ández-Lobato, "Getting a {clue}: A method for explaining uncertainty estimates," in *Proc. Int. Conf. Learn. Representations*, 2021.
[52] J. Borowski et al., "Exemplary natural images explain {CNN} activations better than state-of-the-art feature visualization," in *Proc. Int. Conf. Learn. Representations*, 2021.
[53] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–52.
[54] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: A user study," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 275–285.
[55] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1527–1535.
[56] M. Nourani et al., "Anchoring bias affects mental model formation and user reliance in explainable ai systems," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 340–350.
[57] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf, "Do users benefit from interpretable vision? a user study, baseline, and dataset," in *Proc. Int. Conf. Learn. Representations*, 2022.
[58] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, "Do explanations make VQA models more predictable to a human?," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1036–1042.
[59] J. Colin, T. Fel, R. Cadene, and T. Serre, "What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 2832–2845.
[60] H. Shen and T.-H. Huang, "How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2020, pp. 168–172.
[61] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 20554–20565.
[62] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 9277–9286.
[63] T. Leemann, Y. Rong, S. Kraft, E. Kasneci, and G. Kasneci, "Coherence evaluation of visual concepts with objects and language," in *Proc. Int. Conf. Learn. Representations WS*, 2022.
[64] I. Laina, R. Fong, and A. Vedaldi, "Quantifying learnability and describability of visual concepts emerging in representation learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13112–13126.
[65] Y. Wang, P. Venkatesh, and B. Y. Lim, "Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–28.
[66] D. L. Arendt, N. Nur, Z. Huang, G. Fair, and W. Dou, "Parallel embeddings: A visualization technique for contrasting learned representations," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 259–274.
[67] W. Zhang, M. Dimiccoli, and B. Y. Lim, "Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–32.
[68] J. Gao, X. Wang, Y. Wang, and X. Xie, "Explainable recommendation through attentive multi-view learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3622–3629.
[69] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "Personalized explanations for hybrid recommender systems," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 379–390.
[70] C.-H. Tsai and P. Brusilovsky, "Explaining recommendations in an interactive hybrid social recommender," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 391–396.
[71] A. Balayn, N. Rikalo, C. Lofi, J. Yang, and A. Bozzon, "How can explainability methods be used to support bug identification in computer vision models?," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–16.
[72] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12187–12198.
[73] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proc. Wide Web Conf.*, 2018, pp. 903–912.
[74] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 275–285.
[75] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 392–402.
[76] C. Wang et al., "Do humans prefer debiased AI algorithms? a case study in career recommendation," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 134–147.
[77] N. N. Htun, E. Lecluse, and K. Verbert, "Perception of fairness in group music recommender systems," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2021, pp. 302–306.
[78] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–14.
[79] J. Schoeffer and N. Kuehl, "Appropriate fairness perceptions? on the effectiveness of explanations in enabling people to assess the fairness of automated decision systems," in *Proc. Companion: Companion Pub. Conf. Comput. Supported Cooperative Work Social Comput.*, 2021, pp. 153–157.
[80] T. Donkers, T. Kleemann, and J. Ziegler, "Explaining recommendations by means of aspect-based transparent memories," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 166–176.
[81] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.
[82] U. Kuhl, A. Artelt, and B. Hammer, "Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning," 2022, *arXiv:2205.03398*.
[83] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, "ExplAIn yourself! transparency for positive UX in autonomous driving," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–12.
[84] S. Choi, K. Aizawa, and N. Sebe, "FontMatcher: Font image paring for harmonious digital graphic design," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2018, pp. 37–41.
[85] P. Le Bras, D. A. Robb, T. S. Methven, S. Padilla, and M. J. Chantler, "Improving user confidence in concept maps: Exploring data driven explanations," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.

[86] R. Shang, K. K. Feng, and C. Shah, "Why am I not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations," in *Proc. ACM Conf. Fairness Accountability Transparency*, 2022, pp. 1330–1340.

[87] J. Dodge, A. A. Anderson, M. Olson, R. Dikkala, and M. Burnett, "How do people rank multiple mutant agents?," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 191–211.

[88] D. Das and S. Chernova, "Leveraging rationales to improve human task performance," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2020, pp. 510–518.

[89] G. Bansal et al., "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–16.

[90] V. Lai, H. Liu, and C. Tan, ""why is' Chicago'deceptive?," towards building model-driven tutorials for humans," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[91] S. Feng and J. Boyd-Graber, "What can ai do for me? evaluating machine learning interpretations in cooperative play," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 229–239.

[92] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6618–6626.

[93] K. Z. Gajos and L. Mamykina, "Do people engage cognitively with AI? impact of AI assistance on incidental learning," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2022, pp. 794–806.

[94] M. Liao, S. S. Sundar, and J. B. Walther, "User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–14.

[95] G. Nguyen, D. Kim, and A. Nguyen, "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26422–26436.

[96] M. R. Taesiri, G. Nguyen, and A. Nguyen, "Visual correspondence-based explanations improve AI robustness and human-AI team accuracy," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 34287–34301.

[97] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Syst. Appl.*, vol. 69, pp. 29–39, 2017.

[98] S. Yang, M. Korayem, K. AlJadda, T. Grainger, and S. Natarajan, "Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach," *Knowl.-Based Syst.*, vol. 136, pp. 37–45, 2017.

[99] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, "Towards conversational search and recommendation: System ask, user respond," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 177–186.

[100] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, pp. 362–386, 2020.

[101] H. Cui et al., "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 2090–2096.

[102] Y. Rong, C. Han, C. Hellert, A. Loyal, and E. Kasneci, "Artificial intelligence methods in in-cabin use cases: A survey," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 3, pp. 132–145, May/Jun. 2021.

[103] R. R. Murphy, "Introduction to AI robotics," *Ind. Robot: An Int. J.*, vol. 28, no. 3, pp. 266–267, 2001.

[104] K. Rajan and A. Saffiotti, "Towards a science of integrated AI and robotics," *Artif. Intell.*, vol. 247, pp. 1–9, 2017.

[105] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable AI for robotics," *Sci. Robot.*, vol. 2, 2017, Art. no. eaan6080.

[106] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, vol. 286, pp. 800–809, 2018.

[107] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction," *BMC Med. Res. Methodol.*, vol. 19, 2019, Art. no. 64.

[108] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome: Clin. Res. Rev.*, vol. 14, pp. 337–339, 2020.

[109] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Appl. Soft Comput.*, vol. 91, 2020, Art. no. 106263.

[110] M. Ala'raj, M. F. Abbod, M. Majdalawieh, and L. Jum'a, "A deep learning model for behavioural credit scoring in banks," *Neural Comput. Appl.*, vol. 34, pp. 5839–5866, 2022.

[111] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, p. 38, 2018.

[112] N. Van Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, and V. Kostakos, "Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study," in *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–21, 2019.

[113] T. Sourdin, "Judge V robot?: Artificial intelligence and judicial decision-making," *Univ. New South Wales Law J.*, vol. 41, no. 4, pp. 1114–1133, 2018.

[114] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 469–481.

[115] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Manage. Rev.*, vol. 61, pp. 15–42, 2019.

[116] D. Castelvecchi, "Can we open the black box of AI?," *Nature News*, vol. 538, pp. 20–23, 2016.

[117] M. O. Riedl, "Human-centered artificial intelligence and machine learning," *Hum. Behav. Emerg. Technol.*, vol. 1, pp. 33–36, 2019.

[118] U. Ehsan and M. O. Riedl, "Human-centered explainable AI: Towards a reflective sociotechnical approach," in *Proc. Int. Conf. Human-Comput. Interact.*, 2020, pp. 449–466.

[119] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv: 1702.08608*.

[120] M. Nauta et al., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Comput. Surv.*, vol. 55, pp. 1–42, 2023.

[121] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6021–6029.

[122] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 18770–18795.

[123] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 1069–1078.

[124] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 3, pp. 209–218, Jun. 2019.

[125] Workshop, "ExSS-ATEC: Explainable smart systems for algorithmic transparency in emerging technologies," in *Proc. 25th Int. Conf. Intell. User Interfaces Companion*, vol. 1, 2020.

[126] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst. (TiiS)*, vol. 11, no. 3/4, pp. 1–45, 2021.

[127] Q. Yang, N. Banovic, and J. Zimmerman, "Mapping machine learning advances from HCI research to reveal starting places for design innovation," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–11.

[128] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[129] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, 2020, vol. 58, pp. 82–115.

[130] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Proc. Explainable AI: Interpreting Explaining Visualizing Deep Learn.*, 2019, pp. 5–22.

[131] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.

[132] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, 2019, Art. no. 832.

[133] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.

[134] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–28.

[135] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.

[136] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv: 2006.11371*.

[137] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021.
[138] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," *ACM SIGKDD Explorations Newslett.*, vol. 22, pp. 18–33, 2020.
[139] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Model. User-Adapted Interact.*, vol. 27, pp. 393–444, 2017.
[140] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, pp. 31–57, 2018.
[141] Q. V. Liao and K. R. Varshney, "Human-centered explainable AI (XAI): From algorithms to user experiences," 2021, *arXiv:2110.10790*.
[142] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of Human-AI decision making: A survey of empirical studies," 2021, *arXiv:2112.11471*.
[143] J. J. Ferreira and M. S. Monteiro, "What are people doing about XAI user experience? a survey on ai explainability research and practice," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2020, pp. 56–73.
[144] N. Bevan, "International standards for HCI and usability," *Int. J. Hum.-Comput. Stud.*, vol. 55, pp. 533–552, 2001.
[145] W. Iso, "9241–11: 1998, Ergonomic requirements for work with visual display terminals (VDTs)-Part 11: Guidance on usability," *Int. Org. Standardization*, vol. 45, no. 9, 1998.
[146] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?," explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.
[147] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[148] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[149] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," in *A Practical Guide*, 1st ed., Berlin, Germany: Springer, 2017.
[150] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.
[151] C. Molnar, "Interpretable machine learning," pp. 26–27, 2020.
[152] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019.
[153] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1721–1730.
[154] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor XAI: An ontology-based approach to black-box sequential data classification explanations," in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 629–639.
[155] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust," 2019, *arXiv: 1907.12652*.
[156] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, 2021, Art. no. 103404.
[157] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.
[158] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergonom.*, vol. 4, pp. 53–71, 2000.
[159] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.
[160] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," in *User Modeling User-Adapted Interaction*. Berlin, Germany: Springer, 2012.
[161] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proc. 11th Int. Conf. Ubiquitous Comput.*, 2009, pp. 195–204.
[162] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.
[163] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv: 1812.04608*.
[164] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale (SCS)," *KI-Künstliche Intelligenz*, 2020.
[165] A. Gegenfurtner, E. Lehtinen, and R. Säljö, "Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains," *KI-Ku nstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020.
[166] K. Cotter, J. Cho, and E. Rader, "Explaining the news feed algorithm: An analysis of the "news feed FYI," blog," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2017, pp. 1553–1560.
[167] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–15.
[168] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cogn. Sci.*, vol. 26, pp. 521–562, 2002.
[169] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *ACM Trans. Human-Robot Interact.*, vol. 10, pp. 1–31, 2020.
[170] J. Eccles, "Expectancies, values and academic behaviors," *Achievement Achievement Motives*, vol. 58, pp. 58–74, 1983.
[171] C. S. Hulleman, J. J. Kosovich, K. E. Barron, and D. B. Daniel, "Making connections: Replicating and extending the utility value intervention in the classroom," *J. Educ. Psychol.*, vol. 109, 2017, Art. no. 387.
[172] F. G. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach," *J. Educ. Psychol.*, vol. 84, pp. 429–434, 1992.
[173] K. Ouwehand, A. V. D. Kroef, J. Wong, and F. Paas, "Measuring cognitive load: Are there more valid alternatives to likert rating scales?," *Front. Educ.*, *Frontiers Educ.*, vol. 6, p. 702616, 2021.
[174] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "Pre-registration: Why and how," *J. Consum. Psychol.*, vol. 31, pp. 151–162, 2021.
[175] U. Simonsohn, L. D. Nelson, and J. P. Simmons, "P-curve: A key to the file-drawer," *J. Exp. Psychol.: Gen.*, vol. 143, pp. 534–547, 2014.
[176] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA, USA: MIT Press, 1984.
[177] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, San Francisco, CA, USA: Academic, 2013.
[178] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle," in *Proc. Des. Interactive Syst. Conf.*, 2021, pp. 1591–1602.
[179] F. Y. Kung, N. Kwok, and D. J. Brown, "Are attention check questions a threat to scale validity?," *Appl. Psychol.*, vol. 67, pp. 264–283, 2018.
[180] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, pp. 378–382, 1971.
[181] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring computational user models for agent policy summarization," in *IJCAI: Proc. Conf.*, 2019, Art. no. 1401.
[182] P. Qian and V. Unhelkar, "Evaluating the role of interactivity on improving transparency in autonomous agents," in *Proc. 21st Int. Conf. Auton. Agents Multiagent Syst.*, 2022, pp. 1083–1091.
[183] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.
[184] ChatGPT, Introducing, "OpenAI," 2023. Accessed: Feb. 17, 2023. [Online]. Available: https://openai.com/blog/chatgpt
[185] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.
[186] W. Zhou et al., "Towards interpretable natural language understanding with explanations as latent variables," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6803–6814.
[187] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, "Reframing Human-AI collaboration for generating free-text explanations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 632–658.
[188] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," 2023, *arXiv:2302.07257*.
[189] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! leveraging language models for commonsense reasoning," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4932–4942.
[190] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.

[191] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7786–7795.
[192] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.
[193] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *Proc. Int. Conf. Healthcare Inform.*, 2015, pp. 160–169.
[194] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 33, no. 33, 2011.
[195] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Auton. Robots*, vol. 43, pp. 309–326, 2019.
[196] S. C.-H. Yang, N. E. T. Folke, and P. Shafto, "A psychological theory of explainability," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 25007–25021.
[197] S. C.-H. Yang, W. K. Vong, R. B. Sojitra, T. Folke, and P. Shafto, "Mitigating belief projection in explainable artificial intelligence via Bayesian teaching," *Sci. Rep.*, vol. 11, 2021, Art. no. 9863.
[198] V. Chen, N. Johnson, N. Topin, G. Plumb, and A. Talwalkar, "Use-case-grounded simulations for explanation evaluation," 2022, *arXiv:2206.02256*.
[199] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans," 2022, *arXiv:2208.10264*.

Yao Rong received the MSc degree in electrical and computer engineering from the Technical University of Munich, Germany, in 2019. She is currently working toward the doctoral degree with the Human-Centered Technologies for Learning Group, the Technical University of Munich. From 2022 to 2023, she served as a visiting scholar with the DATA Lab, Rice University. Her research interests lie in human-centered AI, explainable AI, and human-AI interaction technologies.

Tobias Leemann received the MSc degree from the University of Erlangen-Nuremberg, Germany, in 2020. He is currently working toward the PhD degree with the University of Tübingen, Germany where his research is focused on trustworthy machine learning. Specifically, his research interests include the quality assessment of interpretability techniques and the intersections of interpretability, fairness and privacy.

Thai-Trang Nguyen is graduated with a BSc degree in computer science from the University of Tübingen, Germany. She is currently working toward the MSc degree with the same university. Furthermore, she served as a research assistant, the Human-Computer Interaction group from 2019 to 2022.

Lisa Fiedler is currently working toward the BSc degree in media informatics from the University of Tübingen, Germany. Additionally, she works as a student assistant for the Human-Computer Interaction Group at the University of Tübingen.

Peizhu Qian is currently working toward the PhD degree in computer science with Rice University, USA working with Dr. Vaibhav Unhelkar on problems in human-robot interaction, robot transparency, and explainable AI. Her research interest lies in building a mutual understanding between a robot and its human collaborators. Her work applies psychology theories to computational frameworks, enabling robots to communicate their objectives.

Vaibhav Unhelkar received the MS degree in aeronautics and astronautics and the PhD degree in autonomous systems, in 2015 and 2020, respectively. He is an assistant professor of computer science with Rice University, USA where he leads a research group in the emerging area of Human-Centered AI and Robotics. Unhelkar earned his undergraduate degree in aerospace engineering from the Indian Institute of Technology in Bombay, in 2012. From the Massachusetts Institute of Technology, where he worked in the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Tina Seidel received the diploma degree in psychology from the University of Regensburg (Germany) and Vanderbilt University Nashville (USA), in 1998, and the PhD degree with excellence, in 2002 from the Leibniz Institute for Science and Mathematics Education Kiel (Germany). She holds the Friedl Schoeller Chair for Educational Psychology with the School of Social Sciences and Technology, Technical University of Munich, Germany. Her research focuses on teaching and teacher education. She has established a Teacher Research & Training Simulation Center that conducts several research projects funded by the German Science Foundation and the German Federal Ministry of Education and Research.

Gjergji Kasneci received the MSc degree in computer science and mathematics from the University of Marburg, in 2005, and the PhD degree from the University of Saarland - while with the Max Planck Institute - in 2009. He then worked with Microsoft Research Cambridge, the Hasso Plattner Institute, and SCHUFA Holding AG, where he served as CTO from 2017 to 2022. Between 2018 and 2023, he led the Data Science and Analytics Group with the University of Tübingen as an Honorary professor. In 2023, Gjergji Kasneci was appointed professor of Responsible Data Science with the Technical University of Munich.

Enkelejda Kasneci received the PhD degree in computer science from the University of Tübingen, in 2013. She was postdoctoral researcher and a Margarete-von-Wrangell Fellow with the University of Tübingen. She is a distinguished professor for Human-Centered Technologies for Learning with the Technical University of Munich and Core Member of the Munich Data Science Institute. Her research evolves around Human-Centered Technologies and AI systems that sense and infer the user's cognitive state, the level of task-related expertise, actions, and intentions based on multimodal data and provide information for media and assistive technologies in many activities of everyday life, and especially in the context of learning.

APPENDIX A
DATA-DRIVEN BIBLIOMETRIC ANALYSIS

To perform a data-driven bibliometric analysis of the references and citations for all papers[1], we first collected common references from each category. As we had to deal with a large number of papers, a keyword representing the research topic was assigned to each paper. In this way, we could group the papers according to their content. Concretely, the references were extracted directly from the studied papers (in pdf format). The follow-up works that cite each core paper were retrieved from the Google Scholar platform using the Python API ("Scholarly" [1]). The same API was used to extract abstracts from Google Scholar for all references and citations. Based on the paper titles and abstracts, we utilized GPT-4 [2] to tag the papers with keywords and subsequently reviewed the sensibility of these keywords manually. We visualized papers in a 2-dimensional semantic space according to their keyword embeddings using t-SNE [3].

We illustrate the research domains that are fundamental to XAI user studies in Figure 1 (**Left**). Note that for presentation clarity, we only visualized works that were used as references in at least five of the core papers. Similarly to foundations in XAI user studies, we are interested in knowing who will eventually benefit from the findings of XAI user studies. Figure 1 (**Right**) demonstrates the "consumers" of the human-centered XAI core papers (i.e., research domains influenced by the core papers), with each dot representing a research topic. The size of the dots is determined by the number of citations in the set of core papers obtained from this research area.

By studying these two aspects (i.e., foundations and impact), we grasp a clear overview of relevant topics in the research landscape of XAI user studies. More importantly, we can better spot the nascent but pertinent areas for future work such as cognition-driven analysis tools in XAI. We release raw data and code for analyses at https://github.com/yaorong0921/hxai-survey.

A.1 Foundation of XAI User Studies

Through analyzing references in the core papers, we provide XAI researchers with several indispensable literature sources in this field, which can inspire them when organizing their projects. In total, there are over 3000 references from all the core papers, and we pay close attention to the references which are cited at least by ten core papers (ca. 50 papers). In Table 1, we categorize these papers according to their topics. The first group of papers is survey papers about XAI, which are thoroughly discussed in Sec.2 Related Work. For the theory of XAI, Miller et al. [4] propose to build XAI on social sciences such as cognitive science and psychology, while Wang et al. [5] and Liao et al. [6] provide theoretical guidelines for designing XAI frameworks. An important class of references are XAI methods and the most popularly used ones are listed in "XAI Methods". As suggested by [7, 8], the explanations should be sound and complete and thus bring a positive impact on users. Another motivation for XAI is that it should assist users in building mental models of the AI systems [9]. Previous user studies for ML systems or for explainable interfaces that are referenced for comparisons or serve as templates of user study design. In the end, we list several general works about user trust that may go beyond the scope of XAI.

1. In this section, the word "references" refers to sources contained in the references of one of the core papers while "citations" refers to follow-up works that reference one of the core papers

APPENDIX B
MODELS AND EXPLANATIONS IN XAI USER STUDIES

Black-box models are dominant in the current human-AI interaction research area as we can see that more black-box models are studied. Local feature explanations are popularly used such as LIME [18] and SHAP [20]. Figure 2 demonstrates the chronological overview of frequently adopted XAI techniques for black-box models in user studies from the surveyed papers. However, there are many specific explanation types for certain applications. For recommendation systems, content-based and hybrid explanations are widely used explanations. A content-based explanation is a single-style explanation coming from a content-based recommendation system, while a hybrid explanation contains multiple explanation styles such as user-based or item-based, which is provided by a hybrid recommendation system [49, 50, 51]. For instance, Dominguez et al. [52] provide a content-based explanation as *"Painting A is 85% similar to the Painting B that you like"*. Tsai et al. [53], however, use hybrid explanations in textual and visual explanation formats.

APPENDIX C
MEASUREMENT DETAILS

C.1 Trust

Table 3 lists the trust measurement. Most of the works deploy questionnaires to measure user trust (self-reported), where a 7-point or 5-point Likert scale is commonly used. Many works design their own questionnaires [42, 43, 52, 85, 103, 104, 107, 108, 109]. To measure trust in an objective manner, many works choose to use the agreement rate of humans [33, 68, 84, 85].

C.2 Usability

Table 4 demonstrates the measures used for the usability of explanations. We divide usability into five sub-categories: workload (cognitive load), helpfulness, satisfaction, undesired behavior detection and ease of use and others. User perceptions of workload, helpfulness, satisfaction and ease of use are subjective and often measured with questionnaires. However, for debugging tasks, it can be measured objectively such as using the accuracy of the user confirming the correctness of answers from a question-answering model and the time for solving this task [109].

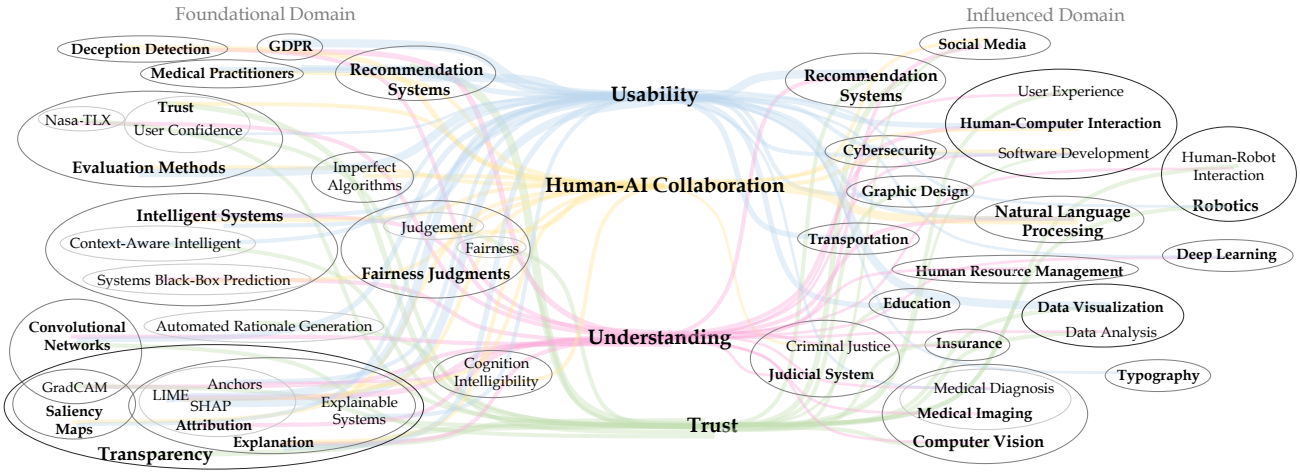


Fig. 1: Illustration of the **foundational** research domains (**Left**): Each dot represents a referenced paper, whose size reflects the number of studied core papers referring to it. Illustration of **influenced** research domains (**Right**): Each dot represents a research topic, whose size refers to the number of papers on the same topic. For a clear depiction, only several important research domains are labeled with text. Lines are used to depict reference links, with thicker lines representing a greater number of links. Core paper categories are in blue (**Middle**). Circles are used to indicate a hierarchical structure of keywords.

Topic	Fundamental works
Surveys of XAI	[10], [11], [12], [13], [14], [15], [16]
Theories for XAI	[4]: social sciences, [5]: theory for XAI design, [6]: a question bank for XAI design
XAI Methods	[17]: a survey, [18]: LIME, [19]: Anchors, [20]: SHAP, [21]: TCAV, [22]: explaining recommendation systems, [23]: intelligible models, [24]: influence function, [25]: counterfactual explanations, [26]: Integrated Gradient (IG), [27]: saliency maps for images, [28]: GradCAM
Principles of Explanations	[7, 8]: completeness and soundness, [9]: helping users build mental models
User studies for ML	[29]: image retrieval algorithm for medical uses, [30]: interactive model
User studies for XAI	[31]: justice perceptions, [32]: fairness [33]: human-AI team, [34]: usability, [35, 36, 37, 38, 39]: understanding, [40, 41, 42, 43]: trust and understanding
Trust	[44]: trust (calibration), [45]: trust in automation, [46]: impact of model accuracy on trust, [47, 48]: impact of system transparency on trust,

TABLE 1: Fundamental works of the core papers (categorized according to topics).

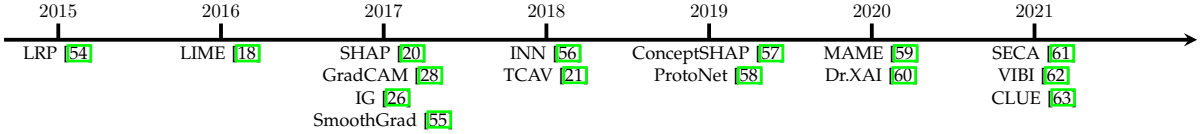


Fig. 2: Chronology of commonly used XAI methods from reviewed papers.

Tasks	Tabular	Image/Video	Text	Other
forward simulation	[37, 64, 65] [19, 63, 66] [40, 67, 68]	[42, 69, 70] [71, 72, 73] [74, 75, 19] (VQA)]	[64, 76]	[77] (Audio)
marginal feature effects	[64, 78, 79] [40, 68]		[64]	
manipulation / counterfactual sim.	[40, 68, 80]	[81]	[76]	
feature importance	[67, 68, 79]	[21, 82]		
failure prediction		[72]		
relative simulation (selection)	[40, 66]			
other	[78] (mental model faithfulness)	[69] (class-wise acc.)		

TABLE 2: Works measuring objective understanding grouped by proxy task/data modality

C.3 Understanding of Explanations

For novel or cognitively challenging types of explanations, it makes sense to verify whether users can make use of the information provided through the explanation. Usually these types of tests are conducted in combination with other measures to establish if the explanations are correctly understood by users and can thus be processed as intended.

In the domain of conceptual explanations [21, 133], such kind of understanding questions are common, to assess semantic coherence of automatically discovered concepts [57, 134, 135, 136]. Assignment tasks, where novel instances should be assigned to existing clusters are commonly used as a proxy to measure the intelligibility [57, 65, 134, 135]. Another option is to assess how well the cluster can be described in natural language which is often referred to as *describability* [134, 135, 136]. Apart from conceptual explanations, Zhang et al. [77] ask multiple choice questions to verify if users understand the differences between the acoustical cues presented and evaluate which cue differences were most noticeable. Wang et al. [114] prompt users

	Studied Paper	Metric	Definition Source	Detail
Observed	[83]	Weight of Advice (WOA)	-	Degree to which the algorithmic suggestion influences the participant's estimate.
	[33, 68, 84, 85, 86, 87]	Agreement rate	-	Percentage of cases in which participants agree with the model. [68] defines the *appropriate trust*, *overtrust* and *undertrust*. [85] defines as *adherence*
Self-reported	[41, 88, 89]	Trust in Automation	[90]	On the 7-point Likert scale. [88] adapts the questions.
	[91, 92]	General trust in technology	[93]	On the 5-point Likert scale.
	[94]	Human-Computer Trust	[95]	On the 7-point Likert scale. [94] adapts the questions.
	[96]	Trust-TAM (Technology Acceptance Model)	[97]	On the 7-point Likert scale. [96] includes other self-designed questions.
	[40]	Trust in human-machine systems	[98]	On the 7-point Likert scale.
	[99]	Unified Theory of Acceptance and Use of Technology Model (UTAUT)	[100]	On the 5-point Likert scale.
	[101]	Human-Robot Collaborative Fluency Assessment	[15]	On the 7-point Likert scale
	[92]	Trusting beliefs and intentions	[102]	On the 7-point Likert scale.
	[42, 85, 103, 104, 105, 106] [43, 52, 87, 107, 108, 109]	Self-designed questionnaire	-	[43, 85, 103] are on the 7-point Likert scale. [42, 104, 107, 108, 109] are on the 5-point Likert scale. [52] rates from 0 to 100. [42, 52, 105, 106, 107, 108, 109] measure one-dimensional trust.
	[110]	Semi-structured interview	-	

TABLE 3: Measures of trust. The measurement is divided into two main groups: "Observed" and "self-reported" trust. The studied core papers using the same measurement are grouped together. The name and the paper reference of the used metrics are listed in the column "Metric" and "Definition Source", respectively. "-" in the column "Definition Source" means that the source is the studied paper. More details about the metrics are given in the last column.

explicitly if the found the explanation easy to understand.

Research questions and Findings. Laina et al. [134] found that feature vectors obtained by contrastive learning approaches such as MoCo [137] or SeLa [138] allow for clusters that are almost as interpretable as human labels. Leemann et al. [136] show the similarity of ResNet-50 embeddings allows to predict how semantically coherent users find a cluster of images. For the acoustical cue, Zhang et al. [77] found that shrillness and speaking rate were most often recognized. Wang et al. [68] found that users reported they understood all types of explanations well without significant differences.

APPENDIX D
FINDINGS

When using explanation types as the evaluation dimension, many works compare their effects without comparing them to a control group (baseline) without explanation methods. Anik et al. [88] argue that many works have proven the usefulness of explanations and therefore no need to include such a control group. Table 6 summarizes the findings of the comparison among different explanations. Table 5 lists results of using other evaluation dimensions beyond explanations.

APPENDIX E
TOWARDS INCREASINGLY USER-CENTERED XAI

In this section, we provide a detailed literature review regarding existing work in pedagogical frameworks, which provides implications for designing future transparent AI systems and human-centered evaluations in Sec. 7.1.

E.1 Expectancy-value Motivation Theory

Human interaction with XAI interfaces can be viewed as an activity where humans learn about the model's inner workings through explanations and then achieve an understanding of the models. The question of how to enhance the efficiency and the outcome of this human learning process is of high importance [147]. This research question is widely considered in educational psychology through the lens of expectancy-value motivation theory [148, 149, 150]. For instance, Hulleman et al. [148] propose to utilize *interventions* to increase the perception of usefulness (utility value) to subsequently increase motivation and final performance. Intervention here refers to identifying the relevance of model explanations to the user's own situation, which can be a prompt question while working with the interface. Moreover, when utilizing model explanations in human-AI collaboration, explanations can be seen as a type of "scaffolding" (prompt during a task) proposed in a conceptual framework in education [151, 152]. Bisra et al. [153] summarize guidelines for effective scaffolding. For instance, different disciplinary descriptions can be used in the scaffolding (explanation prompt) to enhance the user's intuition. Another important, yet often unconsidered point is the role of personality traits in the perception of explanations. For instance, Conati et al. [154] show that the *need for cognition* characteristic, which indicates users' openness towards cognitively challenging tasks, is a determining factor for explanation effectiveness in an intelligent tutoring system. Considering these findings, we see personalized XAI as a relatively underexplored but yet sorely needed research direction.

E.2 Theory of Mind

When interacting with XAI systems, humans form mental models of the machine learning algorithms that reflect their belief of how the algorithms work. The formation of

	Studied Paper	Metric	Definition Source	Detail
Workload	[41, 52, 89, 111, 112]	NASA TLX	[113]	
	[78]	Memory Performance	-	
Helpfulness	[42, 69, 114] [77, 78, 115]	Self-designed questionnaire	-	[42, 69, 114] are on 5-point Likert scale [77, 78, 115] are on 7-point Likert scale
	[116]	Rating	-	Rating from 1 to 5
	[80]	Comparison	-	Users select the most helpful method
Satisfaction	[52, 53, 103, 105, 106, 107, 108]	Self-designed questionnaire	-	[53, 107, 108] are on 5-point Likert scale [103, 105, 106] are on 7-point Likert scale [52] rates from 0 to 100
	[49, 91]	User experience of recommendation system	[93]	[91] adapts the questions on the 5-point Likert scale [49] adapts the questions on the 7-point Likert scale
	[79, 83]	Explanation Satisfaction Scale	[117]	[83] are on 5-point Likert scale [79] are on 6-point Likert scale
Undesired behavior detection	[118]	Number of identified bugs	-	Questions about bug identification and solutions
	[109]	Accuracy (percentage of correct answers) and time	-	Task is to determine the correctness of model answers
	[37]	Deviation between human's and model's predictions	-	Model's predictions are buggy and human's predictions should be different.
	[82]	Accuracy (percentage of correct answers)	-	Task is to identify (ir)relevant features
	[119]	Accuracy of answers	-	Task is to detect model biases or discrimination
	[31, 32, 35, 88, 104, 120]	Rating	-	[104]: to judge whether they receive enough information to judge the model process is unfair or not; The other judge the model is unfair or not.
	[121]	Rating	-	Rating on the unfairness of features
Ease of use and others	[34, 78, 109, 111, 114]	Self-designed questionnaire	-	[109, 114] are on 5-point Likert scale [34, 78] are on 7-point Likert scale
	[122]	AVAM and UEQ-S	[123, 124]	Autonomous Vehicle Acceptance Model Questionnaire (AVAM) [123] User Experience Questionnaire-Short (UEQ-S) [124] Both on the 7-point Likert scale
	[81]	Single Ease Question (SEQ)	[125]	On the 7-point Likert scale
	[118]	User Engagement Scale (UES)	[126]	On the 7-point Likert scale
	[127, 128]	System Causability Scale	[129]	On the 5-point Likert scale
	[89]	System Usability Scale	[130]	On the 5-point Likert scale
	[131, 132]	semi-structured interview	-	-

TABLE 4: Measures of usability. The measurement is divided into five categories. The studied core papers using the same measurement are grouped together. The name and the paper reference of the used metrics are listed in the column "Metric" and "Definition Source", respectively. "-" in the column "Definition Source" means that the source is the studied paper. More details about the metrics are given in the last column.

these mental models comes from observing explanations or examples given to the human, who often subconsciously applies the observations in a few examples to the broader understanding of the whole machine learning system. This incredible ability to infer, rationalize, and summarize other intelligent agents' decisions is known as the Theory of Mind (ToM) [155, 156] in psychology. Based on this theory, Bayesian Theory of Mind (BToM) [157] provides a probabilistic framework to predict the inferences that people make about the mental states underlying other agents' actions [158]. Recent work, at the intersection of XAI and robotics, indicates that humans also attribute ToM to artificial agents that they observe or interact with [159, 160]. Guided by these user-centered results, several works at the intersection of XAI and robotics have utilized BToM to create a simulated user and then use the simulated user to generate helpful explanations. Towards this goal, Huang et al. [161] provide a greedy algorithm for selecting explanations that maximize the simulated user's knowledge of the agent's (a self-driving car in their domain) policy; and Lee et al. [162] provide a related approach where the user is modeled as an inverse reinforcement learner. In addition to selecting the most informative explanations, Qian and Unhelkar [163] utilize a variation of the Monte Carlo tree search to generate a computationally tractable approach to identify the most informative sequence of the explanations, based on the assumption that some explanations might be more effective initially. Thus, while some existing works evaluate the effectiveness of the selected explanations through experiments with human users, the community still lacks an understanding of how robust or realistic BToM is compared to a human's cognitive process particularly for XAI. We also advocate for more probabilistic and computational cognitive models to be utilized in XAI designs. To achieve this, we need experts from cross disciplines to address individual user's needs in an XAI system from cognitive, psychological, and computational perspectives. Lastly, we also encourage XAI researchers to develop solutions to explain *AI-enabled systems* – for instance, robots and autonomous vehicles – which require grounded and user-centered solutions.

E.3 Hybrid Teaching

Teaching strategies for the human-to-human setting have been widely studied and many categorizations exist [164, 165, 166]. One way of categorizing these strategies is through the following three concepts: (1) direct teaching, (2) indirect teaching, and (3) hybrid teaching. *Direct teaching* utilizes direct instructions that are teacher-centered, involve clear teaching objectives, and are consistent with classroom organizations. In XAI applications, direct teaching methods

		Other Evaluation Dimensions	
		Positive	Non-positive / Mixed
Trust		[88]: balanced training data, [43]: high model performance [92]: high quality of explanations [104]: high AI literacy [106]: interactivity [86]: model confidence	[88]: user expertise, insignificant [107]: personal characteristics, insignificant [105]: different topic modeling approaches, insignificant [94]: self-referential pronoun "I" in explanations, negative [40]: user technical literacy, insignificant
Understanding	Obj.	[81]: disentanglement of gen. model [40]: interactivity [80]: ExpO regularization of the model	[37, 81]: high dimensionality, negative [79]: contextualization, insignificant [42]: inductive vs. deductive explanations, insignificant [76]: different ML models, insignificant [70]: user expertise, insignificant [72]: instant feedback, insignificant [69]: timing of model errors, mixed
	Sub.	[81]: disentanglement of gen. model [40]: interactivity [139]: user expertise	[64]: model correctness, insignificant [140]: QuickSort, insignificant [66]: test of understanding, negative
Usability		[81]: significant difference in self-reported difficulty dependent on the generative model [91, 106]: interactivity [111]: Parallel Embeddings [104]: high AI literacy [121]: fair features are "current charges" and "criminal history"	[105]: different topic modeling approaches, insignificant [107]: personal characteristics, insignificant for satisfaction [69]: early encounters of system weaknesses lead to lower explanation usage [37]: clear model is *less* useful in debugging [121]: unfair features are "quality of school life" and "education & school behavior", etc.
Human-AI Collaboration Performance		[141]: low model complexity [42, 142]: showing model prediction	[101]: explanations are positive for novices' performance but negative for experts' [74, 86] Showing predictions, insignificant [86]: model confidence

TABLE 5: User study findings when using **other aspects** (other than the presence of explanation) as evaluation dimensions. Effects on measured quantities are divided into "Positive" where explanation information is given, and "Non-positive / Mixed" where negative impact is marked with underlines.

generate explanations by selecting representative examples of an agent's decisions to convey the patterns in its policy [162, 167, 168, 169, 170, 171]. In contrast, *indirect teaching* is student-centered and encourages independent learning. In the XAI perspective, methods utilizing indirect teaching provide users with tools to actively and independently explore an AI system. Although the goal of direct and indirect teaching methods is the same, namely explaining an AI system to human users, the computational problems solved by these methods are different. Direct teaching focuses on providing guidance (using a computational approach) to assist users in building an understanding of a machine, whereas indirect teaching (often through a user interface) enables users to address individual learning preferences and mitigate individual confusion about the AI. To leverage the advantages of the two teaching strategies, *hybrid teaching* has been widely used in human-to-human teaching with an emphasis on interactivity [172, 173, 174]. In XAI-related work, Qian and Unhelkar [163] provide a hybrid teaching framework by introducing an *AI Teacher* to enable guided interactivity between RL-based AI agents and a user. Their results indicate that hybrid teaching reduces the amount of time for a user to understand an agent's policy compared to direct and indirect teaching, and is more subjectively preferred by the participants. Building on this, future XAI systems can consider using hybrid teaching methods that (i) generate direct instructions to provide guidance to users' understanding of an AI system and (ii) provide methods to allow users to interact with the agent or model enabling active learning.

REFERENCES

[1] S. A. Cholewiak, P. Ipeirotis, V. Silva, and A. Kannawadi, "SCHOLARLY: Simple access to Google Scholar authors and citation using Python," 2021.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020.

[3] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, 2008.

[4] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, 2019.

[5] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *CHI*, 2019.

[6] Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow, "Question-driven design process for explainable ai user experiences," *arXiv preprint arXiv:2104.03483*, 2021.

[7] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *IUI*, 2015.

[8] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *VL/HCC*, 2013.

[9] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? the effects of mental model soundness on personalizing an intelligent agent," in *CHI*, 2012.

[10] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[11] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, 2018.

[12] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *CHI*, 2018.

		Evaluation Dimension: Explanations Effect comparison among **different explanations**
Trust		[84]: example-based explanations are positive in trust building [42]: deductive (rule-based) explanations > inductive (example-based) explanations in decision-making tasks, but contrary in proxy tasks [43]: different explanations positively affect different beliefs of trust [108]: proposed explanation interfaces (different visualizations), SCATTER > RANK and SCATTER > TUNER but insignificant [143] HEX-RL (theirs) > LSTM-attention (for RL agents)
Understanding	Obj.	[77]: Cues and Counterfactuals > Saliency (audio data) [78]: Sparse Lin. > COGAM > GAM [65]: MAME > SP-LIME [63]: CLUE > Sensitivity, Human CLUE, Random (for uncertainty) [70]: Natural images > synthetic (activation prediction) [82]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts [19] Anchors > LIME
	Sub.	[144]: local+global explanation > local/global explanation [43]: example-based explanations (normative/comparative) improve the subj. understanding [64] LIME $\geq$ Composite, Prototypes and others [127]: closest and plausible counterfactuals, difference insignificant [144]: local+global explanation > local/global explanation [143] HEX-RL (theirs) > LSTM-attention (for RL agents) [139]: visual > textual explanations
Usability		[78]: sLM $\leq$ COGAM $<$ GAM, insignificant for self-reported cognitive load [79]: contextualizing/exploration improve user's satisfaction, but no significant impact when interacting both factors [34]: diff. expl. (e.g. local expl., counterfactuals,...) [41]: GAM vs. SHAP, pos. for cognitive load [52]: diff. interfaces, pos. for cognitive load [77]: counterfactual+cues > saliency, pos. for helpfulness [116]: DEAML > EFM (feature-level expl.) > PAV ("people also viewed" expl.) for usefulness in RS [69]: Salient video segments > Confidence scores, Component combinations shown for helpfulness [42]: deductive (rule-based) has higher cognitive load than inductive (example-based) in proxy tasks, deductive (rule-based) > inductive (example-based) in helpfulness in decision-making task [127]: closest and plausible counterfactuals, difference insignificant [49]: text explanation > visual explanations in user experience (e.g., satisfaction) [108]: proposed explanation interfaces (different visualizations), SCATTER > RANK and TUNER > SCATTER in satisfaction, RANK > SCATTER and TUNER > SCATTER in usefulness, but all insignificant [82]: Counterfactuals (INN) = (proposed) Baseline Expl. > Concepts in bias detection [119]: AReS (theirs) > AR-LIME [32]: sensitivity- and case-based explanations are rated as least fair when they expose a bias of the model [145]: acceptance of the gender-aware career recommender > gender-debiased [146]: significant preference for equalizing false positives over equalizing accuracy [104]: the amount of information positively relates with perceived fairness [88]: data-centric explanations that indicate balanced training data raise the fairness rating
Human-AI Collaboration Performance		[42]: both deductive (rule-based) explanations and inductive (example-based) explanations are positive, no significant difference

TABLE 6: User study findings when using model **explanations** as evaluation dimensions and comparing different explanation types on measured quantities.

[13] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, 2018.

[14] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *DSAA*, 2018.

[15] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, 2019.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, 2020.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *CSUR*, 2018.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *KDD*, 2016.

[19] ——, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.

[21] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *ICML*, 2018.

[22] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *CSCW*, 2000.

[23] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *KDD*, 2015.

[24] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*, 2017.

[25] S. Wachter, B. Mittelstadt, and C. Russell, "Counter-

factual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, 2017.

[26] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[29] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *CHI*, 2019.

[30] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *CHI*, 2016.

[31] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *CHI*, 2018.

[32] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: an empirical study of how explanations impact fairness judgment," in *IUI*, 2019.

[33] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *ACM FAccT*, 2019.

[34] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *CHI*, 2019.

[35] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *CHI*, 2018.

[36] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *NeurIPS*, 2016.

[37] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *CHI*, 2021.

[38] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *CHI*, 2009.

[39] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[40] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *CHI*, 2019.

[41] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *CHI*, 2020.

[42] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems," in *IUI*, 2020.

[43] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *IUI*, 2019.

[44] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *ICHI*, 2015.

[45] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, 2004.

[46] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *CHI*, 2019.

[47] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *CHI*, 2016.

[48] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-adapted interaction*, 2008.

[49] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "Personalized explanations for hybrid recommender systems," in *IUI*, 2019.

[50] G. Friedrich and M. Zanker, "A taxonomy for generating explanations in recommender systems," *AI Magazine*, 2011.

[51] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "User preferences for hybrid explanations," in *RecSys*, 2017.

[52] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra, "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images," in *IUI*, 2019.

[53] C.-H. Tsai and P. Brusilovsky, "Explaining recommendations in an interactive hybrid social recommender," in *IUI*, 2019.

[54] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, 2015.

[55] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[56] J.-H. Jacobsen, A. W. Smeulders, and E. Oyallon, "i-revnet: Deep invertible networks," in *ICLR*, 2018.

[57] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *NeurIPS*, 2019.

[58] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *NeurIPS*, 2019.

[59] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," *Advances in neural information processing systems*, vol. 33, pp. 5968–5979, 2020.

[60] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor xai:

an ontology-based approach to black-box sequential data classification explanations," in *ACM FAccT*, 2020.
[61] A. Balayn, P. Soilis, C. Lofi, J. Yang, and A. Bozzon, "What do you mean? interpreting image classification with crowdsourced concept extraction and analysis," in *WWW*, 2021.
[62] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *AAAI*, 2021.
[63] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a {clue}: A method for explaining uncertainty estimates," in *ICLR*, 2021.
[64] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *ACL*, 2020.
[65] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," *NeurIPS*, 2020.
[66] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, "I think i get your point, ai! the illusion of explanatory depth in explainable ai," in *IUI*, 2021.
[67] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *ACM FAccT*, 2022.
[68] X. Wang and M. Yin, "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making," in *IUI*, 2021.
[69] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate, "Anchoring bias affects mental model formation and user reliance in explainable ai systems," in *IUI*, 2021.
[70] J. Borowski, R. S. Zimmermann, J. Schepers, R. Geirhos, T. S. A. Wallis, M. Bethge, and W. Brendel, "Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization," in *ICLR*, 2021.
[71] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: a user study," in *IUI*, 2020.
[72] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, "Do explanations make vqa models more predictable to a human?" in *EMNLP*, 2018.
[73] J. Colin, T. Fel, R. Cadene, and T. Serre, "What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods," in *NeurIPS*, 2022.
[74] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky, "Hive: Evaluating the human interpretability of visual explanations," in *ECCV*, 2022.
[75] H. Shen and T.-H. Huang, "How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, no. 1, 2020, pp. 168–172.
[76] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig, "Explain, edit, and understand: Rethinking user study design for evaluating model explanations," in *AAAI*, 2022.
[77] W. Zhang and B. Y. Lim, "Towards relatable explainable ai with the perceptual process," in *CHI*, 2022.
[78] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, "Cogam: measuring and moderating cognitive load in machine learning model explanations," in *CHI*, 2020.
[79] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki, "Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users," in *IUI*, 2022.
[80] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 526–10 536, 2020.
[81] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez, "Evaluating the interpretability of generative models by interactive reconstruction," in *CHI*, 2021.
[82] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf, "Do users benefit from interpretable vision? a user study, baseline, and dataset," in *ICLR*, 2022.
[83] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, "Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems," in *CHI*, 2022.
[84] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan, "Intuitively assessing ml model reliability through example-based explanations and editing model inputs," in *IUI*, 2022.
[85] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your ai: expertise and explanations," in *IUI*, 2019.
[86] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 295–305.
[87] Y. Rong, N. Castner, E. Bozkir, and E. Kasneci, "User trust on an explainable ai-based medical diagnosis support system," *arXiv preprint arXiv:2204.12230*, 2022.
[88] A. I. Anik and A. Bunt, "Data-centric explanations: explaining training data of machine learning systems to promote transparency," in *CHI*, 2021.
[89] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, "Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles," in *CHI*, 2021.
[90] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, 2000.
[91] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, "Building trust in interactive machine learning via user contributed interpretable rules," in *IUI*, 2022.
[92] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and

J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *CHI*, 2019.

[93] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User modeling and user-adapted interaction*, 2012.

[94] M. Liao and S. S. Sundar, "How should ai systems talk to users when collecting their personal information? effects of role framing and self-referencing on human-ai interaction," in *CHI*, 2021.

[95] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, 2000.

[96] J. Ooge, S. Kato, and K. Verbert, "Explaining recommendations in e-learning: Effects on adolescents' trust," in *IUI*, 2022.

[97] I. Benbasat and W. Wang, "Trust in and adoption of online recommendation agents," *Journal of the association for information systems*, 2005.

[98] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, 1992.

[99] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *IUI*, 2019.

[100] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, 2003.

[101] R. Paleja, M. Ghuy, N. Ranawaka Arachchige, R. Jensen, and M. Gombolay, "The utility of explainable ai in ad hoc human-machine teaming," *NeurIPS*, 2021.

[102] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, 2002.

[103] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll, "Exploring and promoting diagnostic transparency and explainability in online symptom checkers," in *CHI*, 2021.

[104] J. Schoeffer, N. Kuehl, and Y. Machowski, ""there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making," *arXiv preprint arXiv:2205.05758*, 2022.

[105] A. Smith-Renner, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater, "Digging into user control: perceptions of adherence and instability in transparent models," in *IUI*, 2020.

[106] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater, "No explainability without accountability: An empirical study of explanations and feedback in interactive ml," in *CHI*, 2020.

[107] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, "To explain or not to explain: the effects of personal characteristics when explaining music recommendations," in *IUI*, 2019.

[108] C.-H. Tsai and P. Brusilovsky, "Beyond the ranked list: User-driven exploration and diversification of social recommendation," in *IUI*, 2018.

[109] D. H. Kim, E. Hoque, and M. Agrawala, "Answering questions about charts and generating visual explanations," in *CHI*, 2020.

[110] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *CHI*, 2021.

[111] D. L. Arendt, N. Nur, Z. Huang, G. Fair, and W. Dou, "Parallel embeddings: a visualization technique for contrasting learned representations," in *IUI*, 2020.

[112] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," in *IUI*, 2019.

[113] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, 1988.

[114] Y. Wang, P. Venkatesh, and B. Y. Lim, "Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation," in *CHI*, 2022.

[115] W. Zhang, M. Dimiccoli, and B. Y. Lim, "Debiased-cam to mitigate image perturbations with faithful visual explanations of machine learning," in *CHI*, 2022.

[116] J. Gao, X. Wang, Y. Wang, and X. Xie, "Explainable recommendation through attentive multi-view learning," in *AAAI*, 2019.

[117] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[118] A. Balayn, N. Rikalo, C. Lofi, J. Yang, and A. Bozzon, "How can explainability methods be used to support bug identification in computer vision models?" in *CHI*, 2022.

[119] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 187–12 198, 2020.

[120] J. Schoeffer and N. Kuehl, "Appropriate fairness perceptions? on the effectiveness of explanations in enabling people to assess the fairness of automated decision systems," in *CSCW*, 2021.

[121] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *WWW*, 2018.

[122] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, "Explain yourself! transparency for positive ux in autonomous driving," in *CHI*, 2021.

[123] C. Hewitt, I. Politis, T. Amanatidis, and A. Sarkar, "Assessing public perception of self-driving cars: the autonomous vehicle acceptance model," in *IUI*, 2019.

[124] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (ueq-s)," *IJIMAI*, 2017.

[125] J. Sauro and J. S. Dumas, "Comparison of three one-question, post-task usability questionnaires," in *CHI*, 2009.

[126] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short

form," *International Journal of Human-Computer Studies*, 2018.

[127] U. Kuhl, A. Artelt, and B. Hammer, "Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting," *arXiv preprint arXiv:2205.05515*, 2022.

[128] ——, "Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning," *arXiv preprint arXiv:2205.03398*, 2022.

[129] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (scs)," *KI-Künstliche Intelligenz*, 2020.

[130] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, 1996.

[131] P. Le Bras, D. A. Robb, T. S. Methven, S. Padilla, and M. J. Chantler, "Improving user confidence in concept maps: Exploring data driven explanations," in *CHI*, 2018.

[132] T. Li, G. Convertino, R. K. Tayi, and S. Kazerooni, "What data should i protect? recommender and planning support for data security analysts," in *IUI*, 2019.

[133] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *ICML*, 2020.

[134] I. Laina, R. Fong, and A. Vedaldi, "Quantifying learnability and describability of visual concepts emerging in representation learning," 2020.

[135] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *NeurIPS*, 2019.

[136] T. Leemann, Y. Rong, S. Kraft, E. Kasneci, and G. Kasneci, "Coherence evaluation of visual concepts with objects and language," in *ICLR2022 WS*, 2022.

[137] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[138] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, 2020.

[139] M. Szymanski, M. Millecamp, and K. Verbert, "Visual, textual or hybrid: the effect of user expertise on different explanations," in *26th International Conference on Intelligent User Interfaces*, 2021, pp. 109–119.

[140] J. Rebanal, J. Combitsis, Y. Tang, and X. Chen, "Xalgo: A design probe of explaining algorithms' internal states via question-answering," in *IUI*, 2021.

[141] V. Lai, H. Liu, and C. Tan, ""why is' chicago'deceptive?" towards building model-driven tutorials for humans," in *CHI*, 2020.

[142] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?" in *AAAI*, 2021.

[143] X. Peng, M. Riedl, and P. Ammanabrolu, "Inherently explainable reinforcement learning in natural language," in *NeurIPS*, 2022.

[144] M. Radensky, D. Downey, K. Lo, Z. Popovic, and D. S. Weld, "Exploring the role of local and global explanations in recommender systems," in *CHI*, 2022.

[145] C. Wang, K. Wang, A. Bian, R. Islam, K. N. Keya, J. Foulds, and S. Pan, "Do humans prefer debiased ai algorithms? a case study in career recommendation," in *IUI*, 2022.

[146] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *ACM FAccT*, 2020.

[147] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring computational user models for agent policy summarization," in *IJCAI*, 2019.

[148] C. S. Hulleman, J. J. Kosovich, K. E. Barron, and D. B. Daniel, "Making connections: Replicating and extending the utility value intervention in the classroom." *Journal of Educational Psychology*, 2017.

[149] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: a systematic review and meta-analysis." *Psychological bulletin*, 2012.

[150] A. Wigfield and J. Cambria, "Expectancy-value theory: Retrospective and prospective," in *The decade ahead: Theoretical perspectives on motivation and achievement*, 2010.

[151] O. Chernikova, N. Heitzmann, A. Opitz, T. Seidel, and F. Fischer, "A theoretical framework for fostering diagnostic competences with simulations in higher education," *Learning to Diagnose with Simulations*, 2022.

[152] N. Heitzman, T. Seidel, A. Opitz, A. Hetmanek, C. Wecker, M. Fischer, S. Ufer, R. Schmidmaier, B. Neuhaus, M. Siebeck *et al.*, "Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education." *Frontline Learning Research*, 2019.

[153] K. Bisra, Q. Liu, J. C. Nesbit, F. Salimi, and P. H. Winne, "Inducing self-explanation: A meta-analysis," *Educational Psychology Review*, 2018.

[154] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized xai: A case study in intelligent tutoring systems," *Artificial Intelligence*, 2021.

[155] G. S. Becker, *The economic approach to human behavior*, 1976.

[156] S. Baron-Cohen, "Precursors to a theory of mind: Understanding attention in others," *Whiten, Andrew (ed.), Natural theories of mind*, 1991.

[157] C. L. Baker, "Bayesian theory of mind: Modeling human reasoning about beliefs, desires, goals, and social relations," Ph.D. dissertation, 2012.

[158] G. Csibra, "Cognitive science: Modelling theory of mind," *Nature Human Behaviour*, 2017.

[159] T. Hellström and S. Bensch, "Understandable robots," *Paladyn, Journal of Behavioral Robotics*, 2018.

[160] S. lai Lee, I. Y. man Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *ICRA*, 2005.

[161] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, 2019.

[162] M. S. Lee, H. Admoni, and R. Simmons, "Machine teaching for human inverse reinforcement learning," *Frontiers in Robotics and AI*, 2021.

[163] P. Qian and V. Unhelkar, "Evaluating the role of in-

teractivity on improving transparency in autonomous agents," in *AAMAS*, 2022.
[164] L. Julien-Schultz, N. Maynes, and C. Dunn, "Managing direct and indirect instruction: A visual model to support lesson planning in pre-service programs," *The International Journal of Learning: Annual Review*, 2010.
[165] K. A. Nguyen, J. Husman, M. A. T. Borrego, P. Shekhar, M. J. Prince, M. DeMonbrun, C. J. Finelli, C. Henderson, and C. K. Waters, "Students' expectations, types of instruction, and instructor strategies predicting student response to active learning," *IJEE*, 2017.
[166] T. Ruutmann and H. Kipper, "Teaching strategies for direct and indirect instruction in teaching engineering," in *2011 14th International Conference on Interactive Collaborative Learning*, 2011.
[167] D. Amir and O. Amir, "Highlights: Summarizing agent behavior to people," in *AAMAS*, 2018.
[168] S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan, "Establishing appropriate trust via critical states," *IROS*, 2018.
[169] O. Amir, F. Doshi-Velez, and D. Sarne, "Summarizing agent strategies," *Autonomous Agents and Multi-Agent Systems*, 2019.
[170] O. Watkins, S. Huang, J. Frost, K. Bhatia, E. Weiner, P. Abbeel, T. Darrell, B. Plummer, K. Saenko, and A. Dragan, "Explaining robot policies," *Applied AI Letters*, 2021.
[171] O. A. Yotam Amitai, ""i don't think so": Summarizing policy disagreements for agent comparison," 2022.
[172] C. P. Fulford and S. Zhang, "Perceptions of interaction: The critical predictor in distance education," *American Journal of Distance Education*, 1993.
[173] B. Muirhead, "Interactivity research studies," *Journal of Educational Technology & Society*, 2001. [Online]. Available: http://www.jstor.org/stable/jeductechsoci.4.3.108
[174] B. Muirhead and C. Juwah, "Interactivity in computer-mediated college and university education: A recent review of the literature," *Educational Technology & Society*, 2004.

CC BY 4.0 DEED

Attribution 4.0 International

Canonical URL :
https://creativecommons.org/licenses/by/4.0/

See the legal code

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any

purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the

public domain or where your use is permitted by an applicable exception or limitation .

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it

summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- Learn more about our work
- **Learn more about CC Licensing**
- Support our work
- Use the license for your own material.
- Licenses List
- Public Domain List

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- More info

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indication of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- Marking guide
- More info

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WIPO Copyright Treaty.

- More info

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- More info

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- More info

CONTACT US

Creative Commons PO Box 1866, Mountain View, CA 94042

info@creativecommons.org

+1-415-429-6753

SUBSCRIBE TO OUR NEWSLETTER

Your e

SUBSCRIBE

SUPPORT OUR WORK

Our work relies on you! Help us keep the Internet free and open.

DONATE NOW

I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise

Yao Rong[1], **Peizhu Qian**[2], **Vaibhav Unhelkar**[2], **Enkelejda Kasneci**[1]

[1]Technical University of Munich, [2] Rice University
{yao.rong, enkelejda.kasneci}@tum.de, {pqian, vaibhav.unhelkar}@rice.edu

Abstract

Effectively explaining decisions of black-box machine learning models is critical to responsible deployment of AI systems that rely on them. Recognizing their importance, the field of explainable AI (XAI) provides several techniques to generate these explanations. Yet, there is relatively little emphasis on the user (the explainee) in this growing body of work and most XAI techniques generate "one-size-fits-all" explanations. To bridge this gap and achieve a step closer towards human-centered XAI, we present I-CEE, a framework that provides **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise. Informed by existing work, I-CEE explains the decisions of image classification models by providing the user with an informative subset of training data (i.e., example images), corresponding local explanations, and model decisions. However, unlike prior work, I-CEE models the *informativeness* of the example images to depend on user expertise, resulting in different examples for different users. We posit that by tailoring the example set to user expertise, I-CEE can better facilitate users' understanding and simulatability of the model. To evaluate our approach, we conduct detailed experiments in both simulation and with human participants ($N = 100$) on multiple datasets. Experiments with simulated users show that I-CEE improves users' ability to accurately predict the model's decisions (simulatability) compared to baselines, providing promising preliminary results. Experiments with human participants demonstrate that our method significantly improves user simulatability accuracy, highlighting the importance of human-centered XAI.

Introduction

As AI systems receive increasingly important roles in our life, human users are challenged to comprehend the decisions made by these systems. To ensure user safety and proper use of AI systems, experts across disciplines have recognized the need for AI transparency (Yang et al. 2017; Ehsan et al. 2021; Russell 2021). Solutions for AI transparency – e.g., techniques for explainable AI (XAI) – are essential as most AI models can be viewed as a "black box," whose decision-making process cannot be easily interpreted or understood by human users. Among the different settings of XAI, our work focuses on explaining image classification tasks (Barredo Arrieta et al. 2020). Existing XAI techniques for image classification widely use attribution explanations, such as GradCAM (Selvaraju et al. 2017), SHAP (Lundberg and Lee 2017) or LIME (Ribeiro, Singh, and Guestrin 2016). While these techniques inform our work, they all miss one key element: human factors, potentially due to the complexity of modeling human users.

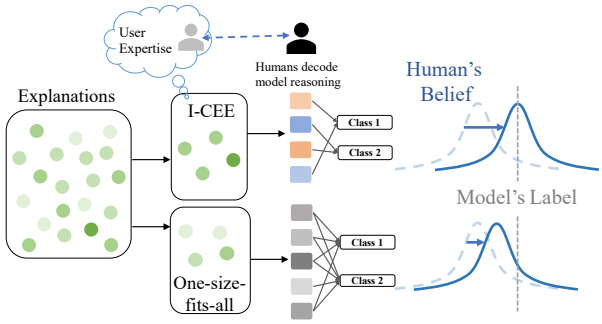


Figure 1: I-CEE tailors the explanation process to each user by considering their expertise. By selecting the most informative explanations based on user expertise, I-CEE can better enhance user simulatability of ML model's decisions.

We advocate that human modeling is critical to XAI research because explainability is inherently centered around humans (Liao and Varshney 2021). A few works focusing on explaining reinforcement learning policies use cognitive science theories to model the human user and generate explanations based on the human model (Baker and Saxe 2011; Huang et al. 2019; Lage et al. 2019b; Qian and Unhelkar 2022). Closer to our focus, the works of Yang, Folke, and Shafto (2022) and Yang et al. (2021) utilize a Bayesian Teaching framework to model human perception and then generate human-centered explanations. One limitation of these works is that all human users are treated the same by the modeling method, presuming that an identical set of explanations will work for *all* users. In contrast, we attempt to generate tailored explanations for each user by modeling their *task-specific expertise*. Our approach to modeling user expertise is informed by human annotator models used in active and imitation learning (Welinder et al. 2010; Beliaev et al. 2022). Similar to these works, our user model aims to capture both the decisions and reasoning process

(expertise in concepts used for image classification) of the human user in the context of a given classification task.

To bridge the research gap that personalization is missing in the explanation process, we propose the framework **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise (I-CEE). Informed by existing XAI methods for image classification, our framework utilizes the *explanation-by-examples* paradigm and provides attribution explanations (local explanations) for a subset of training data. However, in I-CEE, the approach of selecting the example explanations differs and is user-specific. For a given image classification task, I-CEE first discovers a set of m task-relevant concepts. It then models the user's task-specific expertise as a m-dimensional vector, where each entry lies between $[0, 1]$ and represents their expertise in the corresponding concept. Based on this user model, I-CEE finally selects the set of local explanations that can best fill user's knowledge gaps.

As depicted in Figure 1, by selecting the set of local explanations that can best increase the user's task-specific expertise, I-CEE aims to accelerate user's understanding of the decision-making process of the machine learning model. In contrast, most existing work in XAI either selects random or one-size-fits-all local explanations, thereby foregoing the opportunity to accelerate model understanding by providing tailored explanations. The contributions of this work can be summarized as follows:

- We identify the opportunity for tailored explanations for explaining decisions made by image classification models and develop a novel framework named I-CEE that realize this opportunity. This work represents an advancement towards human-centered explanations.
- To evaluate I-CEE, we test the simulatability of explanations generated by our framework on four datasets. Results demonstrate that our framework achieves better simulatability (i.e., users' ability to predict the model's decisions) relative to state-of-the-art XAI baselines[1].
- We evaluate our framework through detailed human-subject studies ($N = 100$). Experimental results indicate that our framework can more effectively help users understand the ML model's decision-making than the state-of-the-art technique Bayesian Teaching (Yang et al. 2021), and is subjectively more preferred by the participants, highlighting the advantages of our framework.

Related Work

Human-centered Explainable AI. Recent surveys indicate a growing activity in XAI research (Doshi-Velez and Kim 2017; Liao and Varshney 2021; Rong et al. 2023). The field recognizes the central role of humans in their explanations, leading to increasing adoption of human-centered evaluations of explanation techniques (Lage et al. 2019a). Besides evaluations, a few techniques have also considered human factors in generating explanations (Lage and Doshi-Velez 2020; Lage et al. 2019b; Huang et al. 2019; Qian and Unhelkar 2022; Yang, Folke, and Shafto 2022). Among these, the most related framework is that of Bayesian Teaching, which focuses on image classification and selects explanations by modeling the users as a Bayesian agent (Yang et al. 2021). However, this work does not model differences between users' reasoning or prior expertise. In contrast, we consider personalized user models to better fit the specific explanation needs of different users. Our design is informed by research in pedagogy and active machine learning.

Pedagogical Theories on Learning from Errors. XAI has been viewed as a teaching process, where the XAI technique serves the role of the teacher and the user that of the student (Qian and Unhelkar 2022). To teach learners effectively, pedagogical research confirms that a teacher needs to assess a learner's prior knowledge and design instructions accordingly (Owens and Tanner 2017; Ambrose et al. 2010). A common indicator of incorrect knowledge is errors, caused by an incorrect association or understanding. To correct the errors, feedback on the correct answers along with explanations have been found to be crucial and most helpful (Metcalfe 2017). These findings in learning sciences have laid the groundwork for our XAI framework, motivating our example selection approach; in particular, I-CEE emphasizes explaining the images on which it estimates the user will make errors. Additionally, as the confidence in an error increases, learning from the error also increases (Butterfield and Metcalfe 2001; Metcalfe and Finn 2011). This is an effect known as the hypercorrection effect. To reflect the hypercorrection effect in our framework, we choose images where the user has low confidence in the correct label (i.e., high confidence in the incorrect label), and argue that using these examples will result in better learning outcomes.

Active Learning. In the context of machine learning (ML), techniques for active learning aim to achieve high model accuracy while minimizing the required labeling effort (Settles 2009; Ren et al. 2021). Active learning is valuable in domains where a limited amount of training data is labeled, and it has been used beyond classification tasks such as in sequence labeling (Settles and Craven 2008) or image semantic segmentation (Sinha, Ebrahimi, and Darrell 2019). While active learning pertains to training machines, we observe that insights from the field are highly relevant for XAI (which seeks to train humans about an AI model). By making this novel connection, we leverage a central component of active learning techniques – *query strategies* – to inform the development and evaluation of I-CEE.

Problem Statement

Consider an ML classifier, denoted as f or the *target model*, trained on dataset $\mathcal{D}$ of image-label pairs $(\mathbf{x}, y)$. The classifier $f : \mathbb{R}^d \rightarrow \{1 : K\}$ maps an input image $\mathbf{x} \in \mathbb{R}^d$ to a label $y \in \{1 : K\}$, i.e., $f(\mathbf{x}) = y$, where K is the number of classes. For a subset of images, the predicted label y may not match the true label y^*. To explain such target models, different feature attribution methods have been proposed that generate local explanations (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). These local explanation assigns each input pixel an importance value, denoted as $\mathbf{e} \in \mathbb{R}^d$, which is usually visualized as a saliency map. In

[1]Code is available at https://github.com/yaorong0921/I-CEE.

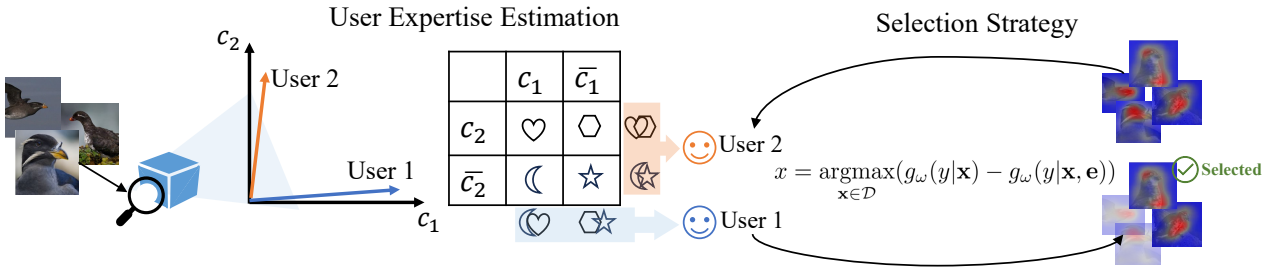


Figure 2: Overview of I-CEE. **Left:** The target model is first projected into a concept space, which is then used to estimate user expertise. Two users are illustrated. User 1 uses the concept c_1 in the reasoning process and can differentiate only two classes (highlighted in blue). Likewise, User 2 is able to distinguish two classes based on c_2 (in orange). **Right:** Based on user models, explanations with images $(\mathbf{x}, \mathbf{e})$ in the training set that maximize Hypercorrection Effect are selected and delivered to the users.

the *explanation-by-example* paradigm, the user is shown a set of images sampled from the training data, its local explanation, and its prediction, i.e., $(\mathbf{x}, \mathbf{e}, y)$. As the user has limited time to understand the model, it is important to select the set of most informative example images.

Within the explanation-by-example paradigm, we consider the problem of selecting the set of most informative example images (and corresponding explanations). Formally, our problem assumes three inputs: the target model f, a data set $\mathcal{D}$ ($|\mathcal{D}| = N$), and a feature attribution method to generate local explanations. Given these inputs, we seek to generate a subset $S \subset \mathcal{D}$ of training data composed of $M \ll N$ images that best facilitate *simulatability*, i.e., help users predict the decisions of the ML model. As the problem objective hinges on a human-centered metric, its successful resolution warrants a human-centered approach.

I-CEE: Image Classification Explanations tailored to User Expertise

We now present our approach to solve this problem: I-CEE, which is composed of two phases (Figure 2). First, our framework models the user by estimating their task-specific expertise (lines 3-4, Algorithm 1). Second, by simulating the user using this model and a query strategy, I-CEE selects informative example images and explanations (lines 5-8).

User Expertise Estimation

The process of a user predicting an ML model's labeling decisions can be viewed as one of image annotation, where the annotators might possess distinct areas of strengths or *expertise* affecting their giving labels (Welinder et al. 2010). For instance, some users find textual patterns to be more recognizable than shapes while others find shapes to be more intuitive. During the annotation process, humans frequently use "concept-based thinking" in reasoning and decision making: identifying similarities among various examples and organizing them systematically based on their resemblances (Yeh et al. 2020; Armstrong, Gleitman, and Gleitman 1983; Tenenbaum 1999). Recognizing these aspects of human reasoning and informed by annotator models proposed in active learning, we model a user by estimating their expertise in applying different task-relevant concepts. We first discover the underlying concepts in the feature space of the target model. Using the discovered concepts, we model a user with a vector representing their ability to utilize each concept when annotating images.

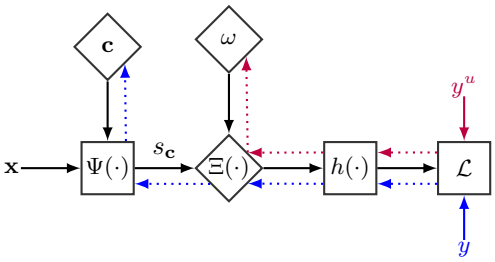


Figure 3: User Modeling: Square nodes are deterministic, while diamond nodes are trainable. Loss back-propagated for concept discovery (Eq. 3) is marked in blue, while that for expertise estimation (Eq. 4) is in red.

Figure 3 provides an overview of the user model. To arrive at the model, I-CEE begins with applying the concept discovery algorithm on the target model (Yeh et al. 2020) that aims to recover m concept $[\mathbf{c}_1, \cdots \mathbf{c}_m]$, such that

$$f(\mathbf{x}) = h(\Psi(\mathbf{x})) = h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}))) \quad (1)$$

where $\Psi(\mathbf{x}) \equiv [\psi(\mathbf{x}^1), \ldots, \psi(\mathbf{x}^T)]$ are T activation vectors, $h(\cdot)$ represents the mapping from the intermediate output of activation vectors to image labels,[2] $s_\mathbf{c}(\cdot)$ is the concept score

$$s_\mathbf{c}(\mathbf{x}) = \langle \psi(\mathbf{x}^i), \mathbf{c}_j \rangle|_{j=1}^m|_{i=1}^T \in \mathbb{R}^{m \cdot T} \quad (2)$$

that estimates the alignment between each concept and activation vector pair, and $\Xi_\theta : \mathbb{R}^{T \cdot m} \rightarrow \mathbb{R}^{T \cdot n}$ is a trainable mapping that converts concept scores back into the activation space. Both the concept vectors and concept scores are unit normalized. For concept discovery (i.e., computing $\mathbf{c}, \theta$), the following cross-entropy loss is minimized:

$$\mathcal{L}_{(\mathbf{c},\theta)} = -\sum_{i=1}^{N} y_i \log(h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}_i)))), \quad (3)$$

[2] Ψ and h can also be viewed as the intermediate and final layers of the image classification neural network, respectively. As h and Ψ are not trained as part of the user model, we do not explicitly denote their parameters (such as weights and biases) in our notation.

Algorithm 1: I-CEE

Input: Target model $f(\cdot)$, data $\mathcal{D}$, user annotation y^u.
Output: A set of example images and explanations $\mathcal{S}$.
Discover concepts by solving Eq. 3.
Estimate user expertise by solving Eq. 4.
for $\mathbf{x} \in \mathcal{D}$ **do**
    Calculate Hypercorrection Effect for $\mathbf{x}$ using Eq. 5.
end for
Return top-K image samples.

where y is the prediction from the target model $f(\cdot)$.

After completing concept discovery (which is a one-time process), the expertise estimation for each user takes place within the concept space. We freeze all model parameters ($\Psi(\cdot)$, $s_{\mathbf{c}}(\cdot)$, $\Xi_\theta(\cdot)$ and $h(\cdot)$) trained using Eq. 3 to learn an expertise vector $\omega \in \mathbb{R}^m$ for each user. The variations among users are manifested through different values of ω, as their diverse domain knowledge influences the way they utilize concepts to arrive at predictions. Concretely, we ask users to annotate images and use ω to simulate their predictions. The expertise vector ω for a user is learned by minimizing the following cross-entropy loss:

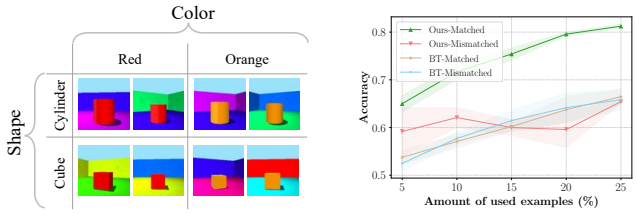
$$\mathcal{L}_\omega = -\sum_{i=1}^{N} y_i^u \log(h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\mathbf{x}_i))), \tag{4}$$

where y^u denotes annotated labels collected from the user. Once ω is learned, we obtain a user model denoted as $g_\omega(\cdot) = h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\cdot))$. If $\omega_1 \approx \omega_2$, it implies that these two users (Users 1 and 2) have very similar "reasoning process" as the utilization of concepts is very similar. Likewise, if $\omega \approx \mathbf{1}_m$, this user employs a very similar reasoning mechanism as the target model f.

Selection Strategy

Our goal is to select a set of informative examples that can most improve the user's simulatability. To estimate the informativeness of the examples, we employ the concept of the hypercorrection effect in educational psychology. As the human needs to learn how the model makes the decision, the model's prediction is viewed as the "correct" answer whereas the human's disagreed initial belief is the "error". Feedback on the correct answer along with explanations has been found to be crucial and most helpful in learning new knowledge (Metcalfe 2017). As the confidence in an error increases, i.e., the confidence in the correct answer decreases, learning from this error example is more effective (Butterfield and Metcalfe 2001; Metcalfe and Finn 2011). To reflect the hypercorrection effect in I-CEE, we choose images where the user has lower confidence in the model's predicted label after knowing the model's reasoning and argue that using these examples will lead to higher learning outcomes. Concretely, I-CEE aims to identify a set of examples $\mathcal{S} \subseteq \mathcal{D}$ which consists of samples with the top maximal Hypercorrection Effect:

$$x = \underset{\mathbf{x} \in \mathcal{D}}{\operatorname{argmax}} \underbrace{(g_\omega(y|\mathbf{x}) - g_\omega(y|\mathbf{x}, \mathbf{e}))}_{\text{Hypercorrection Effect of } \mathbf{e}}, \tag{5}$$



(a) Synthetic dataset (b) (Mis-)Matched expertise

Figure 4: **(a)**: Overview of four classes in the synthetic dataset. **(b)**: User simulatability accuracy when trained with examples that match/mismatch with the user expertise.

where $g_\omega(\cdot)$ represents the user model, $\mathcal{D}$ denotes the training dataset, and $\mathbf{e}$ and y are the local explanation and machine prediction corresponding to the image $\mathbf{x}$.

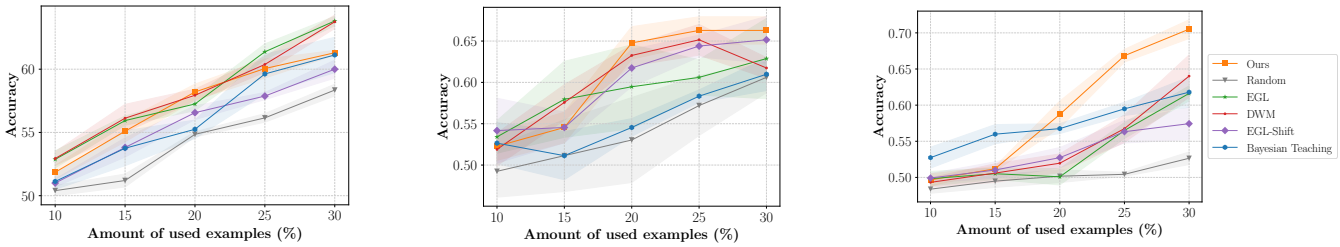
Experiments with Simulated Users

Before conducting a user study, we first evaluate our approach through extensive experiments with simulated users on one synthetic and three realistic image classification tasks. To facilitate reproducibility, Appendix includes more details about the experimental setup.

Synthetic Dataset. We construct a synthetic dataset[3] to validate the design of our proposed method in simulation. This dataset contains four classes and each class is described with two concepts, color and shape, illustrated in Figure 4a. For instance, if a user uses colors to distinguish between different classes (i.e., they have more expertise in using "colors" then "shapes"), then to this user, the red cylinders and red cubes belong to the same class, which differs from the orange ones. Likewise, for a user who has high expertise in using shapes, the cylinders and the cubes are distinguishable for this user regardless of their colors. The other visual features such as angles or background colors are randomly sampled as they are not essential in this decision-making process. For each class, we generate 300 images (80% for training and 20% for testing). We use a ResNet-18 (He et al. 2016) as our classification model and use GradCAM (Selvaraju et al. 2017) for generating explanations. Given their annotation behavior, a simulated behavior is modeled using Eqs. 3-4, i.e., identical to the modeling approach of I-CEE.

Realistic Datasets. We also benchmark I-CEE on three real-world datasets: CIFAR-100 (Krizhevsky, Hinton et al. 2009), CUB-200-2011 (Wah et al. 2011) and German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al. 2012). We construct a simulated user from pre-defined annotations on each dataset who behaves differently from the target model. In particular, for each dataset, our simulated user can distinguish only two classes out of four similar classes. All methods are evaluated based on this user. For instance, on CUB-200-2011, the simulated user labels both Crested and Least Auklet as the same class (Crested Auklet), and Parakeet and Rhinoceros Auklet as the same class

[3]This dataset is based on 3d-shapes (Kim and Mnih 2018).



(a) CIFAR-100 (b) CUB-200-2011 (c) GTSRB

Figure 5: Comparison with baseline algorithms with simulated users on three datasets. The ratio of used examples p (in percentage) is plotted on the x-axis and simulatability accuracy is on the y-axis. (Results averaged over 5 runs.)

(Parakeet Auklet). We use the original training-test splits on these datasets and, similar to the procedure in the synthetic dataset, we use ResNet-50 (He et al. 2016) for classification training and GradCAM for computing explanations.

Baseline Methods

We evaluate I-CEE against a recent human-centered XAI approach: Bayesian Teaching (BT) (Yang et al. 2021). BT simulates a user's behavior (i.e., their prediction of an image class) by deploying a ResNet-50-PLDA (probabilistic linear discriminate analysis (Ioffe 2006)) model. By assuming users perform Bayesian reasoning, it selects example images and explanations to better align user's beliefs to the target model. I-CEE and BT differ in their approaches to both user modeling and example selection.

To evaluate the example selection alone, we also benchmark against query strategies derived from active learning (AL). Unlike traditional AL, in our application of AL query strategies to XAI, the simulated user is the learner and the target model is the annotator. We use Expected Gradient Length (EGL) (Settles, Craven, and Ray 2007), Density-Weighted Method (DWM) (Settles, Craven, and Friedland 2008) as well as a random sampling strategy as baselines. EGL, in the context of this paper, selects samples (x, e) that result in the greatest change to the current model if the annotated label is known. The "change" imparted to the model from the queried samples is measured by the gradient of the objective function with respect to the model parameters. However, the instances chosen by EGL might be outliers that cause significant gradient changes. To alleviate this issue, Settles, Craven, and Friedland (2008) proposes to integrate a density-weighting technique with the query strategy such as EGL. Specifically, each sample is weighted with its average similarity to all other instances in the input dataset. In this work, we extend EGL with the belief shift in the calculated EGL when considering e in the input (denoted as EGL-Shift). Specifically, we compute the difference between EGL of (x, e) and x. With EGL-Shift, we aim to alleviate the influence of an image itself on the training gradient but emphasize the impact of explanations.

Evaluation Metric

To evaluate our method, we use simulatability, which is commonly used as a proxy for testing a user's understanding of the model's decision-making process (Hase and Bansal 2020; Arora et al. 2022; Hase et al. 2020). Simulatability is measured as "to what extent can a user successfully predict a model's prediction." This metric can be used in both simulation experiments and human user studies.

We follow the experimental settings proposed in (Yeh et al. 2018; Koh and Liang 2017) to study the influence of selected examples. Specifically, each method provides an ordered set of example images $\mathcal{S}$, where the ranking is decided by the *informativeness* defined in the respective method. We denote the ratio between number of example images $|\mathcal{S}|$ and the size of training data $\mathcal{D}$ as $p = |\mathcal{S}|/|\mathcal{D}|$. The simulated user is retrained using these example images $\mathbf{x}$ and their corresponding labels $y = f(\mathbf{x})$, where recall that f is the target model. Given the retrained user model g'_ω, we compute the user's accuracy of predicting the model's predictions on the test set, i.e., the simulatability of the user:

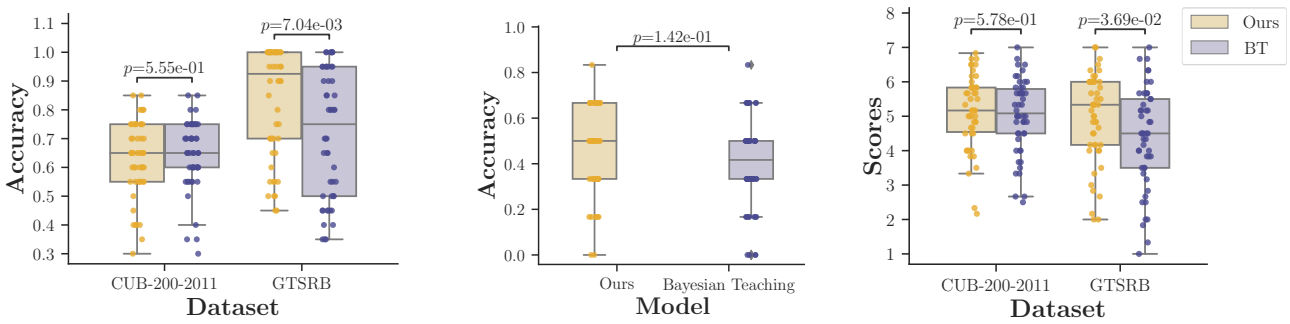
$$\text{Acc} = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{1}(y_i = g'_\omega(\mathbf{x_i})), \tag{6}$$

where N_t is the number of samples in the test set.

Experimental Results

Ablation Study. To validate our model design of $g(\cdot)$, we study (1) whether ω can faithfully reflect the user expertise and (2) the advantages of tailored explanations according to the user expertise. We simulate two users on the synthetic dataset: User 1 only uses color in classification while User 2 only uses shape. We deduce annotations for each user based on attributes for each class (Figure 4a).

After estimating each user, we investigate their expertise vector: ω_1 and ω_2 ($\omega_i \in \mathbb{R}^8$). Each entry in ω_i represents the expertise of the user in one specific concept. The top four largest entries in ω_1 and ω_2 are complementary, corresponding to the fact that each user has the opposite expertise (i.e., each user uses different concepts in the decision-making). To validate the efficacy of the user model via expertise, we run an experiment where we train User 1 using a set of examples specifically chosen based on the User 1 model ("Matched"), against a set of examples chosen for User 2 ("Mismatched"). As demonstrated in Figure 4b, we observe that the simulated user achieves high simulatability accuracy when they receive examples selected according to their expertise ("Ours Matched"). However, if selecting examples



(a) User Accuracy (all predictions) (b) User Accuracy (inaccurate pred., CUB) (c) User evaluation score.

Figure 6: Results of experiments with human users ($N = 100$) comparing I-CEE with the baseline Bayesian Teaching (BT). (a) Simulatability accuracy on all predictions, (b) Simulatability accuracy on images where the target model made inaccurate predictions in the CUB-200-2011 dataset, (c) User's subjective perception of model explanations.

that do not maximize the Hypercorrection Effect tailored to the particular user ("Ours Mismatched"), the simulatability accuracy is low, indicating that such examples fail to provide substantial insights into the target model. Additionally, we compare our user simulation model to that of Bayesian Teaching. We observe little differences between the matched and mismatched settings using the BT framework, suggesting that BT might not be able to accurately simulate the different behaviors of various users. Consequently, it cannot provide examples that effectively improve user simulatability (less performance improvement compared to ours).

Comparison. We compare I-CEE with baselines on three real-world datasets in Figure 5. Evaluation in user prediction accuracy is conducted at $p = [10, 15, 20, 25, 30]\%$. On CIFAR-100, our method always outperforms BT and EGL-Shift but is inferior to EGL and DWM. A potential reason for this result is that the explanation of CIFAR-100 is vague due to the low resolution of images. In this case, Hypercorrection Effect cannot be well captured since explanations are noisy. On CUB-200-2011 and GTSRB, our method outperforms other baselines at most of the percentages. For instance, on CUB our method achieves the best performance after 20%. Note that 20% of the train data consists of 24 images. This is a reasonable number of samples that can be efficiently studied by human users, which we will show in the next section. On GTSRB, we observe an evident performance gap between our method and the competitive baseline BT. A possible explanation for this can be attributed to the architecture of the user model: our model simulates the user via learning ω in the concept space without weakening the capability of the final classifier. On the contrary, BT relies on a PLDA layer to classify images, which can result in suboptimal performance when the latent features of images are highly similar, such as in traffic signs. This is not desirable because humans are good at distilling critical concepts and filtering out similar but irrelevant visual features. With more precise user modeling, our method demonstrates the capability of offering informative learning samples in most of the cases within the simulation experiments.

Experiments with Human Users

We conduct a human user study using the CUB-200-2011 and GTSRB datasets following the same settings as in the simulation experiments. We choose these two datasets as they are more challenging and the images are in higher resolution. We use Bayesian Teaching (Yang et al. 2021) as a baseline since it is the most state-of-art and closest to our focus. Users are first asked to study two classes (among which there are actually four classes) and write down the features used to distinguish between these classes. This step is to let the user think as the pre-defined simulated user, to whom we have tailored model explanations. Then, 20 model explanations selected by our method (experimental group) or Bayesian Teaching (control group) for users are shown, and we ask them to write down the features they use to determine the model prediction. During the evaluation section, participants first receive a test with 15 questions to predict the model's label (images used here are sampled from the test set and include all four classes evenly). We refer to this section as "objective understanding". Then, participants rate their perceived understanding on seven questions on the 7-Likert scale, which we refer to as "subjective understanding". In the user study, we aim to study the following research questions:

- **R1**: Our framework selects informative samples that can increase human understanding of the model.
- **R2**: Human understanding of the model is affected by task domains.

Participants. We recruited 100 participants (average age is 28.8 ± 8.6, 49 females, 50 males, and 1 undefined) using a research platform Prolific[4], and randomly assigned them to one of the two conditions (50 participants/condition). 51 participants have prior experience with AI from using Alexa, Siri, ChatGPT, or from ML-related courses. All participants passed the attention check during the user study. The study protocol has been approved by the Technical University of Munich IRB. At the beginning of the experiment session,

[4]https://www.prolific.co/

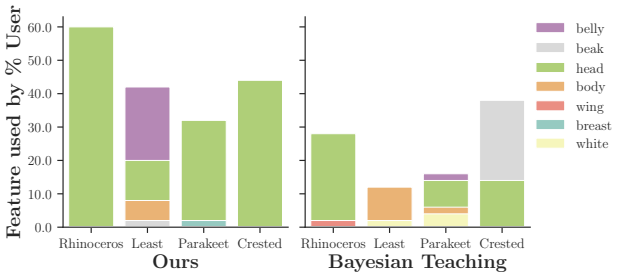


Figure 7: Illustration of features used by human users for distinguishing each class on CUB-200-2011.

we collected informed consent through Prolific. Each participant was compensated with a payment of £4.50 for participation in the user study (within 30 minutes).

Results

Analysis on R1. The results of the simulatability accuracy in each condition on each dataset are shown in Figure 6a. On GTSRB, we observe a statistically significant improvement in using our framework on user simulatability accuracy by 11.5% ($p = 0.007$). On the CUB dataset, we see that users from two conditions achieve similar user prediction accuracy and no significant effect is observed. However, if we inspect the test samples where the target model makes inaccurate predictions (wrong classification) (6 out of 15 images in the test are wrongly predicted), our method demonstrates superior performance compared to BT. Users from the experimental condition achieve an accuracy of 46.3%, whereas users from the control condition achieve 40.3%, as plotted in Figure 6b. These results indicate that users exhibit improved capability in simulating inaccurate predictions from the target model using our method, which is a more challenging task. Additional evidence of the enhancement achieved through our model can be found in Figure 7. We count the words of the features that users think the model uses to distinguish four different classes. When using our framework, the users tend to agree on the same feature (body part of the bird) for each class. For instance, about 68% of the users use "Head" to distinguish Rhinoceros, and about 20% of the users think highly of "Belly" for Least Auklet. Nevertheless, it is more difficult for users in Bayesian Teaching to come to an agreement, for example, for Least Auklet, only around 10% of the participants use "Body" as a feature while other users give diverse descriptions. These results highlight the advantage of the method in improving user understanding of the given target model.

As shown in Figure 6c, the improvement in subjective understanding (rating scores) is not significant on CUB (average rating score is 5.14 in our method and 5.02 in BT). However, we observe that on GTSRB our method surpasses BT significantly with $p = 0.037$. The reason for significant improvement in GTSRB is that our method selects explanations bringing knowledge for distinguishing four classes. But BT chooses examples that reflect important features only for two classes, which hinders users from understanding how the model makes predictions for the other classes.

Analysis on R2. The quantitative result shows that the task domain (dataset) affects the user's objective understanding. However, different tasks influence less subjective understanding, e.g., no significant difference between two datasets when using our method as illustrated in Figure 6c. At the end of the user study, we asked participants for feedback on comparing the perceived helpfulness of model explanations in two datasets. While most of the users in both conditions find the explanations useful, seven users in the experimental condition and fourteen users in the control condition find the explanations on bird species are more helpful than the explanations on road signs. One reason causing this uncertainty in the road sign images is that the salient area is always a circle that covers the road sign, which seems to "be the only one characteristic" for different classes.

Conclusion

We present a human-centered XAI framework, I-CEE, that provides explanations of image classification ML models that are tailored to user expertise. Our framework first discovers task-relevant concepts, uses these concepts to arrive at expertise-based user models, and then selects examples and explanations that help the users to learn the missing concepts so they can accurately predict the machine's image classification decisions. We evaluate our approach through simulation experiments on four datasets, and report on a detailed human-subject study ($N = 100$). In these experiments, we observe that I-CEE outperforms prior art, shows the promise of human-centered XAI, and motivates future research direction for the design of XAI systems.

Limitations and Future Work. Future investigation of our framework can consider the following avenues. First, more complex models of expertise estimation should be studied. In this work, we simulate user expertise by employing the concept-based reasoning approach for image classification proposed in (Yeh et al. 2020). An alternative approach involves utilizing Large Language Models to simulate multiple humans in textual format (Argyle et al. 2023; Aher, Arriaga, and Kalai 2023). Second, the current framework does not consider the sample complexity associated with user expertise estimation. Future work should investigate methods that estimate user expertise with a small number of real-user annotations. Third, we encourage replication of our work to be tested with different datasets, as the power of explanations is dependent on the task domain. Future work should evaluate on datasets that include a more diverse pool of examples, as suggested by some of the participants.

Implications for XAI Systems. This study highlights the importance of personalized XAI, within the explanation-by-example paradigm for image classification. Future work should investigate the potential of personalized XAI in other contexts. We argue that user modeling is essential to provide explanations that target user-specific misunderstanding or confusion. Future XAI systems should leverage and address individual users' preferences and confusion. This involves the development of human-in-the-loop systems, allowing users to actively participate in the process of generating explanations.

Ethical Statement

In this work, we attempt to put human users at the center of XAI design, with the aim of creating AI systems that can be interpreted by non-expert end users. To safeguard user privacy and user rights, we have received approval from University IRB. We believe that only when AI becomes more accessible, acceptable, and usable, can we realize its full potential to empower the world around us.

References

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 337–371. PMLR.

Ambrose, S. A.; DiPietro, M.; Bridges, M. W.; Norman, M. K.; and Lovett, M. C. 2010. *How Does Students' Prior Knowledge Affect Their Learning*, chapter 1, 10–39. John Wiley & Sons.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.

Armstrong, S. L.; Gleitman, L. R.; and Gleitman, H. 1983. What some concepts might not be. *Cognition*, 13(3): 263–308.

Arora, S.; Pruthi, D.; Sadeh, N.; Cohen, W. W.; Lipton, Z. C.; and Neubig, G. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5277–5285.

Baker, C.; and Saxe, R. 2011. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Beliaev, M.; Shih, A.; Ermon, S.; Sadigh, D.; and Pedarsani, R. 2022. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, 1732–1748. PMLR.

Butterfield, B.; and Metcalfe, J. 2001. Errors Committed with High Confidence Are Hypercorrected. *Journal of experimental psychology. Learning, memory, and cognition*, 27: 1491–4.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Hase, P.; and Bansal, M. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, S.; Held, D.; Abbeel, P.; and Dragan, A. 2019. Enabling Robots to Communicate their Objectives. *Autonomous Robots*, 43.

Ioffe, S. 2006. Probabilistic linear discriminant analysis. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, 531–542. Springer.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International Conference on Machine Learning*, 2649–2658. PMLR.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2019a. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 59–67.

Lage, I.; and Doshi-Velez, F. 2020. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*.

Lage, I.; Lifschitz, D.; Doshi velez, F.; and Amir, O. 2019b. Exploring Computational User Models for Agent Policy Summarization. *IJCAI : proceedings of the conference*, 28: 1401–1407.

Liao, Q. V.; and Varshney, K. R. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Metcalfe, J. 2017. Learning from Errors. *Annual Review of Psychology*, 68(1): 465–489.

Metcalfe, J.; and Finn, B. 2011. People's Hypercorrection of High-Confidence Errors: Did They Know It All Along? *Journal of experimental psychology. Learning, memory, and cognition*, 37: 437–48.

Owens, M.; and Tanner, K. 2017. Teaching as Brain Changing: Exploring Connections between Neuroscience and Innovative Teaching. *Cell Biology Education*, 16: fe2.

Qian, P.; and Unhelkar, V. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous

Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1083–1091.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rong, Y.; Leemann, T.; Nguyen, T.-T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; and Kasneci, E. 2023. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Russell, S. 2021. Human-compatible artificial intelligence. *Human-like machine intelligence*, 3–23.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Settles, B. 2009. Active learning literature survey.

Settles, B.; and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, 1070–1079.

Settles, B.; Craven, M.; and Friedland, L. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Settles, B.; Craven, M.; and Ray, S. 2007. Multiple-instance active learning. *Advances in neural information processing systems*, 20.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5972–5981.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332.

Tenenbaum, J. B. 1999. *A Bayesian framework for concept learning*. Ph.D. thesis, Massachusetts Institute of Technology.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23.

Yang, S. C.-H.; Folke, N. E. T.; and Shafto, P. 2022. A psychological theory of explainability. In *International Conference on Machine Learning*, 25007–25021. PMLR.

Yang, S. C.-H.; Vong, W. K.; Sojitra, R. B.; Folke, T.; and Shafto, P. 2021. Mitigating belief projection in explainable artificial intelligence via Bayesian teaching. *Scientific reports*, 11(1): 9863.

Yang, X. J.; Unhelkar, V. V.; Li, K.; and Shah, J. A. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 408–416.

Yeh, C.-K.; Kim, B.; Arik, S.; Li, C.-L.; Pfister, T.; and Ravikumar, P. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33: 20554–20565.

Yeh, C.-K.; Kim, J.; Yen, I. E.-H.; and Ravikumar, P. K. 2018. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31.

Association for the Advancement of Artificial Intelligence
1900 Embarcadero Road, Suite 101, Palo Alto, California 94303
Palo Alto, California 94303 USA

AAAI COPYRIGHT FORM

Title of Article/Paper: I-CEE: Tailoring Explanations of Image Classifications Models to User Expertise

Publication in Which Article/Paper Is to Appear: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24)

Author's Name(s): Yao Rong, Peizhu Qian, Vaibhav Unhelkar, Enkelejda Kasneci

Please type or print your name(s) as you wish it (them) to appear in print

PART A – COPYRIGHT TRANSFER FORM

The undersigned, desiring to publish the above article/paper in a publication of the Association for the Advancement of Artificial Intelligence, (AAAI), hereby transfer their copyrights in the above article/paper to the Association for the Advancement of Artificial Intelligence (AAAI), in order to deal with future requests for reprints, translations, anthologies, reproductions, excerpts, and other publications.

This grant will include, without limitation, the entire copyright in the article/paper in all countries of the world, including all renewals, extensions, and reversions thereof, whether such rights current exist or hereafter come into effect, and also the exclusive right to create electronic versions of the article/paper, to the extent that such right is not subsumed under copyright.

The undersigned warrants that they are the sole author and owner of the copyright in the above article/paper, except for those portions shown to be in quotations; that the article/paper is original throughout; and that the undersigned right to make the grants set forth above is complete and unencumbered.

If anyone brings any claim or action alleging facts that, if true, constitute a breach of any of the foregoing warranties, the undersigned will hold harmless and indemnify AAAI, their grantees, their licensees, and their distributors against any liability, whether under judgment, decree, or compromise, and any legal fees and expenses arising out of that claim or actions, and the undersigned will cooperate fully in any defense AAAI may make to such claim or action. Moreover, the undersigned agrees to cooperate in any claim or other action seeking to protect or enforce any right the undersigned has granted to AAAI in the article/paper. If any such claim or action fails because of facts that constitute a breach of any of the foregoing warranties, the undersigned agrees to reimburse whomever brings such claim or action for expenses and attorneys' fees incurred therein.

Returned Rights

In return for these rights, AAAI hereby grants to the above author(s), and the employer(s) for whom the work was performed, royalty-free permission to:

1. Retain all proprietary rights other than copyright (such as patent rights).
2. Personal reuse of all or portions of the above article/paper in other works of their own authorship. This does not include granting third-party requests for reprinting, republishing, or other types of reuse. AAAI must handle all such third-party requests.
3. Reproduce, or have reproduced, the above article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. The foregoing right shall not permit the posting of the article/paper in electronic or digital form on any computer network, except by the author or the author's employer, and then only on the author's or the employer's own web page or ftp site. Such web page or ftp site, in addition to the aforementioned requirements of this Paragraph, shall not post other AAAI copyrighted materials not of the author's or the employer's creation (including tables of contents with links to other papers) without AAAI's written permission.
4. Make limited distribution of all or portions of the above article/paper prior to publication.
5. In the case of work performed under a U.S. Government contract or grant, AAAI recognized that the U.S. Government has royalty-free permission to reproduce all or portions of the above Work, and to authorize others to do so, for official U.S. Government purposes only, if the contract or grant so requires.

In the event the above article/paper is not accepted and published by AAAI, or is withdrawn by the author(s) before acceptance by AAAI, this agreement becomes null and void.

(1) [redacted]

Author/Authorized Agent for Joint Author's Signature

12.17.2023

Date (MM/DD/YYYY)

Employer for whom work was performed

Title (if not author)

(For jointly authored Works, all joint authors should sign unless one of the authors has been duly authorized to act as agent for the others.)