TUM SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY
TECHNISCHE UNIVERSITÄT MÜNCHEN

Cross-Resolution Face Verification

Martin Knoche

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

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

1. Prof. Dr. Tobias Vogl

Prüfer der Dissertation:

1. Prof. Dr.-Ing. habil. Gerhard Rigoll
2. Prof. Dr.-Ing. Klaus Diepold

Die Dissertation wurde am 24.06.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 22.11.2024 angenommen.

Abstract

Face Recognition (FR) technology has rapidly evolved into a critical tool in various sectors, utilizing sophisticated algorithms to identify or verify a person from a digital image or video frame against a database, thereby enhancing robustness, explainability, and security is crucial. This dissertation delves into the intricacies of FR systems, mainly focusing on the Face Verification (FV) task with cross-resolution images.

The impact of image resolution on Face Verification (FV) is meticulously analyzed and extended to examine facial feature distances within FV models, revealing that the resolution of facial images profoundly affects the performance of state-of-the-art models. Several strategies to bolster the robustness of FR systems against resolution variations are developed, marking a pivotal advancement: Resolution Augmentation Training (RAT), Contrastive Loss Training (CLT), Multi-Branch Contrastive Loss Training (MB-CLT), and Octuplet Loss Training (OLT). These methods significantly enhanced FR models' performance across various benchmark datasets through their unique mechanisms. Notably, the OLT method emerged as the most effective, demonstrating its adaptability to different network architectures. Furthermore, creating a new benchmark dataset, Cross-Quality Labeled Faces in the Wild (XQLFW), addresses the need for a more accurate and standardized evaluation of FV with images of different resolutions. By synthetically deteriorating images from the Labeled Faces in the Wild (LFW) dataset, XQLFW provides a comprehensive evaluation platform for FV models, enabling a more in-depth analysis of their performance on cross-resolution images. In addition to technical advancements, exploring explainability within FR systems yields the development of model-agnostic methods, namely the Confidence Score (C-Score) and the Explanation Map (X-Map). These methods provide meaningful insights into the decision-making processes of FV models, facilitating a more interpretable and transparent evaluation. Lastly, investigating human performance in challenging edge cases for FR systems highlights the potential to synergize human intuition with algorithmic precision. By employing a fusion strategy based on confidence scores, the overall accuracy of FR systems is improved on several benchmarks, emphasizing the value of human-machine collaboration in enhancing FV tasks.

In conclusion, this body of work substantially contributes to the comprehension and enhancement of FR technology, particularly on cross-resolution images. It introduces innovative insights and methodologies that extend the limits of possibility in the field of biometric authentication.

Preface

This dissertation focuses on face verification in unconstrained environments, particularly emphasizing the challenges posed by image resolution. It is the outcome of research conducted during my tenure as a doctoral candidate at the Chair of Human-Machine Communication of the Technical University of Munich from 2018 to 2024. This work aims to integrate the findings into a unified framework, contextualizing, extending and elucidating them for a non-specialist audience. It articulates this collective context and lays the foundational principles in the first two chapters. The subsequent chapters delve into the primary results of the first-authored papers listed on the following pages. A conclusion is drawn in the final chapter, summarizing the essential findings and outlining potential future research directions.

In embarking on this academic journey, I am humbled to present this dissertation as the culmination of years of research, exploration, and dedicated effort. I am immensely thankful for the support I have received along the way. First and foremost, I want to express my gratitude to my advisor, Prof. Gerhard Rigoll, for giving me the opportunity to work on this topic and for his guidance that has been instrumental in shaping the direction of this research. His supervision has been pivotal in molding my academic growth and instilling the pursuit of excellence in me.

I would also like to extend my gratitude to my colleagues at the Chair of Human-Machine Communication for their professional support and the wonderful times we shared. Special thanks to Stefan Hörmann for the enriching technical discussions and the invaluable insights you have provided me. I am profoundly grateful to my parents, Sabine and Richard, who paved the way for this dissertation. Heartfelt thanks go to my friends for always being ready to listen and for their constant encouragement in keeping me motivated. Finally, I am incredibly grateful to my significant other, Julia, for her constant support, understanding, and everlasting patience during the challenging phases of this endeavor. Her presence has provided me with solace and encouragement. Thank you for always being there for me.

Munich, February 3, 2025 *Martin Knoche*

List of Publications and Software Contributions

First-Authored

This dissertation is based on the following publications and software contributions:

- **Martin Knoche**, Stefan Hörmann, and Gerhard Rigoll. "Susceptibility to Image Resolution in Face Recognition and Training Strategies to Enhance Robustness". In *2022 Leibniz Transactions on Embedded Systems*, 8(1)., (pp. 01:1–01:20). [1†] (Chapter 3 and Chapter 4)
  Evaluation protocols:
  `https://github.com/Martlgap/btm-stm`
- **Martin Knoche**, Mohamed Elkadeem, Stefan Hörmann, and Gerhard Rigoll. "Octuplet Loss: Make Face Recognition Robust to Image Resolution". In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1–8). IEEE. [2†] (Chapter 4)
  Code and fine-tuned models:
  `https://github.com/Martlgap/octuplet-loss`
- **Martin Knoche**, Stefan Hörmann, and Gerhard Rigoll. "Cross-Quality LFW: A Database for Analyzing Cross-Resolution Image Face Recognition in Unconstrained Environments". In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1–5). IEEE. [3†] (Chapter 5)
  Code and dataset:
  `https://martlgap.github.io/xqlfw`
  `https://github.com/Martlgap/xqlfw`
- **Martin Knoche**, Torben Teepe, Stefan Hörmann, and Gerhard Rigoll. "Explainable Model-Agnostic Similarity and Confidence in Face Verification". In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshop (WACV-W)* (pp. 711–718). IEEE/CVF. [4†] (Chapter 6)
  Code and the website:
  `https://explainable-face-verification.ey.r.appspot.com`
  `https://github.com/Martlgap/website-x-face-verification`
  `https://github.com/Martlgap/x-face-verification`

- **Martin Knoche**, and Gerhard Rigoll. “Tackling Face Verification Edge Cases: In-Depth Analysis and Human-Machine Fusion Approach”. In *2023 IEEE 18th International Conference on Machine Vision and Applications (MVA)* (pp. 1–5). IEEE .[5†] (Chapter 7)
  Code:
  `https://github.com/Martlgap/human-machine-fusion`
  `https://github.com/Martlgap/human-machine-fusion-survey`

Ideas, text, figures, and experiments originate in majority from the first author. All other authors had important advisory roles, helped with running experiments, supported in programming and assisted in proofreading.

Co-Authored

The author has furthermore contributed to the following publications:

- Stefan Hörmann, Zhenxiang Cao, **Martin Knoche**, Fabian Herzog, and Gerhard Rigoll. “Face Aggregation Network For Video Face Recognition” In *2021 International Conference on Image Processing (ICIP)* (pp. 2973–2977). IEEE. [6†]
- Stefan Hörmann, Zhibing Xia, **Martin Knoche**, and Gerhard Rigoll. “A Coarse-to-Fine Dual Attention Network for Blind Face Completion”. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 01–08). IEEE. [7†]
- Stefan Hörmann, Tianlin Kong, Torben Teepe, Fabian Herzog, **Martin Knoche**, and Gerhard Rigoll. “Face Morphing: Fooling a Face Recognition System Is Simple! ”. In *2022 arXiv preprint arXiv:2205.13796*. [8†]
- Stefan Hörmann, Abdul Moiz, **Martin Knoche**, and Gerhard Rigoll. “Attention Fusion for Audio-Visual Person Verification Using Multi-Scale Features”.. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 281–285). IEEE. [9†]
- Stefan Hörmann, Zeyuan Zhang, **Martin Knoche**, Torben Teepe, and Gerhard Rigoll. “Attention-Based Partial Face Recognition”. In *2021 International Conference on Image Processing (ICIP)* (pp. 2978–2982). IEEE. [10†]
- Stefan Hörmann, **Martin Knoche**, Maryam Babaee, Okan Köpüklü, and Gerhard Rigoll. “Outlier-Robust Neural Aggregation Network for Video Face Identification”. In *2019 International Conference on Image Processing (ICIP)* (pp. 1675–1679). IEEE. [11†]
- Stefan Hörmann, **Martin Knoche**, and Gerhard Rigoll. “A Multi-Task Comparator Framework for Kinship Verification”. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 863–867). IEEE. [12†]

Contents

"A journey of a thousand miles begins with a single step."

— Lao Tzu

Introduction

Face Recognition (FR) (or Facial Recognition) is considered the most natural occurrence in the world for humans; it is engaged every time an individual encounters another person or another face. Regardless of whether the setting is real or virtual, the human brain attempts to immediately recognize the identity behind the face. However, the question arises: What does 'recognizing identity' really mean, and what are the underlying brain mechanisms facilitating this process? The computer vision engineer might ask: How can this ability be transferred to machines?

Let us consider the human brain first. Newborns begin to learn FR from the first days after birth. It is believed that in the first few weeks of life, despite having a significantly limited field of vision, particularly in focus and range, infants begin to develop the ability to recognize other humans [13]. Through the repeated observation of both familiar and unfamiliar faces, a rapid development in FR is observed in infants, particularly starting from 4 months of age. At this age, they begin to develop a face 'schema', viewing faces as a special class of stimuli, and their ability to distinguish faces, including by gender and emotional expression, becomes more robust. [14] This capability progressively matures during childhood, reaching near-full development by the age of 10, with minimal further enhancement thereafter [15].

In the field of computer vision, which pertains to machine perception, the objective is to impart the ability to 'see' to machines, transferring and potentially exceeding human capabilities to these devices. With the progress and developments in deep learning, the principle of learning has, in many cases, been successfully transferred to machines. With social media and the growing amount of labeled face data available in the field of FR, this has led to the development of *Convolutional Neural Network* (CNN) architectures that can be trained to recognize faces.

The first attempts to use CNNs for FR and face detection were made in the late 1990s [16, 17]. However, it was in the 2010s that the first CNN architectures were developed that could be trained on a large scale to recognize faces with sufficient accuracy. Early approaches like DeepFace [18] or [19] were able to achieve promising accuracies due to raising available computation power and data available. Consequently, more and more models have been developed and trained, which have advanced to a level where they can identify faces, often outperforming humans in FR tasks [5†, 20, 21].

The fact that faces can nowadays be identified automatically with sufficiently good accuracy has led to enormous technical developments. The range of use cases for such systems has increased tremendously in recent years. To mention just the most prominent examples: FR in smartphones,

which is now used by almost all manufacturers, or the application of FR in social networks to automatically tag people in photos. However, utilizing FR is not limited to these areas. The "global adoption of AI surveillance is increasing rapidly around the world" [22] and is often equipped with FR systems.

As the technology matures, the business of large-scale FR is taking shape. This is exemplified by companies like Clearview AI Inc., which have amassed enormous data collections — approximately 3 billion photos by 2020 — by scraping the internet. These developments enable governments and companies to search for individual faces within a vast array of digital images. This capability potentially facilitates identity verification or surveillance activities on a global scale [23].

As Smith and Miller said "The expanding use of biometrical FR raises a number of pressing ethical concerns for liberal democracies that need to be considered" [24] — The fundamental and growing influence of these systems on society cannot be understated, highlighting the necessity for their reliable functioning and resistance to simple manipulations. Given these significant ethical concerns, examining the technical challenges inherent in FR technology becomes crucial. Understanding these technical obstacles is key to assessing the feasibility and reliability of these systems in practical applications.

1.1. Motivation

Human faces are recognized as unique biometric characteristics. Despite being subject to alterations through makeup, jewelry, and hairstyles, their distinctiveness remains intrinsic to each individual. This sets them apart from other established biometrics like iris patterns and fingerprints. The increasing popularity of FR, exemplified by technologies such as Apple's FaceID to unlock mobile devices, is highlighted by its ease of use in various applications. Furthermore, faces are not only encoded with identity but also convey additional biometric information such as gender, race, and age, along with insights into head pose, gaze direction, and emotional states.

This dissertation focuses on the encoded identity information in facial images. Nowadays, FR systems can identify people from facial images with high accuracy, and thus, automatic systems are widely used. However, the performance, robustness, evaluability, and explainability of these systems are still not sufficient for many applications, and the following paragraphs will elaborate more on that:

The performance of FR systems can be seen two-fold: In terms of accuracy and computation, *i.e.*, speed and memory usage. The accuracy of systems plays a vital role in their usage in probably all real-world applications and is, therefore, a key aspect to consider FR systems. Additionally, speed and, to some extent, storage capacity are critical, as systems are increasingly expected to operate on mobile devices. For instance, automated border control systems, progressively implemented in numerous airports, utilize FR technology for identifying individuals. These systems must be both rapid and reliable to avoid extended waiting times at border controls. Their efficiency and dependability are essential for surpassing the effectiveness of manual checks, a fundamental criterion for their adoption to ever more airports.

Closely related to accuracy is the robustness against deviations from the straightforward case. Depending on the application domain, robustness can be a limiting factor. For instance, in dynamic scenarios where a camera moves towards a person or vice versa, early recognition is desirable for a timely and reasonable response. If recognition only occurs when the face is directly in front of the camera, the response options become severely limited. In this case, the system should be robust to the distance the individual is from the camera.

The measurability of this robustness is a pivotal aspect in the characterization of FR algorithms, providing a crucial basis for assessing and enhancing the resilience of these systems. Such measurability is necessary to compare and evaluate the robustness of systems, thereby hindering their deployment in security-relevant applications. For instance, in the context of automatic identity tagging in movies, knowing the limits and robustness of such a system is essential to deciding whether to spend the money to use the system on a vast amount of data.

The interaction between humans and machines is increasingly becoming critical, particularly regarding sensitive decisions (such as in law enforcement). In such scenarios, it remains imperative that humans make the final decision in borderline cases. However, the challenge lies in reliably identifying these borderline cases and assessing human proficiency in this context. Because of their exceptional ability to identify individuals even in complex situations, Super-recognizers play a vital role here [25]. As surveillance expands, leading to a surge in image data, the frequency of borderline cases also increases. This situation necessitates not only more precise and reliable systems but also a more detailed analysis of borderline cases and enhanced explanations or descriptions of why a case is identified as borderline by the system.

In a nutshell, the aspects above are crucial for the acceptance of FR systems and their widespread adoption. More robust systems can unlock new opportunities here, notably in law enforcement, such as mobile FR with drones, which typically operate at greater distances from subjects. Or also in the widespread search for missing or dementia-affected individuals.

This study primarily concentrates on the technical dimensions of FR systems, endeavoring to augment their robustness, evaluability, and explainability. While the ethical aspects are not the central focus of this research, it is imperative to recognize their significance in the realm of FR. The rapid advancement and increasingly pervasive deployment of FR technologies necessitate thoroughly considering the associated ethical challenges. These challenges encompass issues like bias, data collection, transparency, protection of vulnerable populations, security, and accountability [26]. Recent scholarly discussions have extensively addressed these ethical concerns in FR systems [27, 28, 29]. Furthermore, Fisher *et al.* [26] and Ketley *et al.* [30] have proposed a specific code of ethics aimed at tackling these critical issues, highlighting the growing acknowledgment of ethical considerations in the development and application of FR technologies. Consequently, considering ethical aspects in research and deployment of FR technology is essential to ensure that these systems are used responsibly and do not infringe upon fundamental human rights.

1.2. Challenges

The significance and broader implications of FR technology having been established, the focus is now shifted to the technical challenges inherent in this field. Its expanding applications in real-world scenarios present not only technical hurdles but also highlight critical legal and ethical considerations. Given that FR technology plays a crucial role in individual identification, its implications for data protection and privacy are profound. Moreover, its use in law enforcement and expanded surveillance intensifies concerns regarding potential misuse and abuse.

Nevertheless, beyond these societal and ethical dimensions, FR confronts various technical challenges that are pivotal in its development and have recently attracted substantial research interest. To facilitate a better understanding, this section draws a comparison between humans and machines in the context of FR abilities, highlighting their differences and similarities. A critical distinction is that humans, during their FR developmental phase, encounter only a limited number of people and thus faces. In contrast, with advancing technological capabilities and the massive amount of collected data, machines have access to significantly more data, including up to billions of identities. Notably, one must also consider the current research on artificially generated faces, which opens up the possibility of generating an infinite number of identities—however, not just the quantity but also the quality of data matters during the learning process. While machines can access more data, the variety differs significantly from the variety a human being sees in its early FR development phase.

Having initially examined the overall data landscape, the focus is now on the image level. Starting with a straightforward problem that is easy to solve for machines, human proficiency in FR can be quickly challenged by simple manipulations, such as rotating an image 180 degrees, which significantly diminishes the recognition ability of most people. This likely stems from our usual experience of seeing faces in a specific orientation. Through basic detection of eyes, nose, and mouth, images can be normalized to present a consistent orientation to the machine, thus overcoming this challenge. However, more complex obstacles remain and can be categorized into four fundamental challenges in FR: Pose, age, occlusion, and quality. These challenges are illustrated in Figure 1.1 and briefly described below.

Pose. *E.g.*, consider the angle of view on faces. The third column of Figure 1.1 illustrates this by comparing a profile picture to a frontal shot. It is evident that humans constantly encounter faces from a variety of angles. However, large labeled image databases do not always encompass this range of variability. *E.g.*, official documents like passports and driver's licenses typically feature faces captured frontally. Similarly, profile pictures on social media are predominantly frontal shots.

Age. In comparison, age disparity presents itself as a significant challenge. For instance, during the first ten years, crucial to developing FR skills, a human is exposed to a maximum age difference corresponding to this time span. However, machines can be presented with a much broader range of age differences. As time progresses, an increasing number of images capturing longer spans of a person's life are digitally documented. This is exemplified in the first column of Figure 1.1, illustrating the complexities machines face in adapting to the wide spectrum of age variations. This factor significantly influences the effectiveness of FR technology.

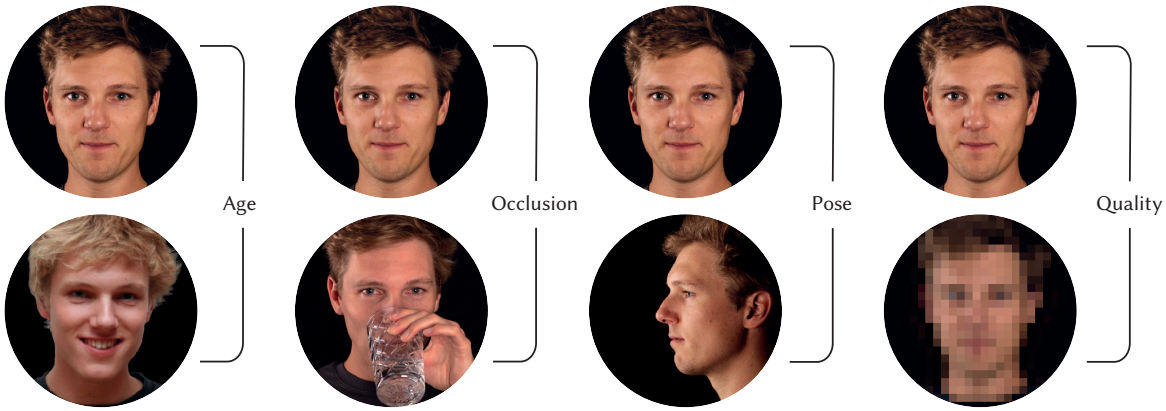


Figure 1.1.: Four example cases depicted to illustrate challenges in face recognition.

Occlusion. The recent pandemic has brought a significant issue: the increased difficulty in recognizing individuals wearing masks. This situation exemplifies the challenges of partial occlusion in FR systems, as seen in the second column of Figure 1.1. A threshold exists in human and machine perception where, once a face is no longer discernible, FR ceases to function effectively. From the outset, humans have encountered various occlusions, such as glasses, hats, and, more recently, masks covering the mouth and nose. These obstructions are often visible in the digital realm and can be synthetically generated with relative ease. This highlights the complexities in ensuring accurate FR in partial face coverings, both in human experience and machine processing.

Quality. Distance to the face emerges as a critical factor complicating identification processes for both humans and machines. For humans, there is an inherent limitation to recognizing individuals beyond a certain range. This constraint is mirrored in machine-based systems, where the resolution of an image inherently limits the amount of detail and information it can convey. An example is given in the last column of Figure 1.1. It is pertinent to acknowledge that the range of digitally available data, particularly concerning facial distance, is quite restricted. The majority of these data, predominantly sourced from the internet, fail to include images captured from extensive distances. This absence is attributed to the reduced facial clarity in such images, which subsequently hinders the recognition of individual identities. Consequently, a notable dearth of digital images taken from long distances exists. When evaluating digital images in the context of distance, considerations generally pivot toward image resolution or overall quality. The term 'quality' in this context extends to include aspects such as lighting, noise levels, contrast, and other relevant variables. Another challenge in the digital domain is the emergence of transferable adversarial attacks [i]. These sophisticated attacks target machine learning systems and are uniquely designed to be undetectable by human observers.

The foundational step naturally involves measuring performance and analyzing robustness to subsequently demonstrate improvements. This means initially addressing the foundation, which is the efficient measurement of image resolution robustness in FR systems. Subsequently, efforts can be

[i] Adversarial attacks are deliberate manipulations of input images with subtle changes to fool the system into incorrect face identification or recognition failure.

directed towards enhancing robustness, ideally without compromising performance in other domains. A significant challenge in this domain is the scarcity of labeled data that specifically reflects this issue, as typically, there are only a few labeled images with low resolution. Consequently, the synthetic generation or modification of data must be considered.

Particularly in relation to deep learning algorithms, understanding the explainability and traceability of decisions is crucial. Typically, these systems are akin to 'black boxes' initially yielding only a numerical output. A decision for verification or rejection is then made based on a certain threshold value. However, the underlying question remains: How is this decision arrived at? Which parts of the image are especially pivotal? Finally, it is important to identify the weaknesses of a system and thereby filter out cases where automatic systems fail and are not reliable. This raises the question of the system's certainty: How confident is it in its decision-making, and would a human possibly be more certain in the same scenario?

In summary, this chapter has outlined the key challenges that FR technology currently faces. The subsequent chapter will build upon this foundation, outlining specific objectives and methodologies designed to enhance the accuracy and reliability of FR systems in the face of these complexities.

1.3. Scope and Research Questions

The overarching aim of this dissertation is to delve into various aspects of FR, with a particular emphasis on the influence of different resolutions on the images being compared, *i.e.*, *Cross Resolution* (CR). This exploration is crucial for understanding and enhancing the efficacy of FR systems in diverse conditions. To achieve this, this work sets forth several key research objectives, each addressing a distinct facet of the broader topic. These objectives have been crystallized into five research questions, which will guide the investigative process and methodology of this study:

Research Question 1: *How does image resolution impact the performance of face recognition systems?*

A fine-grained analysis of the impact of image quality on FR performance is described in Chapter 3 and in [1†]. Several benchmarks of state-of-the-art FR systems are presented and their degrading performance on images with varying resolutions is thoroughly analyzed. Additionally, a closer look on the feature distances is conducted and discussed.

Research Question 2: *What strategies can be developed to enhance the robustness of face recognition systems against variations in image resolution?*

Chapter 4 presents three training strategies published in [1†, 2†] to enhance robustness to image resolution. The first strategy is based on a simple data augmentation technique, *i.e.*, the training data is augmented with images of lower resolution. The second strategy is based on a siamese network architecture, whereas each branch is trained with images of different resolutions. The latter strategy is based on data augmentation in combination with a specific loss function which penalizes the distances of genuine and imposter image pairs at the same time during training. Each strategy is

evaluated on several benchmarks and compared to the state-of-the-art FR systems.

Research Question 3: *How can a new benchmark be designed to measure the performance of face recognition systems on images with different resolutions more accurately and precisely than current methods?*

A novel benchmark dataset is presented in Chapter 5 and in [3†]. The dataset originates from the Labeled Faces in the Wild (LFW) dataset [31] and was modified to include image pairs with difference in image quality. The chapter presents the dataset and the benchmark protocol. Furthermore, it evaluates multiple state-of-the-art FR systems on the dataset and compares the results to the default *Labeled Faces in the Wild* (LFW) benchmark.

Research Question 4: *What methods can be established to provide clearer explanations for the decisions made by face recognition systems?*

Chapter 6 and [4†] describe two methods to make the predictions of *Face Verification* (FV) systems more explainable. The chapter explains an algorithm for calculating a *Confidence Score* (C-Score) for the output of an arbitrary FR model. Furthermore, it presents the derivation of model-agnostic similarity maps, which visualize regions in the images that are important for the decision of an FR system. Finally, a web platform is introduced, which includes the implementation of those explanation methods and allows to interactively explore the maps and scores calculated for several FV benchmark datasets.

Research Question 5: *How do humans perform in borderline cases of face recognition, and can an algorithm be developed to effectively fuse human and machine decisions to improve overall accuracy in face recognition tasks?*

The last contribution of this dissertation shifts the focus to human capabilities in FR and explores how humans make different decisions compared to state-of-the-art FR systems and is described in Chapter 7 and [5†]. Special attention is given to borderline cases, which are analyzed in detail. The chapter concludes by presenting a simple algorithm that fuses the decisions of humans and machines, demonstrating that this fusion can lead to an overall improvement in results.

Aside from the above listed main contributions of this dissertation, the foundational context for image processing, particularly w.r.t. resolution and quality is established in Chapter 2. Furthermore, the fundamental processes and characteristics of state-of-the-art FR systems are introduced, which serve as the basis for the main body of this work. After presenting the fivefold contributions, Chapter 8 concludes the findings of this work and draws the main implications w.r.t. the research objectives stated above. Finally, it draws an outlook to interesting directions for future work. The graphical overall structure of this dissertation is also illustrated in Figure 1.2.

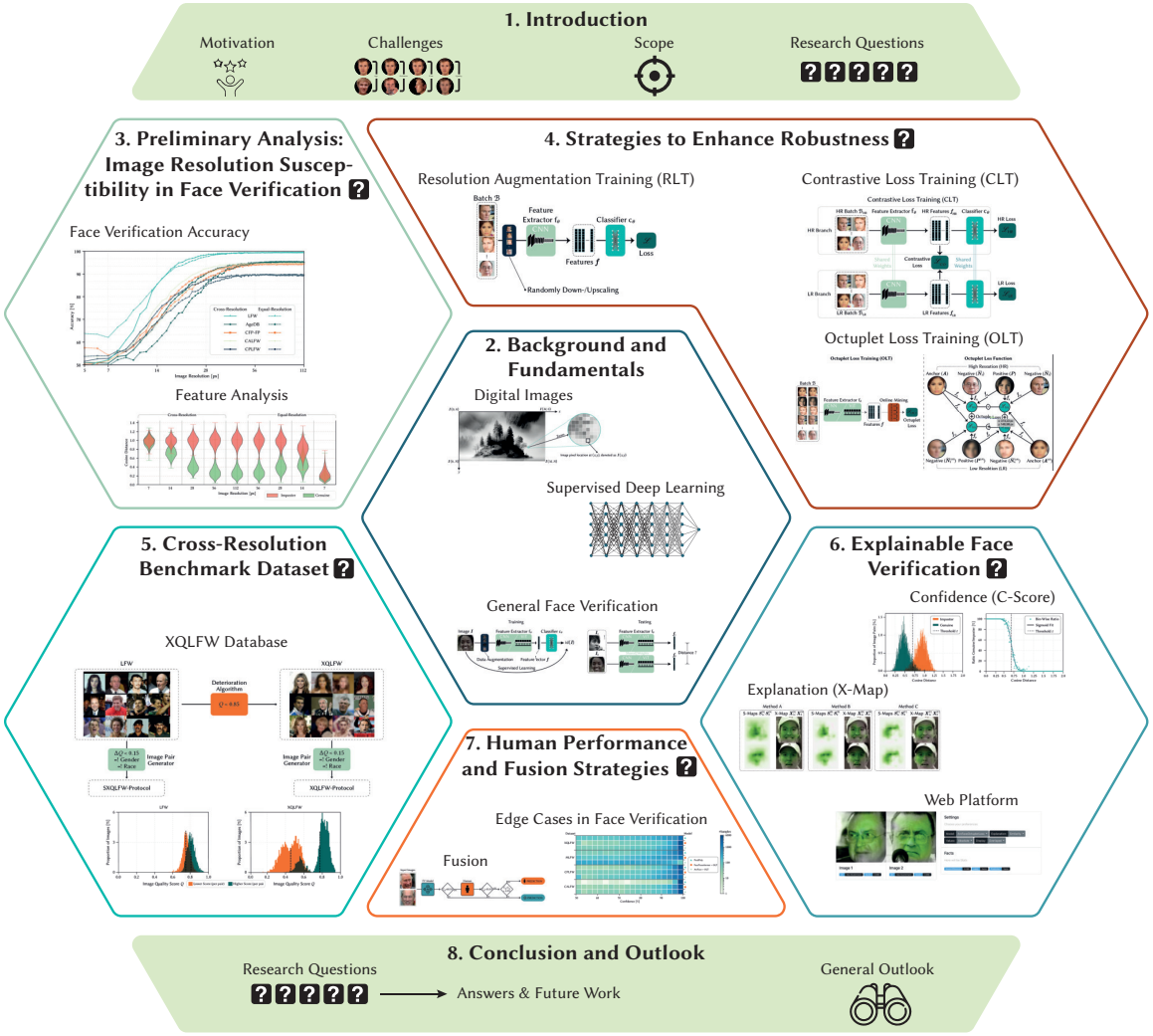


Figure 1.2.: Graphical structure of the dissertation.

“There is no desire more natural than the desire for knowledge.”

– Michel de Montaigne

Background and Fundamentals

In this chapter, an exploration of the fundamentals of image resolution in the context of digital image processing is introduced. Then, the basics of supervised *Deep Learning* (DL) are explained. Finally, the field of *Face Verification* (FV) is introduced, covering a typical state-of-the-art *Face Recognition* (FR) pipeline for training and evaluation of an FR system. Morover, popular datasets used for training and testing are presented. The mathematical notation utilized throughout this work is based on *International Organization for Standardization* (ISO) 80000–2 [32] and is summarized in Appendix A.

2.1. Digital Images

A digital image can be described as a scan of the real world, projecting a 3D scene onto a 2D image plane with a grid of pixels. Each pixel represents color and luminance, offering a snapshot of reality, albeit with the inherent loss of depth information. The 2D discrete, digital image $\boldsymbol{I}(x, y)$ represents the response of a capturing sensor at a series of fixed positions in 2D Cartesian coordinates $x \in \{1, 2, \dots, M\}$, $y \in \{1, 2, \dots, N\}$, and is derived from the 2D continuous spatial signal through a sampling process frequently referred to as discretization. The discretization process takes place in imaging sensors, typically performing a localized averaging of the continuous signal (world) across a small, often square-shaped area, in the receiving domain (image). The discrete image $\boldsymbol{I}$ can be described as a function of two variables, which are the spatial coordinates x and y. The image size is consequently defined by the number of pixels in the horizontal and vertical directions, which are M and N, respectively. Figure Figure 2.1 illustrates the layout of a digital grayscale image. The spatial resolution of the image is described then $M \times N$ px, whereas the total number of pixels is calculated as $M \cdot N$. Note that in this work, the term 'pixel' is abbreviated as px.

The value of the function at any point (x, y) is the intensity of light captured in the image at that point. In a grayscale image, the intensity is a single value, whereas in a color image, the intensity is typically a vector of three values, representing the red, green, and blue channels. There also exist other color models, such as the *HSV* model, which is based on hue, saturation, and value or the *CYMK* model, which is based on cyan, yellow, magenta, and black. Common values for the intensity are in the range of $[0, 255]$, where 0 represents black and 255 represents white. The values in between are then shades of gray. This range is also called 8-bit grayscale, due to the fact that each pixel is represented by 8 bits. In mathematical notations of images $\boldsymbol{I} \in \mathbb{R}^{M \times N \times K}$, the numbers of channels K,

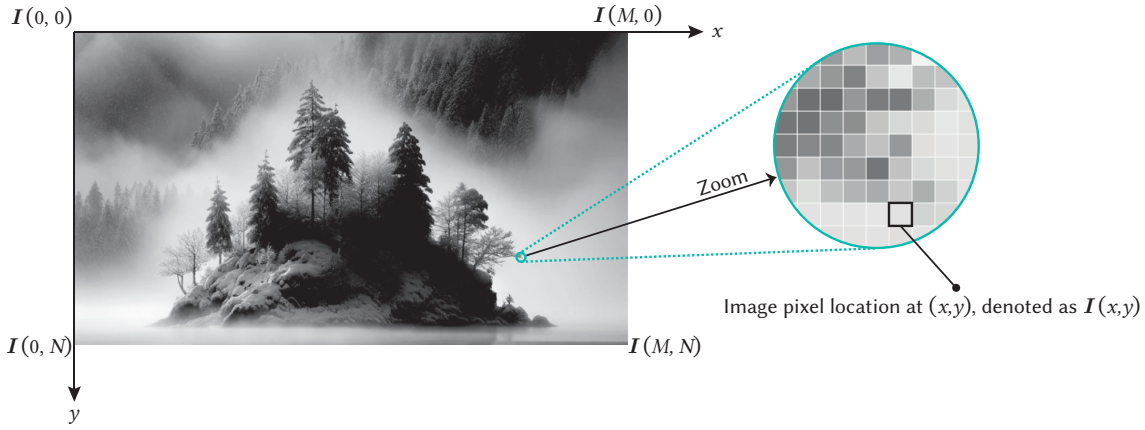


Figure 2.1.: Layout of a digital grayscale $M \times N$ px image, with a zoomed-in section displaying individual pixels. Image generated with DALL·E 3[i] in 2024.

to encode the color information is often omitted, since it is not relevant for the operations discussed in this dissertation.

2.1.1. Resolution

The general concept of the word 'resolution' related to digital images can be specified in terms of four quantities:

Spatial Resolution is the number of pixels in the horizontal and vertical directions, which are M and N, respectively. In this work, the spatial resolution is also called *image size* and is measured in pixels. The higher the spatial resolution, the more pixels are used to represent the image, resulting in a larger file size.

Bit Resolution or ***Color Depth*** is the number of bits used to represent the intensity of a pixel. This corresponds to a bit resolution of 8 bits and the full continuos spectrum is represented by $2^8 = 256$ discrete values. Higher bit resolutions allow for the representation of more values (information), which in turn increases the file size of the image.

Color Resolution means the number of channels in the image. This can be considered as the third dimension of an image. A grayscale image has one value for each pixel, whereas a *Red Green Blue* (RGB) color image has three values, representing the red, green, and blue channels. All experiments covered by this work use exclusively RGB color images and a bit resolution of 8 bits per channel. For the sake of simplicity, the number of channels is typically omitted in this work when denoting the image dimension.

[i] DALL·E is an artificial intelligence program developed by OpenAI that can generate digital images from textual descriptions.

Raw-resolution describes the original image resolution as captured by a camera. Hence, it is independent of posterior up-scaling to a larger spatial resolution. This does naturally not increase the information content of the image although the image contains a high (spatial) resolution. *E.g.*, a low raw-resolution image describes an image necessarily upscaled from a lower original spatial resolution image (see Section 2.1.3). Since all FR processing in this work deals with 112×112 px images and to avoid clutter, the shorter term 'resolution' is used to denote the raw-resolution of an image in the remaining chapters. It is important to mention that this work does not consider motion or lense blur in images. Additionally, the depth of field effect created by the camera's aperture setting is overlooked. The focus is strictly on the face.

2.1.2. Scaling

The term scaling is used to describe the process of changing the spatial resolution of an image. The process of increasing the spatial resolution of an image $\boldsymbol{I}$ is described as *upscaling* or sometimes called up-sampling. In contrast, the process of decreasing the spatial resolution of an image is called *downscaling* or sometimes called down-sampling. In the following, both processes are described in more detail.

Downscaling decreases the image size and therefore reduces the amount of information in the image. For an image $\boldsymbol{I} \in \mathbb{R}^{M \times N}$, the process of downscaling $\mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{\lfloor \frac{M}{s} \rfloor \times \lfloor \frac{N}{s} \rfloor}$ can be formulated as a function $\downarrow_s(\cdot)$:

$$\boldsymbol{I}^* = \downarrow_s \left(\boldsymbol{I} \right), \tag{2.1}$$

with $\boldsymbol{I}^*$ being the downscaled image and the scaling factor $s \in \mathbb{N}_{>0}$. To obtain a downscaled image, the pixel values of the input are mapped to the output. If those pixel coordinates are not an integer, interpolation is required to estimate the actual pixel value. The simplest way to downscale an image is to throw pixels away, which is called *nearest-neighbor* interpolation. Other methods like bilinear, bicubic, or Lanczos [33] interpolation are also used. But aliasing artifacts (or Moire patterns) are present independent of the chosen methods. They occur when the sampling frequency is too low to represent the signal accurately. To avoid aliasing effects, typically a low-pass filter, *e.g.*, a Gaussian filter, is applied before downscaling. Figure 2.2 shows an example of downscaling an 6×6 px image $\boldsymbol{I}_1$ by a factor of 3 using nearest-neighbor interpolation on the left side, resulting in an 2×2 px image $\boldsymbol{I}_1^*$.

Upscaling increases the image size by adding pixels, but this does not increase the amount of information in the image. For a square-sized image $\boldsymbol{I} \in \mathbb{R}^{M \times N}$ with an arbitrary spatial resolution of $M \times N$ px, the process of upscaling $\mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{M \cdot s \times N \cdot s}$ can be formulated as a function $\uparrow_s(\cdot)$ with $s \in \mathbb{N}_{>0}$ being the scaling factor:

$$\boldsymbol{I}^* = \uparrow_s \left(\boldsymbol{I} \right), \tag{2.2}$$

with $\boldsymbol{I}^*$ being the upscaled image. To obtain an upscaled image, the pixel values of the input are mapped to the output. Nearest-neighbor interpolation is the simplest way to upscale an image. Here, the output pixel values are filled with the nearest input pixel value. Figure 2.2 shows in the right an example of upscaling an 2×2 px image $\boldsymbol{I}$ by a factor of 3 using nearest-neighbor interpolation. The resulting image is then a 6×6 px image $\boldsymbol{I}^*$. Other methods incorporate more information around the pixel. *E.g.*, bicubic interpolation uses a weighted average of the 16 nearest pixels and the bilinear

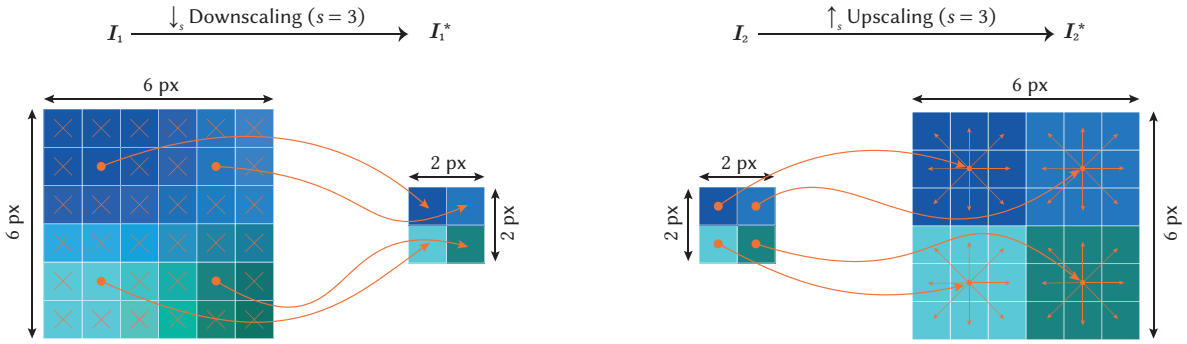


Figure 2.2.: Illustration of nearest-neighbor interpolation of an image.

approach uses a weighted average of the four nearest pixels. In today's digital image processing, additional methods such as Lanczos, Mitchell-Cubic, Gaussian, or Area interpolation exist, but are not covered in this work. The term upscaling without any further clarification, refers to bilinear upscaling in this work.

Besides interpolation, there exist also DL techniques and are referred to as *Super Resolution* (SR). The goal of SR is to recover a *High Resolution* (HR) image from an *Low Resolution* (LR) image. If only one image is available, this is called *Single Image Super Resolution*, whereas if multiple images are available, this is called *Multi Image Super Resolution*. The interested reader is referred to the work of Lepcha *et al.* [34] or Li *et al.* [35] for a more detailed overview of SR.

2.1.3. Raw Resolution

It is important to understand that spatial resolution is not directly a metric for informational content or quality of an image. To address this issue and describe the informational content of an image or in other words the quality, the term *raw-resolution* was introduced (see Section 2.1.1). This characteristic is illustrated in Figure 2.3 using the well-known camera obscura [i] model and capturing two image shots $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ of a person standing at different distances d_1 and d_2 to the camera.

According to the pinhole camera model, the distance d of the face to the camera with a focal length of f is inversely proportional to the height h^* of the face in the captured image:

$$\frac{h^*}{f} = \frac{h}{d}\,. \tag{2.3}$$

A larger distance between face and camera results in a smaller face in the image. In Figure 2.3, the face of the person standing closer to the camera (d_1) fills an area of about 56×56 px in the captured image $\boldsymbol{I}_1$. The face of the person standing further away from the camera (d_2) occupies an area of only 14×14 px in $\boldsymbol{I}_2$. Assuming that an FR system requires a fixed input image size of 112×112 px, the facial images must be upscaled to meet this requirement. Here, they need to be upscaled by a

[i] From Latin: camera obscura 'dark chamber'

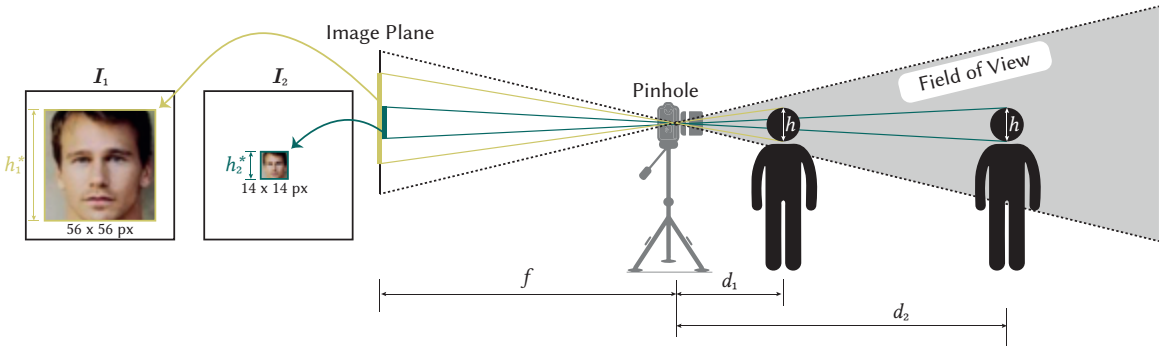


Figure 2.3.: Illustration with the pinhole model of captured images of two people with different distances to a camera.

factor of 2 and 8, respectively. Figure 2.4 visualizes the upscaling results by leveraging two different interpolation methods, nearest-neighbor and bicubic interpolation. Although all images share the same spatial resolution of 112×112 px, the upscaled images contain different levels of information. The upscaled variants $\boldsymbol{I}_1^*$ of $\boldsymbol{I}_1$ contains significantly more information, since it was only upscaled by a factor of 2, or in other words, the raw-resolution of the original image is 56×56 px. In contrast, the upscaled variants $\boldsymbol{I}_2^*$ of $\boldsymbol{I}_2$ contain less information, since the original image was upscaled by a factor of 8, *i.e.*, the raw-resolution of the original image is 14×14 px. This example highlights the necessity of considering the raw-resolution of an image to describe the quality of an image.

An HR image denotes an image with arbitrary spatial resolution but necessarily captured by a camera with the same or higher resolution. In contrast an LR image denotes an image with arbitrary spatial resolution but necessarily upscaled from a lower resolution image, in other words being characterized by a lower raw-resolution.

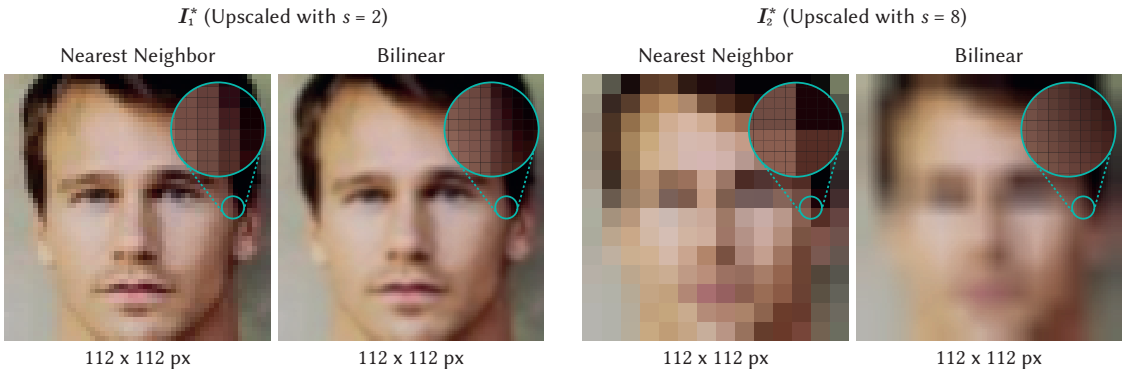


Figure 2.4.: Illustration of the raw-resolution in two upscaled images. Both images are upscaled with nearest-neighbor and bilinear interpolation.

In general, the term *image quality* is comprising of several types of image quality degradation and besides the raw-resolution also includes sharpness, noise, artifacts, and other types.

2.1.4. Raw Resolution Assessment

If the raw-resolution of an image is known, it can be used to describe the degree of degradation, *e.g.*, with the scaling factor obtained to the desired spatial resolution. A higher scaling factor means a higher degradation. This is typically the case for synthetically degraded images or when the information about the original raw image resolution is present. However, there exist also images, where it is unknown, whether and how much the image was upscaled. In this case, the raw-resolution can only be estimated. Dependent on the upscaling technique, the image could be very blurry, as it is the case for bicubic or bilinear upscaling. In contrast, the image could be very sharp, as it is the case for upscaling with nearest-neighbor interpolation (see Figure 2.3). The following two approaches are used in this work:

The *Blind Referenceless Image Spatial Quality Evaluator* (BRISQUE) score introduced by Mittal *et al.* [36] is a no-reference image quality assessment metric. It quantifies the 'naturalness' of images, which is affected by distortions. BRISQUE uses scene statistics of locally normalized luminance coefficients without a distortion-specific model. The method evaluates the empirical distribution of these normalized luminances and their products in the spatial domain. BRISQUE is distinct from other approaches (see [37] for an overview), as it does not require transformation to another coordinate frame like *discrete cosine transform* or *wavelet*, operating purely in the spatial domain.

Stochastic Embedding Robustness Face Image Quality (SER-FIQ) is a DL-based unsupervised approach to estimate face image quality based on the robustness of stochastic embeddings introduced by Terhörst *et al.* [38]. It operates without explicit quality labels. The methodology involves generating multiple stochastic embeddings from random sub-networks of an FR model and then measuring the variations in these embeddings. High robustness (low variation in embeddings) indicates higher image quality, while lower robustness (high variation) suggests lower quality.

2.2. Supervised Deep Learning

*Artificial Neural Network*s (ANNs) aim to emulate human cognitive processes by applying logical rules. Initially, an ANN consisted of just one processing unit, or *neuron*, which could only predict binary outcomes from multiple binary inputs [39]. However, modern ANNs, with their millions of neurons, have long outperformed humans in various tasks. In particular, ANNs have been the driving force behind the recent advances in computer vision. In the field of FR, 2014 was the year of breakthroughs, with the introduction of DeepFace [18] and DeepID [40]. Research from then on has shifted to DL-based approaches.

Very roughly speaking, DL can be divided into three subfields, *deep supervised learning*, *deep unsupervised learning*, and *deep reinforcement learning*. To follow the remaining part of this dissertation this section aims to give a brief overview of supervised learning in the context of DL.

2.2.1. Artificial Neural Networks

An ANN is a computational model inspired by the human brain. It consists of multiple layers of interconnected nodes, which are called *neurons*. Each neuron is a simple computational unit that takes multiple inputs, processes them, and produces an output. The output of a neuron is typically a non-linear function of the weighted sum of its inputs. The weights are the learnable parameters of the network. The output of a neuron is then passed to the next layer of neurons. In the following, specific ANN architectures, typically applied in the realm of FR are discussed. A typical architecture for supervised DL FR models is two-fold, consisting of a *feature extraction* network and a *classification* network. The feature extraction network is responsible for extracting features from a given image, whereas the classification network is responsible for classifying the extracted features. Considering the size of these networks, the feature extraction network is typically larger than the classification network. The following paragraphs elaborate the mathematical description of both networks:

Feature Extraction Networks. A feature extractor network maps a facial image $\boldsymbol{I} \in \mathbb{R}^{M \times N \times K}$ to a feature vector representation $\boldsymbol{f} \in \mathbb{R}^{F}$. This can be simplified as a parametrized differentiable function f_θ utilizing a set of learnable parameters θ:

$$\mathrm{f}_\theta \;:\; \mathbb{R}^{M \times N \times K} \;\rightarrow\; \mathbb{R}^{F} \quad , \quad \boldsymbol{I} \;\mapsto\; \mathrm{f}_\theta\big(\boldsymbol{I}\big) = \boldsymbol{f}\,. \tag{2.4}$$

This network f_θ is typically composed of multiple layers, where the i-th layer $\mathrm{g}^{(i)}_{\theta^{(i)}}$ consumes the output of the previous $i-1$-th layer, *i.e.*,

$$\mathrm{f}_\theta = \mathrm{g}^{(N)}_{\theta^N} \circ \mathrm{g}^{(N-1)}_{\theta^{N-1}} \circ \ldots \circ \mathrm{g}^{(1)}_{\theta^1}\,, \tag{2.5}$$

with θ being the set of all parameters of the underlying layers in the network. In the field of FR feature extraction is typically done via *Convolutional Neural Network* (CNN) or *Vision Transformer* (ViT) [41] architectures, which will be described in more detail in Section 2.3.2.1. The purpose of these architectures is to obtain a meaningful representation of facial identity in a lower-dimensional space.

Classification Networks. The purpose of a classifier networks in the context of FR is to classify facial features to a specific class, *i.e.*, an identity. This means to map a feature vector $\boldsymbol{f} \in \mathbb{R}^{F}$ to a class label l, which is represented by a vector $\boldsymbol{y} \in \mathbb{R}^{L}$ containing posterior-probabilities for L different classes. More simply explained, a parametrized differentiable function c_θ is utilizing a set of learnable parameters θ:

$$\mathrm{f}_\theta \;:\; \mathbb{R}^{F} \;\rightarrow\; \mathbb{R}^{L} \quad , \quad \boldsymbol{f} \;\mapsto\; \mathrm{c}_\theta\big(\boldsymbol{f}\big) = \boldsymbol{y}\,. \tag{2.6}$$

Instead of directly outputting a class decision, the network outputs class posterior-probabilities and in the context of multi-class classification, the last layer typically employs a *softmax* activation function on the i-th element of $\boldsymbol{y}$:

$$\mathrm{SOFTMAX}(\boldsymbol{y}_i) = \frac{e^{\boldsymbol{y}_i}}{\sum_{l=1}^{L} e^{\boldsymbol{y}_l}}\,, \tag{2.7}$$

which ensures that the output of the network contains a valid probability distribution, *i.e.*, the elements of $\boldsymbol{y} \in [0, 1]$ sum up to 1. For any input feature $\boldsymbol{f}$, the function f_θ generates an output vector $\boldsymbol{y}$ containing posterior-probabilities $\boldsymbol{p}$, leading to the following decision rule:

$$\mathbb{R}^L \rightarrow l \in \{1, 2, \dots, L\} \quad , \quad \boldsymbol{I} \mapsto \arg\max_{\boldsymbol{p}} \boldsymbol{p}(l \mid \boldsymbol{I}) \,. \tag{2.8}$$

This section outlined the fundamental principle of ANN structures, that are used to solve supervised learning tasks. Moving on to loss functions in the next sections, it is crucial to to grasp how they have been instrumental in extending the boundaries of accuracy in FR technologies over recent years.

This outlined the fundamental principle of ANN structures, that are used to solve supervised learning tasks.

2.2.2. Loss Functions

It is crucial to grasp how loss functions have been instrumental in extending the boundaries of accuracy in FR technologies over recent years. This sections aims to provide a brief overview of loss functions in the context of DL and FR.

Let D be the number of available known input-output data pairs, a dataset can be described as $\mathcal{X} := \left\{(\boldsymbol{I}_1, \hat{\boldsymbol{y}}_1), \dots, (\boldsymbol{I}_D, \hat{\boldsymbol{y}}_D)\right\}$ containing D images $\boldsymbol{I}$ with associated one-hot encoded ground truth class labels $\hat{\boldsymbol{y}}$, *i.e.*, 1 for the correct class and 0 for all other classes. In the context of FR, the objective then is to approximately compute the output of the $(D+1)$-th input data $\boldsymbol{I}_{D+1}$ without leveraging explicit knowledge of the network function but instead by using the knowledge of the D input-output data pairs. This is done by minimizing a differentiable loss function $\mathscr{L}$ (also known as *training criterion*, *objective function*, or *cost function*) to find the optimal parameters θ^* of a model containing the parameters θ:

$$\theta^* = \arg\min_{\theta} \mathscr{L}\left(\mathcal{X};\, \theta\right), \tag{2.9}$$

such that the difference between the predicted output $\boldsymbol{y}$ and the ground truth $\hat{\boldsymbol{y}}$ is minimized. In the following sections, the general principle of loss functions for classification and deep metric learning are introduced.

2.2.2.1. Deep Metric Learning

In general, *deep metric learning* or *similarity learning* maximizes discriminability of feature vectors, *i.e.*, reducing the distance between feature vectors corresponding to the same class and increasing the distance between the feature vectors corresponding to different classes. Commonly utilized distance metrics are the euclidean distance and the cosine distance. Considering two n-dimensional vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$ the euclidean distance $\mathrm{d}_{\mathrm{EUC}}$ is calculated as follows:

$$\mathrm{d}_{\mathrm{EUC}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \sqrt{\sum_{i=1}^{n} \left(v_1^{(i)} - v_2^{(i)}\right)^2}, \tag{2.10}$$

where $v_1^{(i)}$ and $v_2^{(i)}$ are the i-th elements of the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, respectively. The cosine distance $\mathrm{d}_{\cos} \in [0, 2]$, on the other hand measures the angular distance in feature space and is calculated as:

$$\mathrm{d}_{\cos}(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{\|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\|}\,. \tag{2.11}$$

Deep metric learning distinguishes itself by utilizing distances rather than directly employing class labels from the dataset. Instead, it depends on the binary information indicating whether pairs of images share the same or different identities.

The primary loss functions facilitating this approach are contrastive loss and triplet loss. They are explained in the subsequent paragraphs:

Contrastive Loss. The contrastive loss function was initially formulated by Chopra *et al.* [42] and first employed in the field of FR by Sun *et al.* [40, 43]. The contrastive loss function aims to minimize the feature distance between two faces of the same identity and maximize the feature distance between faces of different identities. Considering a set of D facial images $\mathcal{X} := \{\boldsymbol{I}_1, \boldsymbol{I}_2, \dots, \boldsymbol{I}_D\}$ with corresponding class labels, a set of all possible triplets $\mathcal{T}$ can be defined according to the following rule:

$$\mathcal{T} := \left\{ \left(Y, \boldsymbol{f}_1, \boldsymbol{f}_2\right) \;:\; \boldsymbol{f} = \mathrm{f}_\theta\left(\boldsymbol{I}\right),\, \boldsymbol{I} \in \mathcal{X},\, \boldsymbol{I}_1 \neq \boldsymbol{I}_2,\, Y = 0 \text{ if } \mathrm{id}\left(\boldsymbol{I}_1\right) = \mathrm{id}\left(\boldsymbol{I}_2\right) \text{ else } Y = 1 \right\}. \tag{2.12}$$

Y is a then a binary label indicating whether the two images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ belong to the same identity or not. The contrastive loss function $\mathscr{L}_{\mathrm{CO}}(\,\cdot\,;\theta)$ is then defined as:

$$\mathscr{L}_{\mathrm{CO}}\left(\mathcal{T};\theta\right) = \frac{1}{|\mathcal{T}|} \sum_{(Y,\boldsymbol{f}_1,\boldsymbol{f}_2)\in\mathcal{T}} (1-Y)\frac{1}{2}\mathrm{d}\left(\boldsymbol{f}_1, \boldsymbol{f}_2\right) + Y\frac{1}{2}\left[0, \alpha - \mathrm{d}\left(\boldsymbol{f}_1, \boldsymbol{f}_2\right)\right]_+, \tag{2.13}$$

with $\mathrm{d}\,(\,\cdot\,,\,\cdot\,)$ being an arbitrary distance function and α a margin. The role of the margin, α, is to establish a hypersphere with radius α around each point, $\boldsymbol{f}$, in the embedding space. The algorithm aims to ensure that no points from a different class fall within this hypersphere. It effectively sets a target separation distance between dissimilar pairs. If the distance between a pair of dissimilar items is greater than α, the loss for that pair is 0, *i.e.*, the model is not penalized, as it has achieved the desired outcome.

This approach, which concentrates on pairs of images, yields absolute feature distances. However, this method leads to complications arising from varying intra-class variances among different identities. Therefore the loss function was extended to the triplet loss function.

Triplet Loss. Schroff *et al.* [44] opted for the triplet loss in the field of FR. This method works by simultaneously minimizing the distance between the features of an anchor face $\boldsymbol{I}_{\mathrm{A}}$ and a positive face $\boldsymbol{I}_{\mathrm{P}}$ (belonging to the same identity) while maximizing the distance between $\boldsymbol{I}_{\mathrm{A}}$ and a negative face $\boldsymbol{I}_{\mathrm{N}}$ (from a different identity), effectively considering triplets of the images $\left(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{P}}, \boldsymbol{I}_{\mathrm{N}}\right)$. Figure 2.5 visually explains the repelling and attracting mechanism of the loss function to the features in the feature space. Considering a set of facial images $\mathcal{X}$ with corresponding class labels, a set of all possible triplets $\mathcal{T}$ within a dataset $\mathcal{X}$ can be defined according to the following rule:

$$\mathcal{T} := \left\{ \left(\boldsymbol{f}_{\mathrm{A}}, \boldsymbol{f}_{\mathrm{P}}, \boldsymbol{f}_{\mathrm{N}}\right) \;:\; \boldsymbol{f} = \mathrm{f}_\theta\left(\boldsymbol{I}\right),\, \boldsymbol{I} \in \mathcal{X},\, \mathrm{id}\left(\boldsymbol{I}_{\mathrm{A}}\right) = \mathrm{id}\left(\boldsymbol{I}_{\mathrm{P}}\right),\, \mathrm{id}\left(\boldsymbol{I}_{\mathrm{A}}\right) \neq \mathrm{id}\left(\boldsymbol{I}_{\mathrm{N}}\right),\, \boldsymbol{I}_{\mathrm{A}} \neq \boldsymbol{I}_{\mathrm{P}} \right\}. \tag{2.14}$$

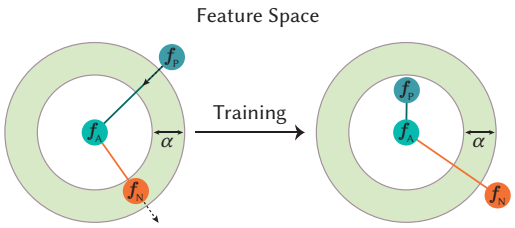


Figure 2.5.: Illustration of the triplet loss mechanism on the features of an anchor, positive, and negative image.

The equation for the triplet loss function $\mathscr{L}_{\text{TRI}}(\,\cdot\,;\theta)$ is then defined as:

$$\mathscr{L}_{\text{TRI}}(\mathcal{T};\theta) = \frac{1}{|\mathcal{T}|} \sum_{(\boldsymbol{f}_A,\boldsymbol{f}_P,\boldsymbol{f}_N)\in\mathcal{T}} \Big[0,\, \mathrm{d}\big(\boldsymbol{f}_A,\, \boldsymbol{f}_P\big) - \mathrm{d}\big(\boldsymbol{f}_A,\, \boldsymbol{f}_N\big) + \alpha\Big]_+, \tag{2.15}$$

where $\mathrm{d}(\,\cdot\,,\,\cdot\,)$ is the distance function and α is a margin. Triplet loss incorporates the margin α to significantly enhance the learning of discriminative features, playing a vital role in deep metric learning scenarios. The margin serves as a buffer zone, establishing a necessary gap between positive and negative pairwise feature distances in the feature space. It ensures that the model goes beyond simply distinguishing between similar and dissimilar samples; it encourages the model to maintain a clear and robust distinction. This clear separation is crucial for the model's ability to generalize well to unseen data, thereby improving its overall performance and reliability. Without a margin, there is a risk of the model collapsing into trivial solutions, where it might map diverse inputs to similar feature vectors, leading to poor differentiation. The margin forces the model to learn more complex and representative features, avoiding such trivial mappings. Additionally, this margin is a powerful tool in preventing overfitting, as it requires the model to focus on learning meaningful patterns in the data, rather than fitting to noise or minor, irrelevant variations.

A significant challenge in training with such loss functions lies in the selection of suitable pairs, as highly dissimilar images contribute minimally to the training process. To address this issue, several approaches have been proposed, *e.g.*, the work of Shrivastava *et al.* [45] introduced an online hard mining algorithm for object detection. Hermans *et al.* [46] then adapted this approach to the domain of FR.

2.2.2.2. Classification

In contrast to deep metric learning, the classification task aims to assign a class label y to a given input image. The most common loss function for multi-classification training of ANNs is the cross-entropy loss function. After connecting the feature extraction network f_θ with the classification network c_θ (see Figure 2.6) in series, the cross-entropy loss function is used to train the whole network end-to-end. Considering a set of facial images $\mathcal{X}$ with L corresponding one-hot encoded class label vectors $\hat{\boldsymbol{y}}$, according to the maximum likelihood estimation, the output probability vector by the classifier ANN

for a specific class is then maximized by minimizing the negative log-likelihood, which results in the cross-entropy loss or *log loss*:

$$\mathscr{L}_{\text{CE}}\big(\mathcal{X};\,\theta\big) = -\frac{1}{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\sum_{l=1}^{L}\hat{\boldsymbol{y}}_i^{(l)}\log \mathrm{c}_\theta\Big(\mathrm{f}_\theta\left(\boldsymbol{I}_i^{(l)}\right)\Big)\,, \tag{2.16}$$

where $\hat{\boldsymbol{y}}_i^{(l)}$ denotes the ground truth class label of the i-th image from the l-th class and $\boldsymbol{y}_i^{(l)}$ is the posterior-probabilities of the i-th image for the l-th class.

Training a network with the cross-entropy loss function is a common practice in the field of FR and has been successfully applied in various works [40, 43, 44, 47]. The cross-entropy loss function is moreover a powerful tool for training the feature extraction network, as it allows the network to learn discriminative features for the classification task. In contrast relative comparison of features in deep metric learning, the cross-entropy loss function aims for a common feature representation. Hence, it can also be used for direct classification tasks, where the network is trained to predict the class label of an input image.

2.2.3. Optimization

Training a ANN involves a non-convex optimization problem where finding the minimum of the loss function cannot be analytically determined. The state-of-the-art method for training ANNs, which utilizes specific loss functions, is through the technique of *backpropagation*. Although the foundational concepts of backpropagation were initially developed by Henry Kelley and Arthur Bryson in the early 1960s [48, 49], it was not until the 1980s that this algorithm was widely applied to ANNs in their current form [50]. The predominant algorithm incorporated today for optimizing ANNs is *Stochastic Gradient Descent* (SGD). In practical scenarios, computing gradients for the entire dataset simultaneously is generally impractical for large datasets due to the extensive memory requirements. Consequently, the gradients for a network function f_θ are computed using randomly selected subsets of the dataset (mini-batches) $\mathcal{B} \subset \mathcal{X}$:

$$\nabla_\theta \mathscr{L}\big(\mathcal{B};\,\theta\big) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\nabla_\theta \mathscr{L}\big(\mathcal{B}^{(i)};\,\theta\big)\,, \tag{2.17}$$

where $\mathcal{B}^{(i)}$ denotes the i-th sample in the mini-batch $\mathcal{B}$. In each iteration, the parameters θ are updated accordingly:

$$\theta \leftarrow \theta - \eta\nabla_\theta \mathscr{L}\big(\mathcal{B};\,\theta\big)\,, \tag{2.18}$$

where η denotes the learning rate, which controls the step size of the gradient descent. A small learning rate η results in gradual parameters updates, often leading to slow convergence towards the minimum of the loss function. Conversely, a large learning rate η may cause the updates to overshoot minima, resulting in oscillations around the minimum. Exposing the network to the same data repeatedly across multiple *epochs* is beneficial for navigating the complex loss landscape inherent to high-dimensional optimization problems. This iterative learning process helps the network avoid becoming trapped in local minima or saddle points, thereby enhancing its ability to find more globally

optimal solutions. Nowadays, optimization approaches utilize more sophisticated techniques, such as SGD with momentum, which uses the exponential moving average over all previous gradients to enhance the effective learning rate, η, when all gradients align in the same direction, and to reduce η in the presence of oscillating gradients. Another widely used technique is the *Adaptive Moment Estimation* (ADAM) [51] optimizer. It is an algorithm for gradient-based optimization that adaptively estimates lower-order moments to improve convergence and is used for training networks in the scope of this dissertation. However, this work does not provide detailed coverage of several other proposed extensions such as *Adaptive Gradient Algorithm* (ADAGRAD) [52] or *Root Mean Square Propagation* (RMSProp) [53]. These methods have been shown to improve the convergence speed and stability of the optimization process, particularly in the context of DL.

2.3. Face Verification

FV is a subfield of FR. Typically, FR encompasses more than just the stages of verification or identification; it also includes the detection and preprocessing of faces. When examining the phase centered on identity, it is essential to distinguish between two primary tasks. Face identification is the task of assigning an identity label to an image of person, not necessarily known, whereas FV is the task of deciding whether two facial images belong to the same person or not. In other words, FV is a binary classification task. In the following, first, the general approach of FV is described, then the most prominent training and testing datasets in the context of FV are introduced. Finally, the most common evaluation metrics are presented.

2.3.1. Cross-Resolution

In general, w.r.t. image resolution one can differentiate between two types of FV tasks: *Equal Resolution* (ER) and *Cross Resolution* (CR) FV.

The term ER describes the task of FV with images of the same raw-resolution. HR images provide a wealth of details, making it easier to extract fine-grained facial features, which are crucial for accurate FV. While HR FV is ideal, in real-world scenarios, obtaining HR images is not always be possible, *e.g.*, in surveillance or mobile environments. LR images are common in surveillance footage, where many cameras do not capture high-quality images due to large distances to the face. A primary challenge in using LR FV is the difficulty in extracting reliable facial features, stemming from the inherent lack of detail. This limitation can significantly impair the accuracy of the verification process.

The term CR is then introduced to describe the task of FV with images of different raw-resolutions, *e.g.*, one image of high raw-resolution and one image of low raw-resolution. CR FV is highly relevant in real-world applications, offering a more flexible approach to FV across varied conditions and sources.

2.3.2. General Approach

Nowadays, state-of-the-art FV approaches are based on DL. This section aims to give a brief overview of the general approach of FV using DL. First, the most recent and widely used DL architectures

for FV are introduced. Then, the preprocessing steps, including face detection, and alignment are introduced. Finally, the training and testing phase are described.

2.3.2.1. Deep Learning Architectures

Nowadays, deep learning methods in the field of FR are based on two main architectures, CNNs and ViTs, which are explained in the following:

Convolutional Neural Network (CNN) architectures include a variety of different network architectures. Early works [19, 40] used deep CNN architectures to learn discriminative features from facial images. These architectures are typically composed of multiple convolutional layers, which are then stacked on top of each other. A convolutional layer is a special form of ANN that processes data through a grid-like topology, primarily designed to recognize patterns in images by applying filters that capture spatial hierarchies and features. Further development of CNN-based approaches lead to the introduction of residual networks, such as *Residual Layer Network* (ResNet) [54] and ResNet-V2 [55], which are still widely used in FR. These architectures introduce a novel approach to mitigating the vanishing gradient problem encountered in the training of deep neural networks, facilitating the construction and effective training of networks with substantially increased depth. Through the incorporation of residual blocks that implement shortcut connections, ResNet allows the direct propagation of gradients from deeper to shallower layers by enabling the addition of the input to a block directly to its output. This design principle aids in preserving the strength of the gradient throughout the network, thereby enhancing the learnability of deep architectures. Additionally, it empirically demonstrates significant performance enhancements across a variety of complex computer vision tasks. This evidences the capacity of deeper networks to achieve superior learning outcomes when equipped with mechanisms to effectively bypass the challenges of depth. Prominent works utilizing the ResNet approach include [40, 43, 44, 47, 56, 57] and are still considered state-of-the-art. Although, CNN-based approaches have led to remarkable performance in FR, they still suffer a critical limitation when capturing long range relations among facial regions.

Vision Transformer (ViT), introduced by Dosovitskiy *et al.* [41] addresses this issue. They applied the principles of transformers, originally designed for natural language processing tasks, to computer vision by treating images as sequences of patches. These patches are linearly embedded, alongside positional encodings, to maintain spatial information, and then processed through a series of transformer blocks that leverage self-attention mechanisms to capture global dependencies within the image. This architecture allows the network to dynamically adjust the focus on different parts of the image, facilitating a more flexible and comprehensive understanding of visual data compared to traditional convolutional approaches, and has demonstrated remarkable performance on various image recognition tasks. Recent works [58, 59, 60, 61] show promising results in the field of FR using ViT architectures.

Both main architectures, CNNs and ViTs, are employed in the feature extraction part of the FR pipeline. These features represent an intermediate state in a bigger network, which is then used to classify the identity of a person from the given image. In general, an FR model can be divided into two parts: The feature extraction and classification as shown in Figure 2.6.

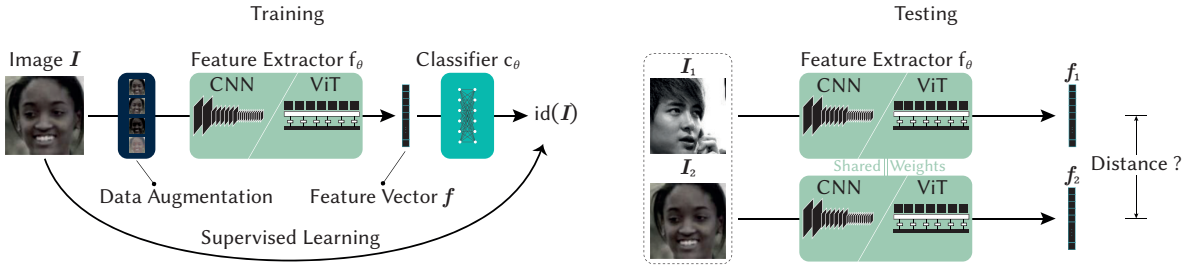


Figure 2.6.: General Approach of face recognition. The process of typical training illustrated on the left. The testing/application of the trained model shown on the right. Sample images from MS1M-V2 dataset [47, 62].

2.3.2.2. Preprocessing

Facial images are often taken in uncontrolled settings and are loosely cropped, *i.e.*, they may include backgrounds or other faces. While joint face detection and recognition has been suggested in various studies *e.g.*, [63, 64, 65], its effectiveness is not yet at an ideal level. To reduce the impact of background distractions, most FR algorithms first detect faces to help the FR system to concentrate solely on recognizing the face. Furthermore, the alignment or spatial normalization of the detected faces, enhances training speed and overall performance. This alignment is feasible because all human faces have roughly similar proportions and the position of facial features like eyes, nose, and mouth is consistent. By standardizing image resolution, there is no need for algorithms to handle varying input resolutions. As facial sizes and distances are relatively uniform, the network can avoid learning a multitude of filter combinations for different scales. This reduces redundancy in the network and allows it to focus on more critical features for identity recognition. The whole process of aligning an image $\boldsymbol{I}$ to an aligned image $\boldsymbol{I}^*$ is illustrated in Figure 2.7.

This work follows recent work [47, 57, 66, 67] and use the most prominent face detection algorithm *Multi-Task Cascaded Convolutional Networks* (MTCNN) introduced by Zhang *et al.* [68]. The MTCNN is a cascaded CNN, which is able to detect faces and facial landmarks in an image. The MTCNN is composed of three stages. The first stage is a proposal network, which is trained to propose regions of interest. The second stage is a refinement network, which is trained to refine the regions of interest. The third stage is a output network, which is trained to output the bounding boxes and facial landmarks.

After facial landmark detection, the face is aligned. There are two main approaches to align a face, *i.e.*, mapping the retrieved facial landmarks to the landmarks of a face template: Affine and rigid transformation. The difference between the two is that affine transformation includes translation, scaling, shearing, and rotation, whereas rigid transformation only includes scaling, rotation and translation. The aligned face is then cropped and resized to a fixed spatial resolution. All methods in this work use a spatial input image size of 112×112 px and to avoid clutter in the following, all facial images $\boldsymbol{I}$ in this work are assumed to be cropped and aligned. Note that the facial landmark detection and alignment is applied in both training and testing scenarios.

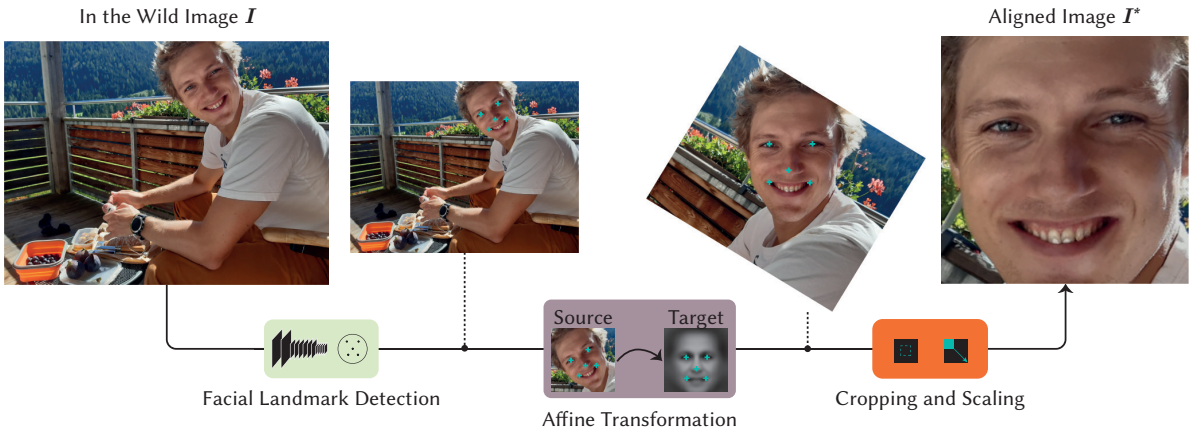


Figure 2.7.: Flowchart of a typical face recognition pre-processing pipeline including face detection and alignment.

Additionally, data augmentation techniques are commonly applied in the field of FR to increase the variance of the training samples. These techniques include, *e.g.*, horizontal flipping, adjusting brightness, saturation, and contrast, as well as applying Gaussian blur, motion blur, noise injection, and compression, which further enhance the robustness of the network.

2.3.2.3. Training

The primary objective when training an FR network is to cultivate a distinctive feature representation for each identity, as illustrated in Figure 2.6. To facilitate this, a classifier, typically a fully connected layer within an ANN, is integrated following the feature extractor network. This classifier is trained end-to-end with a dataset comprising images and their corresponding class labels. During this process, the feature extractor network indirectly learns to map the input images into a designated feature space, which is often 512-dimensional. Within this space, the representations of identical identities are engineered to converge, whereas those of dissimilar identities are positioned distantly. This division of the network into distinct segments *i.e.*, the feature extractor and the classifier, is strategically advantageous. The feature extractor, in particular, offers versatile applicability across various FR tasks beyond mere identity classification, *e.g.*, emotion or attribute classification. However, it is pertinent to note that while the entire system is trained at classifying known identities, *i.e.*, those included within the training dataset, it lacks the capability to recognize unknown identities. This limitation underscores the critical role of the feature extraction, which is foundational for any extension of the model to accommodate new, previously unseen identities.

A key aspect of this architecture is a bottleneck between feature extraction and classification. The dimension is typically much smaller than the amount of classes, forcing a dense feature representation

of identities. This bottleneck is crucial for the network to learn a compact representation of the input data, which is essential for the network to generalize well to unseen data.

In the last years, promising approaches such as ArcFace [47], CosFace [56], GroupFace [57] have emerged, which continue to be considered state-of-the-art and are widely used. For a comprehensive overview of these approaches, the interested reader may refer to the work of Hörmann [69].

2.3.2.4. Testing/Application

Through end-to-end training of the entire system, the feature extractor learns to identify and encode features specific to different identities, transforming unknown identities into unique feature vectors. By comparing these vectors, it is possible to determine their similarity and whether they represent the same or different identities.

During testing or application/deployment, the model is used to decide whether two faces belong to the same person or not. This is a binary classification task and typically utilized by taking the features of the feature extraction part of the network and compare them using a distance metric (see Section 2.2.2.1).

The model is trained using augmented, aligned images, while only aligned faces are used during testing to guarantee a deterministic evaluation. Nevertheless, some FR approaches [56, 70] perform horizontal flipping and concatenate the feature vectors of the flipped and original image to slightly improve the performance. However, in this work, no data augmentation is employed at test time.

2.3.3. Datasets

In this section, a brief overview of the most common datasets used in the field of FR is given. As introduced in Section 2.2.1, supervised learning is based on optimizing a correspondence between input images and labels. In the context of FR, this involves images of faces along with their corresponding identity labels. It should be noted that the term ‘dataset’ is also referred to as ‘database’ and is used interchangeably throughout this work. Using a dataset for testing is also termed ‘benchmark’. In training datasets, identity labels are leveraged to train the network, whereas in testing datasets, these labels are used to form image pairs that either belong to the same person or to different individuals. These datasets can be categorized into two types based on their usage: Training datasets and benchmark datasets.

2.3.3.1. Training

Training datasets are used solely to train the network and can be quantitatively characterized by the size, the deepness or shallowness, and the balance of the dataset. Size in this context refers to the total amount of images in the dataset, and can be further distinguished into the number of identities and the number of images per identity. Deep means that the dataset contains a lot of images per identity, whereas shallow means that the dataset contains only a few images per identity. Together with the balance, which refers to the distribution of the number of images per identity, these characteristics are

Table 2.1.: Overview and key statistics of widely used training datasets in the field of face recognition.

Dataset	Year	# Identities	# Images	Availability[ii]	# Images / ID		
					Min	Avg	Max
WebFace260M [74]	2021	4 M	260 M	Public	–	65	–
WebFace42M (cleaned) [74]	2021	2 M	42 M	Public	–	21	–
CelebFaces+ [19]	2014	10 k	203 k	Public	1	20	35
CASIA-WebFace [75]	2014	10 k	494 k	Public	2	47	804
UMDFaces [76]	2017	8 k	368 k	Public	–	45	–
VGGFace [77]	2015	2 k	2.6 M	Public	1 000	1 000	1 000
VGGFace2 [78]	2018	9 k	3.3 M	Public	87	364	843
MS1M [62]	2016	100 k	10 M	Public	–	100	–
MS1M-V2 [47]	2019	86 k	5.8 M	Public	2	68	602
MS1M-IBUG [79]	2017	85 k	3.8 M	Public	–	45	–
MS1M-Glint [80]	2018	87 k	3.9 M	Public	–	44	–
Asian-Celeb [80]	2018	94 k	2.8 M	Public	–	30	–
MegaFace2 [81]	2017	600 k	4.7 M	Public	–	7	–
Glint360k [82]	2021	360 k	17.1 M	Public	–	47	–
IMDB-Face [83]	2018	59 k	1.7 M	Public	–	29	–
Facebook [18]	2014	4 k	4.4 M	Private	–	1.1 k	–
Facebook [84]	2015	10 M	500 M	Private	–	50	–
Google [44]	2015	8 M	200 M	Private	–	25	–
MillionCelebs [85]	2020	600 k	18.8 M	Private	–	30	–

important to ensure that an FR system is robust and generalizes well. Bansal *et al.* [71] investigated the impact of deep and wide datasets on the performance of FR systems. They found that it is better to use deeper datasets for deeper networks, while wider datasets are preferred for shallower networks.

Additionally, the analysis from Zhou *et al.* [72] shows that adding more identities to a given dataset boosts performance, but only if the added identities contain a certain number of images. Zhang and Deng [73] demonstrated that a dataset with a uniform distribution performs better than datasets with imbalanced samples, where some identities have few samples and others have many. Therefore, it is crucial for a dataset to have a sufficient number of images for each identity and to maintain a uniform distribution of images per identity. Table 2.1 provides an overview of the most common training datasets used in the field of FR and there quantitative characteristics.

To ensure the effectiveness of FV systems, it is important to also consider qualitative characteristics of the dataset. Noisy data, such as images with occlusions, LR, or poor lighting, can significantly impact the performance of the system. Additionally, it is important to address bias and ensure that the distribution of the training dataset aligns with the distribution of the test dataset. *E.g.*, when evaluating the recognition of children, the training dataset should include images of juvenile faces. Other qualitative characteristics include the racial and gender diversity of the dataset.

While the gender distribution is relatively balanced in VGGFace2 (59.3% male versus 40.7% female [86])[iii], the ethnicity distribution is substantially biased towards Caucasians (74.2%) with Asians (6.0%) Indians (4.0%) and Africans (15.8%) being underrepresented [87]. This imbalance is even

[ii] As of writing in February 2024
[iii] Assuming binary genders

more pronounced in the *Microsoft 1 Million* (MS1M) [47, 62] dataset. This should be kept in mind when evaluating the performance of an FR system.

While the gender distribution in VGGFace2 is relatively balanced, with 59.3% male and 40.7% female [86][iv], the ethnicity distribution shows a substantial bias towards Caucasians (74.2%), leaving Asians (6.0%), Indians (4.0%), and Africans (15.8%) underrepresented [87]. This imbalance is even more pronounced in the MS1M dataset [47, 62]. Such disparities should be considered when evaluating the performance of an FR system.

Another important aspect of the dataset is the retrieval of the images and thus the quality or correctness of the labels. Due to large scale automatic web scraping techniques to collect the images, the quality of the images and the correctness of the labels is not always guaranteed. This is especially true for the MS1M dataset, which is a large scale dataset with 5.8 million images of 85 thousand identities. The dataset is collected from the web and thus contains a lot of noisy data. The dataset is also known for containing many duplicate images. This is a common problem in large scale datasets, as the same image can be found on different websites. Celebrities often exemplify this phenomenon, as they are frequently photographed by different individuals and their images subsequently appear on various websites.

Wang *et al.* [83] demonstrated that large datasets are very susceptible to label noise. They estimated that the MS1M dataset contains $\sim$ 50% label noise, making it challenging to train an FR system on this dataset. Furthermore, Bansal *et al.* [71] and Wang *et al.* [83] have both confirmed that the presence of noise in a dataset negatively impacts the performance of FV systems. This finding highlights the importance of developing robust training strategies to mitigate the effects of noise. The MS1M dataset, although acknowledged for its high level of noise, presents an opportunity to address this challenge.

Jin *et al.* [88] and Wang *et al.* [83] have shown that cleaning datasets, particularly MS1M, leads to better outcomes. Furthermore, Deng *et al.* [47] enhanced the label quality of challenging samples in MS1M by using annotators familiar with the ethnicity, resulting in the *Microsoft 1 Million Version 2* (MS1M-V2) dataset. This dataset has remained the most popular training dataset for FR in recent years, as evidenced by its use in several studies [57, 66, 67, 89, 90, 91] and is utilized in this dissertation for a fair comparison.

Recently, very large datasets such as Glint360k [82] from Beijing Geling Shentong Information Technology Co., WebFace260M [74], and WebFace42M (cleaned) [74] are introduced aiming to close the gap between academia and industry. However, their analysis is out of scope for this work.

Several non-public datasets from companies like Facebook, Google, and Microsoft are also used for training and listed in the bottom section of Table 2.1. However, these datasets are not publicly available and are not further discussed in this work.

[iv] Assuming binary genders

2.3.3.2. Testing

Benchmark datasets are used to evaluate the performance of an FR system. The task of FV (see Section 2.3) is a binary classification task, and thus the testing datasets contain defined pairs of images belonging to the same (genuine) or different (imposter) identities. Typically, every benchmark dataset was created to investigate a particular task. Table 2.2 gives an overview of the most common benchmark datasets used in the field of FR and their characteristics.

The *Labeled Faces in the Wild* (LFW) dataset [31], initially released in 2007, was the first dataset to include images taken in uncontrolled (in the wild) environments and was widely used to measure FV performance. However, the hight accuracy (99.83%) of FV systems on this dataset has reached saturation and is now considered relatively easy to achieve. One contributing factor to this high accuracy is that imposter pairs often have different gender and ethnicity, while genuine pairs have the same gender and ethnicity. Additionally, the age gap between faces in genuine pairs is usually smaller than the age gap between faces in imposter pairs. The overall quality and image resolution of the dataset is also very high and the faces show only little variation in pose.

In response to this, new datasets were created to address these limitations. *E.g.*, *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92] and *Celebrities in Frontal-Profile — Frontal-Profile* (CFP-FP) [93], evaluate FV performance under varying head poses. Sengupta *et al.* [93] also released the *Celebrities in Frontal-Profile — Frontal-Frontal* (CFP-FF) dataset, which does not contain any pose variations. The *Cross-Age Labeled Faces in the Wild* (CALFW) [94] and *Age Database* (AgeDB) [95][v] focus on ensuring a similar age gap among genuine and imposter pairs. AgeDB addresses a huge variety of age from 1 to 101 years. To measure the robustness of FV approaches against variation of image raw-resolution and quality, QMUL-SurvFace [96] dataset was introduced by the Queen Mary University of London.

In Chapter 5 the *Cross-Quality Labeled Faces in the Wild* (XQLFW) dataset is introduced, which is based on the LFW dataset and designed to address the image quality and raw-resolution limitations of the LFW dataset. Deng *et al.* [97] introduces the *Similar-Looking Labeled Faces in the Wild* (SLLFW) dataset, which contains images of similar looking people as imposter pairs, making the task of FV more challenging. Considering adversarial attacks, the *Transferable Adversarial Labeled Faces in the Wild* (TALFW) dataset [98] was introduced to evaluate the robustness of FV systems against adversarial attacks. To address the issue of racial bias in FR systems, the *Racial Faces in the Wild* (RFW) dataset [87] was introduced, which contains images of people from different racial groups. Very recently, more and more people wearing face masks due to the COVID-19 pandemic and the *Masked Labeled Faces in the Wild* (MLFW) dataset [99] was introduced to evaluate the robustness of FV systems against face masks.

While these datasets serve their purpose and are widely used, their relatively low number of pairs restricts the performance analysis. To mitigate this, larger datasets, such as MegaFace [100] and TrillionPairs [80] have been introduced, which are used to evaluate the performance of FR systems. However, the analysis of FV performance on the latter datasets is out of scope for this work. Other benchmark datasets, such as the *Intelligence Advanced Research Projects Activity Janus Benchmark*

[v] In this work the AgeDB-30 protocol is used, *i.e.*, an age gap of 30 years

Table 2.2.: Overview and statistics of popular face verification benchmark datasets.

Dataset	Year	# Identities	# Images	# Pairs	Description
LFW [31]	2007	4 281	7 701	6 000	–
CALFW [94]	2017	2 997	7 167	6 000	Age Gap
CPLFW [92]	2018	2 296	5 984	6 000	Pose Variation
XQLFW [3$^+$]	2021	3 743	7 263	6 000	Quality
SLLFW [97]	2017	2 810	6 091	6 000	Similar Looking
MLFW [99]	2022	2 997	12 000	6 000	Masked Faces
RFW [87]	2019	11 430	40 607	24 000	Racial Diversity
AgeDB [95]	2017	568	12 000	6 000	Age Gap
CFP-FP [93]	2016	500	7 000	7 000	Frontal-Pose
CFP-FF [93]	2016	500	7 000	7 000	–
YTF [104]	2011	1 447	3 226	5 000	Faces from Youtube Videos
QMUL-SurvFace [96]	2018	5 319	10 051	10 638	Surveillance Uncontrolled

(IJB) datasets IJB-A [101], IJB-B [102], and IJB-C [103] are focusing on the face identification task, thus not being further discussed in this dissertation.

2.3.4. Evaluation

In general, an FR system can be evaluated with a left-out subset of the training dataset, *i.e.*, using the same identities, but strictly assuring the network has not seen those images during training. This then evaluates the whole system including the classification part. However, the main goal of FR is to be able to compare unknown identities. Therefore only the feature extractor network is evaluated.

As briefly mentioned in the beginning of this chapter, the evaluation of FR systems is typically done using the FV task and unknown identities.

After extracting the features of two images the distance between the two feature vectors is calculated (see Section 2.2.2.1). Besides the cosine distance (see Equation (2.11)), the Euclidean (L2) distance is also widely used. Independently from the metric, the distance is then compared to a certain threshold. If the distance is below the threshold, both images are classified as the same person, otherwise as different persons. Since the labels of the image pairs in the testing data are known, using the optimal threshold would be somewhat unfair. Thus, a 10-fold cross validation is usually applied, *i.e.*, the testing data is split into 10 equally distributed folds, and the threshold is optimized for each fold, user the data of the other nine folds. This results in 10 folds of testing data, with each fold having a different threshold. The metrics are then averaged over the 10 folds.

To measure the performance of FV approaches, the following metrics are commonly used:

Accuracy is the ratio of the number of correct predictions to the total number of predictions. It is defined as:

$$\text{ACCURACY} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2.19}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

False Acceptance Rate (FAR) is the ratio of the number of false positives to the total number of negative samples. It is defined as:

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} . \tag{2.20}$$

False Rejection Rate (FRR) is the ratio of the number of false negatives FN to the total number of positive samples. It is defined as:

$$\text{FRR} = \frac{\text{FN}}{\text{FN} + \text{TP}} . \tag{2.21}$$

True Acceptance Rate (TAR) is the ratio of the number of true positives TP to the total number of positive samples. It is defined as:

$$\text{TAR} = \frac{\text{TP}}{\text{TP} + \text{FN}} . \tag{2.22}$$

Receiver Operating Characteristic (ROC) is a curve plot of the TAR (sensitivity) against the FAR (specificity) for different threshold values, providing a tool to assess the trade-offs between benefiting from TPs and suffering from FPs. It is utilized for comparing the diagnostic ability of classifiers under varying thresholds, facilitating the identification of an optimal balance between detecting TPs and avoiding FPs.

Equal Error Rate (EER) is the point on the ROC curve, where the FAR is equal to the FRR. It serves as a valuable indicator of the overall accuracy of a biometric system, providing a single measure to compare the performance of different systems. An EER is particularly useful for evaluating systems where the cost of FPs is comparable to that of FNs, as it offers a balanced metric for assessing system reliability.

2.3.5. Performance Overview

To conclude this chapter, Table 2.3 presents a brief overview of the performance of the most prominent FV approaches. This also serves to classify the methods presented in this work [$1^{\dagger}$, $2^{\dagger}$, $5^{\dagger}$] within the state-of-the-art. The performance is measured in terms of accuracy (see Equation (2.19)) and is reported on the most common benchmark datasets.

It is evident, that the algorithms' capabilities have surpassed humans on LFW by a substantial margin for many years [31]. In addition to their superior efficiency in processing images, algorithms have demonstrated clear advantages over human labeling in FR datasets. These results highlight the need to employ alternative approaches, such as using ethnicity-specific annotators [47] or super-recognizers [25], to minimize label noise and improve dataset quality.

In the recent years, research shifted towards ViT architectures and more and more methods such as [$2^{\dagger}$, 59, 60, 91, 108, 109, 111] are based on ViT architectures.

The table clearly shows, why dataset LFW is saturated and provides no more insights, and the need for other more challenging dataset is there. On datasets besides LFW, the differences between methods and the substantial improvements in the last years become apparent. And also the strength

Table 2.3.: Face verification accuracy of state-of-the-art face recognition methods on various benchmark datasets. Numbers marked with * are calculated using (re)-implementations.

		Verification Accuracy [%]					
Method	Year	LFW [31]	CALFW [94]	CPLFW [92]	XQLFW [3†]	CFP-FP [93]	AgeDB [95]
CurricularFace [66]	2020	99.80	96.05	93.13	–	98.36	98.37
BroadFace [67]	2020	99.85	96.20	93.17	–	98.63	98.38
CosFace [56]	2018	99.73	95.76	92.28	–	98.12	98.11
GroupFace [57]	2020	99.85	96.20	93.17	–	98.63	98.28
CenterLoss [70]	2016	99.28	85.48	77.48	–	–	90.72
SphereFace [105]	2017	99.42	90.30	81.40	–	–	–
VGGFace2 [78]	2018	99.43	90.57	84.00	–	–	–
ArcFace VPL [106]	2021	99.83	96.12	93.45	–	99.11	98.60
ArcFace + OLT [2†]	2023	99.55*	96.03*	92.75*	93.27*	93.19*	93.42*
MagFace [89]	2021	99.83 / 99.63*	96.15	92.87	76.95*	96.19	97.82
ArcFace [47]	2019	99.83 / 99.50*	95.45 / 93.85*	92.08 / 88.37*	74.22*	98.27 / 92.27*	98.15 /95.10*
ArcFace BTM [1†] (RAT)	2019	99.30*	94.10*	86.70*	83.60*	91.82*	90.90*
ProdPoly [107]	2020	99.83	96.23 / 96.03*	93.32 / 92.75*	86.90*	98.99	98.47
ArcFace ST-M1 [1†] (CLT)	2019	97.30*	83.85*	82.60*	90.97*	90.37*	81.43*
ArcFace ST-M2 [1†] (MB-CLT)	2019	95.87*	83.85*	82.60*	90.82*	90.37*	81.43*
FaceTransformer [91]	2021	99.80 / 99.70*	94.93*	91.58*	87.88*	–	–
Part fViT-B [59]	2022	99.83	–	–	–	99.21	98.29
CFormerFaceNet [60]	2023	99.75	95.73	90.20	–	–	97.12
SwinFace [108]	2023	99.87	96.10	93.42	–	98.60	98.15
TransFace-L [109]	2023	99.85	–	–	–	99.32	98.62
FaceTransformer + OLT [2†]	2023	99.73*	94.65*	91.38*	95.12*	97.21*	96.05*
🤖 (Reported in [110])	2009	97.53	–	–	–	–	–
🤖 + FaceTransformer + OLT [5†]	2023	–	94.93*	91.73*	95.28*	–	–

and weaknesses of those methods. With a focus on CR FR it is evident, that there are huge differences in performance between the methods on the XQLFW dataset.

The analysis further reveals that the ArcFace + *Resolution Augmentation Training* (RAT), ArcFace + *Contrastive Loss Training* (CLT), and ArcFace + *Multi-Branch Contrastive Loss Training* (MB-CLT) methods show significant decrease in performance on all other datasets than XQLFW. However, their improvement compared to their baseline model (ArcFace) is substantial. The ArcFace + *Octuplet Loss Training* (OLT) approach is then putting a step on top and improves drastically XQLFW performance, while yielding considerably accuracy on all other datasets.

“The important thing is not to stop questioning. Curiosity has its own reason for existing.”

— Albert Einstein

Preliminary Analysis: Image Resolution Susceptibility in Face Verification

The majority of *Face Recognition* (FR) systems are trained and tested on *High Resolution* (HR) images. However, in practical applications, the image resolution can vary significantly due to different image capture mechanisms or sources. This chapter first investigates the differences of reduced and HR images on the image pixels itself. Then, it analyzes the impact of image resolution on the verification performance of a state-of-the-art FR model on various synthetically downscaled datasets. Note, that the analysis is done with post-preprocessing, *i.e.*, the images are already aligned and cropped before being synthetically deteriorated. Here the *Equal Resolution* (ER) and *Cross Resolution* (CR) scenarios are considered. Furthermore, the feature distances for every 2-image test pair are analyzed. The chapter concludes with limitations and a short discussion of the implications of the results. Note that the experiments and key findings in this chapter are based on the publication of Knoche *et al.* [1†].

3.1. Related Work

Early works [112] investigated the impact of image resolution on classical FR (non-*Deep Learning* (DL)) systems. Their findings are considered limited, due to the fact that the classical systems are very different from state-of-the-art DL approaches and therefore the results may not be directly transferable. Li *et al.* [113] did a comprehensive analysis on *Low Resolution* (LR) FR for the ER scenario. However, they do not provide a detailed analysis the behavior of FR models on decreasing image resolution. In [114] the authors also focused on an analysis of LR FR and proposed a *Single Image Super Resolution* method to mitigate the effect of LR. However, they did not consider the CR scenario. Marciniak *et al.* [115] illustrate the influence of image resolution on the *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR) of FR systems. However they do not provide results on the accuracy (accuracy) of the system. Moreover, they analyzed the effect of different light emission angles and pose variations.

3.2. Pixel-Level Differences

To get a better insight of what exactly happens to the content of a facial image when performing the synthetical reduction of image resolution, this section presents an experiment from [1†] to analyze the

pixel-level differences between HR and LR images. Therefore, synthetic downscaling is applied to all images of the *Labeled Faces in the Wild* (LFW) [31], *Cross-Age Labeled Faces in the Wild* (CALFW) [94], *Celebrities in Frontal-Profile — Frontal-Profile* (CFP-FP) [93], *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92], and *Age Database* (AgeDB) [95]. Afterwards, the average image $\overline{\boldsymbol{I}}$ for each dataset $\mathcal{X}$ is calculated as follows:

$$\overline{\boldsymbol{I}} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{I} \in \mathcal{X}} \boldsymbol{I} \,. \tag{3.1}$$

Additionally the mean pixel differences between the HR and LR images are computed. The raw-resolution reduction is performed using downscaling $\downarrow_s(\cdot)$ to four different image dimensions with $s \in \{16, 8, 4, 2\}$, followed by upscaling $\uparrow_s(\cdot)$ them back to their original resolution of 112×112 px using bicubic interpolation, to retain the spatial image resolution but alter the raw-resolution. The overall process can be formulated per dataset $\mathcal{X}$ as follows:

$$\overline{\boldsymbol{D}}_r = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{I} \in \mathcal{X}} \left(\left| \uparrow_s \left(\downarrow_s (\boldsymbol{I}) \right) - \boldsymbol{I}_i \right| \right), \tag{3.2}$$

with r denoting the used raw-resolution to calculate the difference image $\overline{\boldsymbol{D}}_r$.

Figure 3.1 visualizes the results after calculating the mean difference images. The mean original HR images $\overline{\boldsymbol{I}}$ are quite different across all datasets. One can clearly see that pose variations in CPLFW dataset result in more blurred areas in the image. In contrast, the CALFW and LFW dataset images seem to be very accurately aligned and show a almost clear and detailed average face. Interestingly,

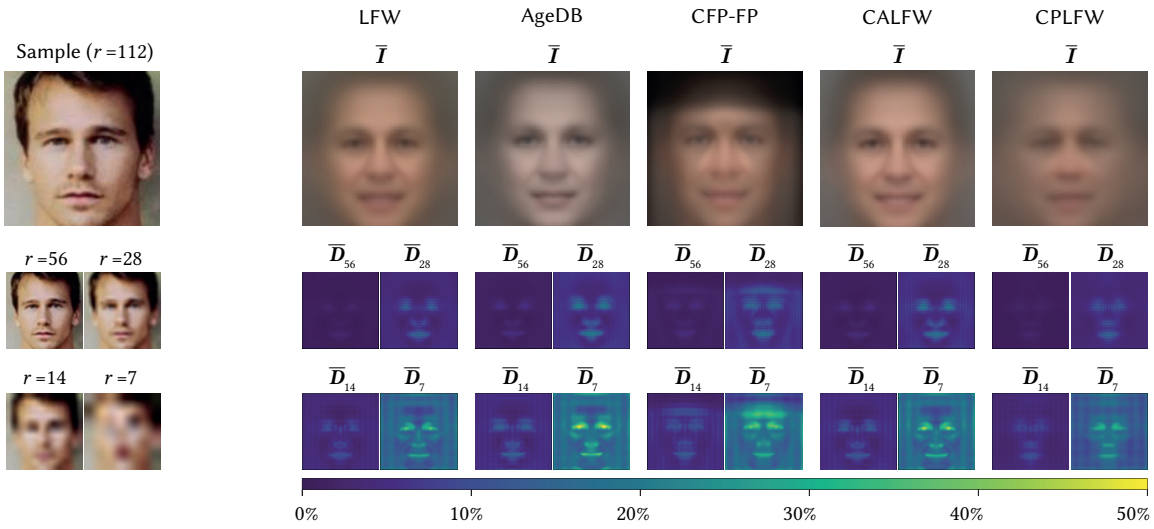


Figure 3.1.: Left: A sample image and it's downscaled variants, taken from *Microsoft 1 Million Version 2* (MS1M-V2) [47, 62]. Right: Illustration of the average mean pixel differences after the resolution-reduction process in comparison to the mean original resolution images for several datasets. Adapted from [1†].

the background in the CFP-FP dataset is very dark compared to other datasets. This is probably due to alignment of faces which are not entirely captured by the camera and thus lead to black image regions after aligning the faces to the target landmarks (see also Section 2.3.2.2). The mean face of the AgeDB dataset is clearly more bright than the other datasets. This is probably due to the fact that the dataset contains more old photos, which are captured in grayscale and thus have a brighter appearance. Also, the pose variation can be seen in the average face of CPLFW, as the outer areas of the face got blurred out.

In the mean difference images, as expected, eye, nose, and mouth regions are heavily affected by the resolution reduction process in all datasets. High detail information in those regions is lost. This is valuable information for FR. The effect is consistent across all datasets. The variation increases for lower resolutions. The maximum deviation of a single LR image pixel concerning its counterpart pixel in the HR image is about 50%. There are slightly visible artifacts in a grid style manner occurring in all pixel-difference images. These might be some aliasing artifacts, which could not entirely be removed by the anti-aliasing method of the applied bicubic interpolation algorithm.

3.3. Face Verification Accuracy

To investigate the effect of image resolution on *Face Verification* (FV) accuracy (see Section 2.3.4), the ArcFace [47] approach was re-implemented with Tensorflow [116] and trained on the MS1M-V2 [47, 62] dataset.

Figure 3.2 visualizes the accuracy for five datasets across different image resolutions (using the same protocol as introduced in Section 3.2) in CR (solid lines) and ER (dashed lines) scenario. CR means that each input pair of images consists of one HR and one LR image. ER indicates that both images in a pair have the same resolution.

Focusing on the accuracy for the maximum resolution of 112 px, the performance on LFW dataset is best with 99.50% accuracy. The CPLFW dataset shows the worst performance with about 88.37% accuracy. The CALFW, AgeDB, and CFP-FP datasets are in between with 93.85%, 95.10%, and 92.27% accuracy, respectively.

Looking at lower resolutions, the accuracy decreases significantly for all datasets. The accuracy is quite consistent across all datasets until about 56×56 px. Then the accuracy drops significantly until it reaches a plateau at about 7×7 px and lower. Note that an accuracy of 50% is the random guess level. It is not straightforward to set a distinct threshold at when the performance exactly starts dropping, since it is a continuous process. However, the *European Norm* (EN) (DIN EN 62676–4:2016 [117]) sets the minimum required pixel resolution for face identification in surveillance applications to one pixel per 4 mm of the face. Assuming an average inter-ocular distance of 63 mm [118] this would then be a raw-resolution of about 48×48 px considering aligned faces used in this work. For detecting a human face in a scene, the minimum required size is given as 40 mm per pixel, which would be a raw-resolution of about 5×5 px. But this in fact is too low to perform any facial landmark detection and thus an alignment at this resolution will not be possible. This should be kept in mind, when considering the image resolution effects as they might be even more significant when considering

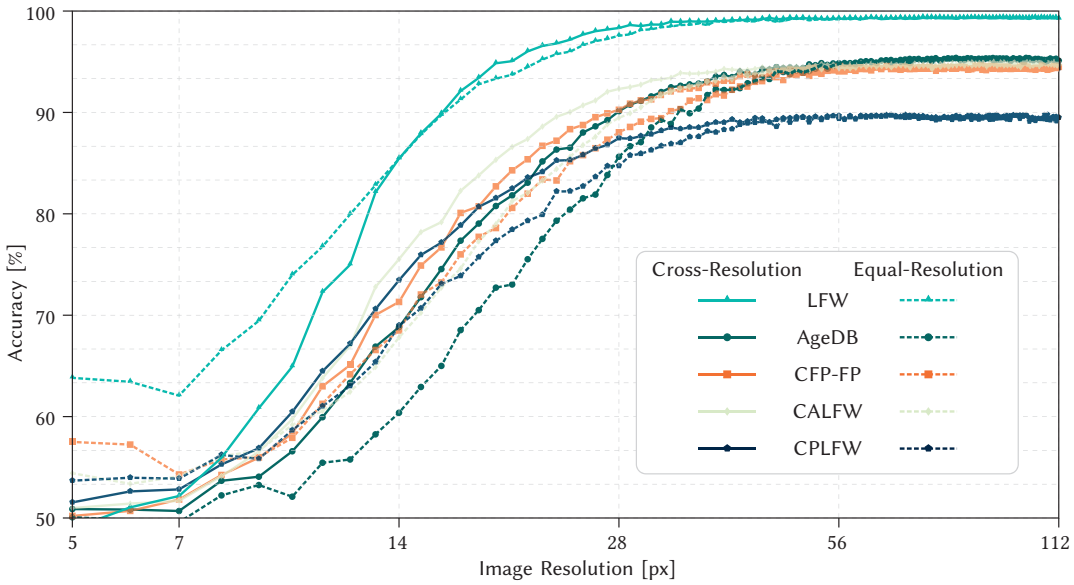


Figure 3.2.: The figure captures the face verification performance of an ArcFace model re-implementation, as adapted from [1†], which has been trained on the MS1M-V2. The model's accuracy is evaluated across a series of synthetically downscaled images from five renowned benchmark datasets, under both cross-resolution and equal-resolution synthetic down-scaling protocols. The datasets included in this analysis are LFW [31], CALFW [94], CFP-FP [93], CPLFW [92], and AgeDB [95]

the complete pipeline including face detection and alignment. Considering the gap between CR and ER accuracy, the LFW dataset has the largest difference, whereas all other datasets show opposite behavior. For a better understanding, what reasons may cause this large decrease of accuracy, a closer look at the extracted features from the ArcFace model is conducted in the next section.

3.4. Feature Distances

As described in Section 2.3.4, the distance between two feature vectors is crucial for the verification accuracy. To analyze the feature distances more in detail, the average cosine distance between the feature vectors of all genuine and imposter image pairs in the LFW dataset is calculated for various raw-resolutionsṪhe results are visualized in Figure 3.3. Here also the CR and ER scenarios are considered.

One can divide the behavior roughly into three sections: 1) For resolutions larger than about $r = 16$ pixels, feature distances between genuine and imposter image pairs seem to be independent of the image resolution. The average distance between genuine pairs is quite low about 0.3 and the distance

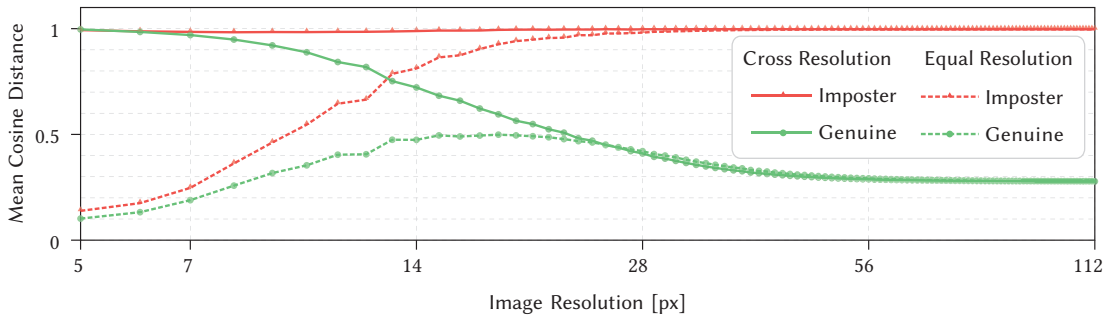


Figure 3.3.: The graphic presents the average cosine distances calculated between the feature vectors of individual images within all genuine and all imposter pairs from the LFW dataset [31], under the protocols of cross-resolution and equal-resolution synthetic down-scaling. These feature were generated using a re-implementation of the ArcFace model, adapted from [1†], trained on the MS1M-V2 dataset.

for imposter pairs is about 1.0. 2) Between resolutions of about $r = 60$ pixels and $r = 20$ pixels, which can be considered as mid-range resolutions, in both CR and ER scenarios, the distance between genuine image pairs tends to increase. However, the distance for imposter image pairs stays roughly at the same level. 3) The last section can be considered as LRs with $r \leq 20$ pixels. Interestingly, distances for both scenarios show a contrary behavior. On the one hand, the mean feature distance for CR genuine pairs is increasing towards 1. This is consistent with the accuracy decreasing towards 50%, which is, in terms of verification, merely guessing. On the other hand, in the ER case, genuine and imposter feature distances decrease towards 0.1. This also coincides with low accuracy scores in that resolution range.

For the CR scenario, the model is not able to extract more meaningful features for the very LR images. Hence, this results in a large distance between features because the HR image features are still very distinctive. However, in the ER scenario both images are somehow unfamiliar to the network and the extracted features are pretty similar. To underline this statement and analyze the distribution of features more fine-grained, the distributions of genuine and imposter feature distances for five specific image resolutions are visualized in a violin plot for the LFW dataset (see Figure 3.4).

The center violin plots represent the feature distance distribution for HR image pairs. Distances for genuine and imposter image pairs are clearly distinguishable. The genuine distances are mainly in a range between 0.1 and 0.6, whereas imposter distances are mostly in the field of 0.6 and 1.4. Both classes can be separated effectively with a threshold of about 0.6, and thus, the accuracy for only HR images is best (see Figure 3.2). To the left side, distributions for the CR scenario are shown. On the right side, ER feature distributions are plotted. In both procedures, for a image resolution of 56×56 px no significant difference can be noticed. Interestingly, the peak feature distance for genuine image pairs even exceeds the maximum distance for imposter pairs in the CR scenario at very LR 5×5 px. In other words, the resolution has more impact on the distance than the identity itself. The gap between

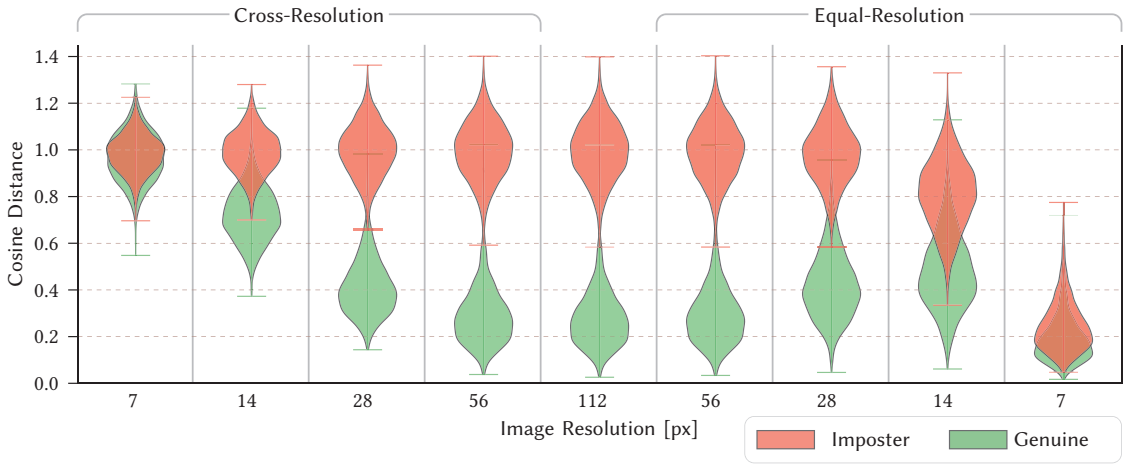


Figure 3.4.: The cosine feature distances, divided into genuine and imposter pairs, for a re-implementation of the ArcFace model adapted from [1†], are calculated for various synthetically downscaled images of the LFW dataset [31]. The analysis includes cross-resolution image pairs (left) and both image pairs of the same resolution (right). The model was trained on the MS1M-V2 dataset.

CR and ER accuracy for very LRs is therefore reasonable. This behavior explains a higher accuracy for very LRs in the ER scenario compared to CR scenario. Further experiments with CFP-FP, AgeDB, CALFW, and CPLFW datasets show the same trend.

3.5. Conclusion

This chapter recapitulates the analysis in [1†] and reveals that the image resolution has a significant impact on the performance of a state-of-the-art FR model. Especially for image resolutions below 28×28 px the effect is severe. However, this effect is measured after preprocessing, *i.e.*, that the images are already aligned and cropped. In a real-world scenario, the effect might be even more significant, due to the limitations of face detection at very LRs. A comprehensive analysis of face detection on LR images is out of the scope of this work. The interested reader may refer to the work of Hayashi and Hasegawa [119].

A closer look to the feature distances shows that the model is not able to extract discriminative features for very LR images. This results in a large distance between features because the HR image features are still very distinctive. However, in the ER scenario both images are unfamiliar to the network and the extracted features are pretty similar. This is also reflected in the accuracy scores.

The results of this chapter lay the basis for the following chapter which covers the development of training strategies to mitigate the effect of the image resolution on the performance of FR models, especially for the CR scenario.

“The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.”

— Stephen Hawking

Strategies to Enhance Face Verification Robustness

This chapter is divided into four parts. The first part presents work related to the field of resolution-robust *Face Verification* (FV). Then, three approaches for robustness-aware training are introduced, which were published in the Leibnitz Transactions on Embedded Systems — Special Issue on Embedded Systems for Computer Vision (2021) [1†]. Furthermore, a fine-tuning method for robustness-enhancing of existing models is presented, which has been published at the IEEE conference series on Automatic Face and Gesture Recognition (2023) [2†]. In the last part of this chapter, the discussions and findings of both publications are summarized and extended.

4.1. Related Work

In general, *Face Recognition* (FR) approaches for *Cross Resolution* (CR) scenarios can be categorized into two groups: 1) Transformation-based approaches, which transform images before feature extraction. 2) Non-transformation-based methods, which directly use the original images for feature extraction [120]. In the following two subsections, an overview of the most relevant related work is given. The interested reader is referred to the work of Wang *et al.* [121] for a more comprehensive review of *Low Resolution* (LR) and CR FR.

4.1.1. Transformation-Based Approaches

Transformation-based methods aim to focus on either first mapping the input images into a common space before processing with FR models or projecting the extracted features into a common space. The left side of Figure 4.1 illustrates common approaches with an example image pair.

The upper left flowchart of Figure 4.1 shows the feature extraction into an *High Resolution* (HR) feature space, which requires the LR image to be transformed into an HR image, before fed through a feature extraction network (typically a *Convolutional Neural Network* (CNN) or *Vision Transformer* (ViT)), prior trained on HR images. The upscaling can be performed with classical methods, such as bicubic interpolation, or with *Deep Learning* (DL)-based methods, such as *Super Resolution* (SR) approaches. The latter approaches have gained significant attention in the last years, as they are able to generate more high-quality HR images from LR images. Jiang *et al.* [122] provided an exhaustive

review of face SR in general. With a focus on FR, prior guided [123, 124, 125, 126, 127] and attribute constrained [128, 129, 130, 131] face SR approaches were proposed. However, they aim for a visually pleasant reconstruction ignoring identity-related information. Therefore, numerous works [132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144] leveraged FR networks to ensure face feature similarity and optimized the SR to preserve identity information. Zangeneh *et al.* [145] proposed a two branch deep CNN. While the LR branch consists of a SR network combined with a feature extraction network, the HR branch is only a feature-extraction network. Both branches are trained in three different training phases. For testing, images are fed through the branches depending on their resolution. A similar approach was used in [146], in which they trained a U-Net [147] with a combination of reconstruction and identity preserving loss in order to super-resolve multi-scale LR images. For feature extraction, they utilized a pre-trained Inception-Residual-Network (iResNet) [148]. To cope with weakly labeled datasets, Hsu *et al.* [149] apply an identity-preserving contrastive loss, whereas Kazemi *et al.* [150] utilize an adversarial FV loss. Very recently, Ghosh *et al.* [151] presented an end-to-end supervised resolution enhancement and recognition network using a heterogeneous quadruplet loss metric to train a *Generative Adversarial Network* (GAN), which super-resolves images without corrupting the discriminative information of the LR images.

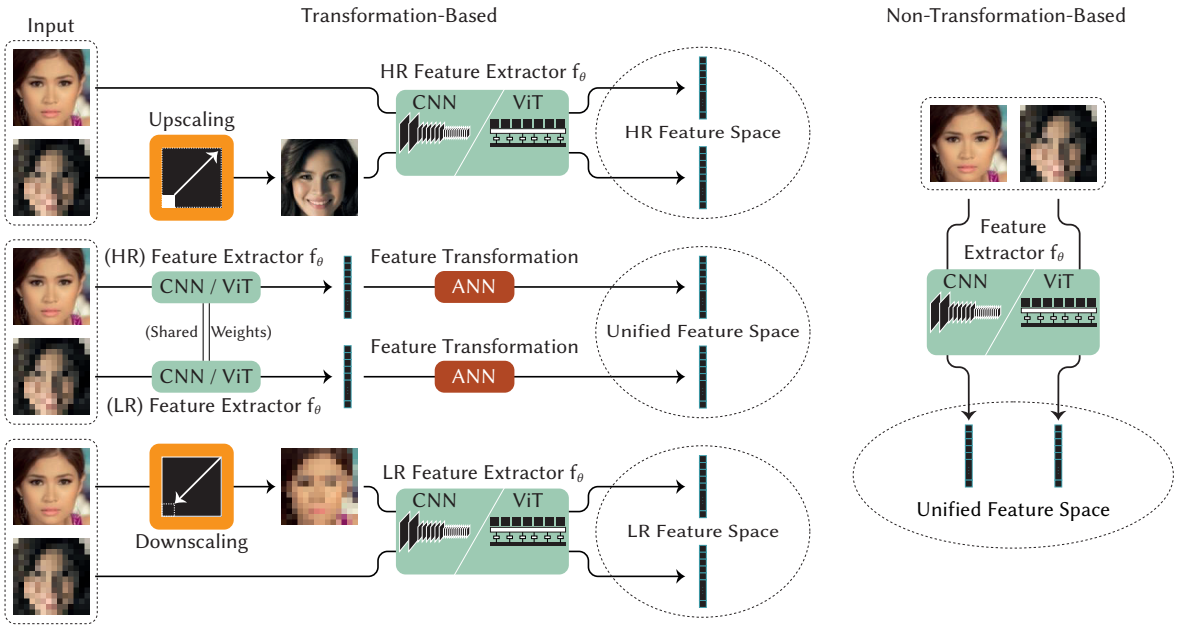


Figure 4.1.: Transformation-based approaches (left) either project the images into a common space before feature extraction (top/bottom left), or transform the extracted features afterwards into common space via a Artificial Neural Network (ANN) (center left). Non-transformation-based (right) approaches aim to learn directly scale-invariant image features. Images taken from *Microsoft 1 Million Version 2* (MS1M-V2) [47, 62].

The center flowchart of Figure 4.1 shows the feature extraction through either two separate feature extraction networks, which focus on HR and LR images independently, or through a single feature extraction network, which is trained with both HR and LR images. Consequently, the extracted features are then transformed via another network into a common feature space. Lu *et al.* [152] presented, *e.g.*, a deep coupled *Residual Layer Network* (ResNet) [54] model, containing one trunk network and a two branch network. The trunk network extracts features and the two branches networks transform HR and the corresponding LR features to a space where their difference is minimized.

The lower flowchart of Figure 4.1 illustrates the straightforward method. The HR images are simply transformed into an LR image, before fed through a feature extraction network, prior trained on LR images.

4.1.2. Non-Transformation-Based Approaches

Non-transformation-based methods try to directly project features from arbitrary image resolutions into a common feature space. The right side of Figure 4.1 illustrates those approaches with an example image pair. Here, the left image is considered an HR image and the right image an LR image. Both images are then extracted from a single network and directly projected into a unified feature space. In [153], Zeng *et al.* presented a resolution-invariant deep network and trained it directly with unified LR and HR images. However, they used only resolutions in the range of 24 to 60 pixels. Massoli *et al.* [154] proposed a student-teacher network approach and showed that their approach can be more effective concerning to preprocessing images with SR techniques. In [145], this was accomplished by a non-linear coupled mapping architecture using two deep CNNs. [154] approached the problem differently with a student-teacher method. In [155], Talreja *et al.* proposed an attribute-guided CR FR model utilizing a coupled GAN and multiple loss functions. Ge *et al.* [156] focused on low computational costs and introduced a new learning approach via selective knowledge distillation. A two-stream technique, comprising a large teacher model and a lightweight student model, is employed to transfer selected knowledge from the teacher model to the student model. Sun *et al.* [157] proposed a shared classifier between HR and LR images to further narrow the domain gap. To fully exploit intermediate features and loss constraints, they embed a multi-hierarchy loss into intermediate layers, reducing the distance of intermediate features after the max-pooling layer and avoiding an over-utilization of intermediate features.

Zeng *et al.* [153] presented a resolution-invariant deep network and trained it directly with unified LR and HR images. In [158], the authors applied a CR contrastive loss on higher-level features of two separate network branches, with each branch focusing precisely on one resolution (high and low). The following two methods go one step further: [159] tackled the problem with a deep siamese network structure and combined a classification loss with a CR triplet loss. Zha and Chao [160] also applied CR triplet loss, but in contrast to [159], they used a two-branch network similar to [158]. In the recent past, [161] proposed a multi-scale parallel deep CNN feature fusion architecture. In contrast to most other FR systems, they provide an end-to-end approach and directly predict the similarity score of two input images. However, they do not report performance on CR datasets such as *Cross-Quality Labeled Faces in the Wild* (XQLFW) [3†] or *Labeled Faces in the Wild* (LFW) [31], nor

do they provide any code or models. Very recently, Li *et al*. [162] proposed a novel deep rival penalized competitive learning strategy for LR FR. However, they did not test their approach on CR images. The work of Terhörst *et al*. [163] pursued a distinct goal. They focused on a more general quality-aware FR, *i.e*., they do not concentrate solely on the physical image quality but also consider pose and age variations. Their approach combines a quality-aware comparison score, utilizing model-specific face image qualities, with an FR model based on a magnitude-aware angular margin loss. A rather unusual method in this field of research but still relevant is the work of Zhao [164], which shows a new technique for correlation feature-based FR. However, his approach relies on video streams and does not provide any quantitative results.

4.2. Robustness-Aware Training

This section covers three training strategies for image resolution robust FV models. First, the methodologies of three approaches are introduced. Then, the experimental settings are described for two training scenarios: 1) A two-resolution scenario, which incorporates only one specific LR during training and 2) A multiple-resolution scenario, where LR images with multiple resolutions are used during training. Afterwards, the results are presented and compared with other relevant works. Although this section is mainly based on [1†], the structure is rearranged and the namings of the approaches are slightly adapted to fit the context of this dissertation.

4.2.1. Methodology

In the rapidly evolving field of FR, the robustness of algorithms against variations in image resolution remains a critical challenge. The methodologies described in the subsequent sections are designed to address this challenge by introducing innovative training strategies that enhance the model's ability to accurately recognize faces across a wide range of image resolutions. These strategies leverage the principles of *Resolution Augmentation Training* (RAT) (named BT in [1†]), *Contrastive Loss Training* (CLT) (named ST / ST-M1 in [1†]), and *Multi-Branch Contrastive Loss Training* (MB-CLT) (named ST-M2 in [1†]), each of which contributes uniquely to a model's resilience against resolution variability.

4.2.1.1. Resolution Augmentation Training

Motivated by [153, 154], the straightforward CR RAT approach to tackle the varying image resolutions is introduced. The left part of Figure 4.2 illustrates this approach. The architecture consists of a single branch, containing a feature extraction CNN network followed by a modified fully connected layer utilizing *Additive Angular Margin Loss* (ArcFace) [47]. Instead of applying only HR images, as holds for a baseline training, in this approach half of the images per batch $\mathcal{B}$ are synthetically reduced in their resolution, to make the network more robust about image resolution. This results in $\mathcal{B}$ containing HR and LR at the same time.

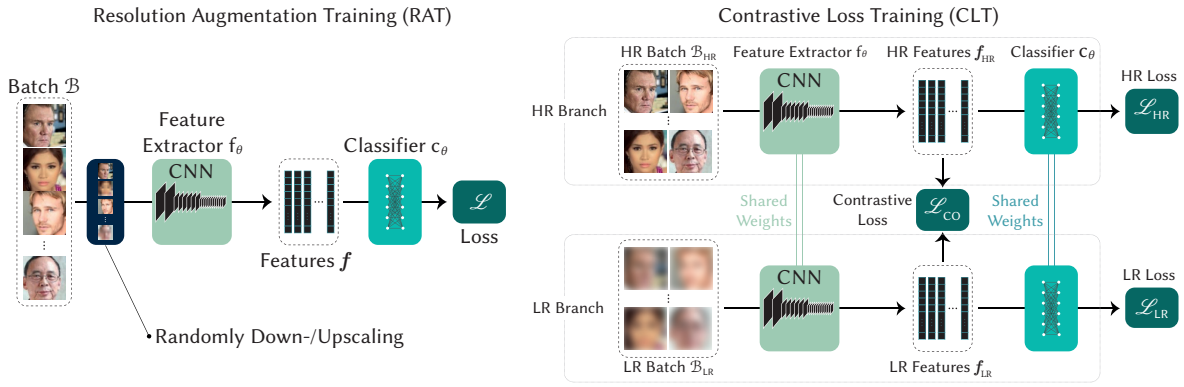


Figure 4.2.: The left side illustrates the RAT [1†] approach. Images are randomly augmented with downscaling during training. The right side highlights the CLT [1†] approach. A contrastive loss between the high- and low resolution branch is calculated. Cross-entropy losses are calculated for both branches. Sample images taken from MS1M-V2.

4.2.1.2. Contrastive Loss Training

Inspired by Tang *et al.* [165], a siamese network structure is build with a simplified version of the contrastive loss $\mathscr{L}_{CO}$ function (see Section 2.2.2.1). Given that in this scenario, only the resolution of the images is varied, the contrastive loss is solely calculated for positive pairs, *i.e.*, Y in Equation (2.13) is always 0. The right part of Figure 4.2 illustrates the proposed architecture. The approach consists of two branches, one for HR images and one for LR images. The branches share their weights, *i.e.*, only one CNN network and a classification layer is utilized. The amount of weights is thus equal to the RAT approach. The input batch $\mathcal{B}$ is duplicated and images are downscaled online during training. The batches are then denoted as $\mathcal{B}_{HR}$ and $\mathcal{B}_{LR}$. Both batches are fed through the network for each training step and the losses $\mathscr{L}_{CE}^{[HR]}$ and $\mathscr{L}_{CE}^{[LR]}$ are calculated, respectively. Additionally, a contrastive loss $\mathscr{L}_{CO}$ is calculated between the extracted features $\boldsymbol{f}_{HR}$ and $\boldsymbol{f}_{LR}$ of the HR and LR branches, enforcing the network to learn features that are invariant to image resolution. As distance measure the cosine distance (see Section 2.2.2.1) is used for the contrastive loss function. Choosing the cosine distance is reasonable, since evaluation is also done by calculating the cosine distance between image pairs. The total loss $\mathscr{L}$ is then the sum of the cross-entropy losses $\mathscr{L}_{CE}^{[HR]}$, $\mathscr{L}_{CE}^{[LR]}$ and the contrastive loss $\mathscr{L}_{CO}$:

$$\mathscr{L} = \mathscr{L}_{CE}^{[HR]} + \mathscr{L}_{CE}^{[LR]} + 25 \cdot \mathscr{L}_{CO}\,. \tag{4.1}$$

Due to the cosine distance ranging from $[0, 2]$, a factor of 25 is applied to balance the contrastive loss with the classification losses approximately. In this scenario, the computational effort doubles, as the network must process both HR and LR images.

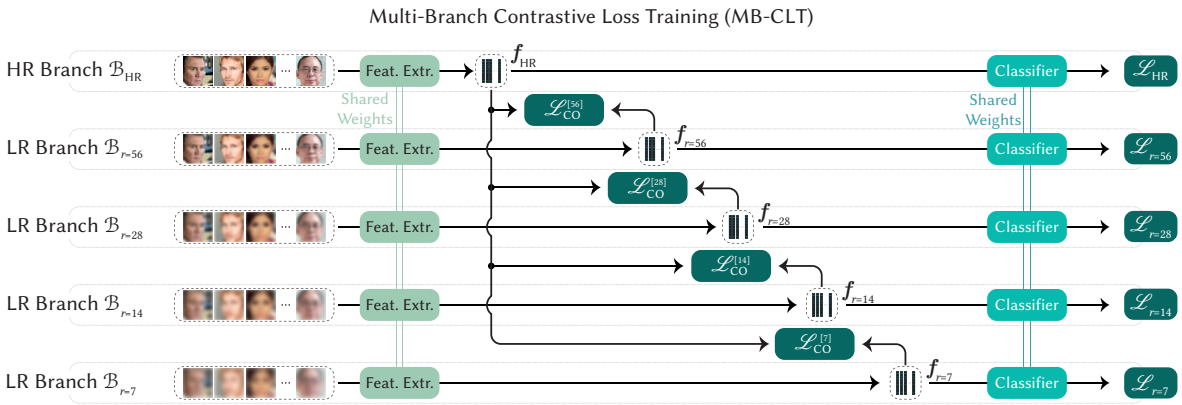


Figure 4.3.: Flowchart of the Multi-Branch Contrastive Loss Training (MB-CLT) [1†] approach. The network is arranged in a siamese structure, which is trained with five different resolutions simultaneously. The feature distance loss is calculated between for each low resolution branch compared to the high resolution branch. The classification loss is calculated for each branch. Images taken from MS1M-V2.

4.2.1.3. Multi-Branch Contrastive Loss Training

Extending the CLT approach with the addition of more branches for specific image resolutions culminates in the MB-CLT method. The architecture of this method is depicted on the right side of Figure 4.3. The integration of multiple branches allows for concurrent training across various LRs. Similar to the CLT, the cross-entropy loss $\mathscr{L}_{CE}$ is computed for each branch. Additionally, a cosine distance contrastive loss $\mathscr{L}_{CO}$ is calculated between the features of all LR branches $\boldsymbol{f}_r$ and the HR branch $\boldsymbol{f}_{HR}$. All branches are unified by sharing weights, thus employing a single CNN network and a classification layer. Consequently, the total amount of weights remains equivalent to those in the RAT and CLT methods. The overall loss $\mathscr{L}$ is then computed as follows:

$$\mathscr{L} = \sum_{r \in \mathcal{R}} \mathscr{L}_{CE}^{[r]} + 25 \cdot \sum_{r \in \mathcal{R}} \mathscr{L}_{CO}^{[r]}, \tag{4.2}$$

with $\mathcal{R} := \{7, 14, 28, 56\}$ being the raw-resolutions for the LR images. The computational cost is about five times larger than the RAT training method, since the network needs to infer additionally all LR images.

4.2.2. Experimental Settings

In this section, first the baseline feature extraction network for the trainings is described in more detail. Secondly, the downscaling method is described. Finally, two scenarios for the experiments are introduced.

4.2.2.1. Network and Training Details

The in [47] used 50-layer deep ResNet-50 [54] architecture is utilized as the feature extraction network. The weights of this network are pre-trained on ImageNet [166]. This backbone network consists of four blocks, which are repeated several times and containing in total 50 convolutional layers. The image dimension within the layers is decreasing, and the image depth is increasing from $112 \times 112 \times 3$ px input to $4 \times 4 \times 2\,048$ px at the end. After flattening this output from the backbone network, a dropout layer is added. Finally, following [56, 70, 105] a bottleneck layer (512-dimensional fully connected layer) outputs the feature vectors. The classifier consists of a the fully connected classification layer as utilized in the ArcFace [47] approach with the dimension of the number of identities in the training set (87k).

For training, the MS1M-V2 [47, 62] dataset is used, containing about 5.8M images from about 87k identities. Online data augmentation is performed as described in Section 2.3.2.2 using random brightness and saturation variations combined with left-right flipping. All training parameters are set according to [47] except for a smaller batch-size of 128 due to hardware limitations. The learning rate is set to 0.01 and is decreased by a factor of 10 after epoch 9 and epoch 13. In total, all approaches are trained similar to [47] for 16 epochs with momentum *Stochastic Gradient Descent* (SGD) optimizer. The dropout rate and weight decay are set to 0.5 and $5 \cdot 10^{-4}$, respectively.

Training the RAT architecture with solely HR images is referred to as the baseline network in this context.

4.2.2.2. Reduction of Image Resolution

Following [54], the ResNet-50 architecture requires an input image size of 112×112 px. As proposed in Section 4.2.1, a downscaling $\downarrow_s(\cdot)$ method is utilized to reduce the image resolution (see Section 2.1.2). To retain the required input image size, the downscaled images are then upscaled $\uparrow_s(\cdot)$ subsequently using bicubic interpolation. This process can be formulated as:

$$\boldsymbol{I}_r^* = \uparrow_s \left(\downarrow_s \left(\boldsymbol{I}_{\mathrm{HR}} \right) \right), \tag{4.3}$$

where r denotes the resulting reduced raw-resolution (see also Section 2.1.3).

The resulting low raw-resolution images $\boldsymbol{I}_r^*$ then have the same resolution as the original images, but depending on r the high frequency information is removed, thus simulating LR images. For both scaling processes, bicubic interpolation is applied. To reduce unwanted artifacts, typically stemming from the downscaling process, standard anti-aliasing techniques are also utilized. Figure 4.4 visualizes the process of resolution reduction for an example image. The left image $\boldsymbol{I}_{\mathrm{HR}}^{[112]}$ is a sample from the MS1M-V2 dataset with a original resolution 112×112 px. The center image $\boldsymbol{I}^*$ is the downscaled ($r = 14$) LR image with the dimension 14×14 px. The right image $\boldsymbol{I}_{\mathrm{LR}}^{[14]}$ is the upscaled low raw-resolution image with the dimension 112×112 px, but contains a raw-resolution of $r = 14$, *i.e.*, 14×14 px.

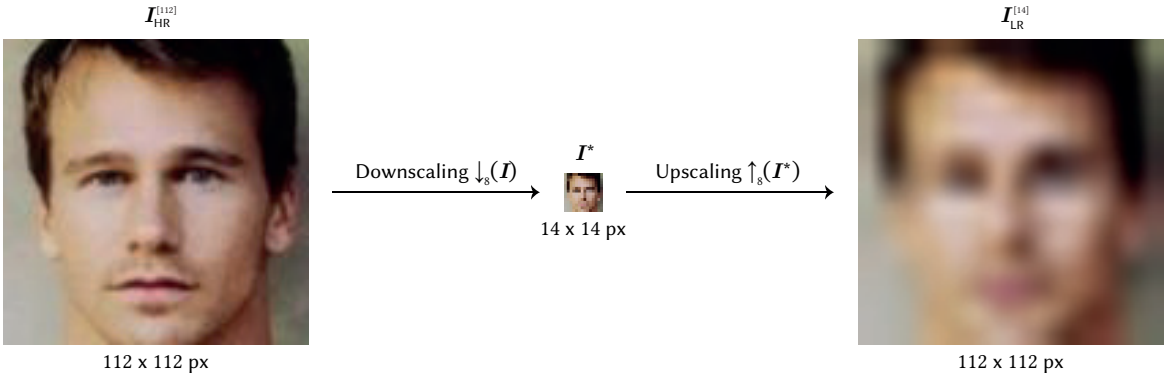


Figure 4.4.: Illustration of the synthetic image resolution reduction via bicubic down- and upscaling. Sample image taken from the MS1M-V2 dataset.

4.2.2.3. Training and Testing Scenarios

Two training scenarios are conducted. The first scenario utilizes the RAT and CLT method, which are trained with one specific raw-resolution r during the whole training process. These are referred to as RAT-r and CLT-r accordingly. To develop a model that can handle arbitrary resolutions more effectively, a second scenario involving simultaneous training with multiple resolutions is conducted. In this scenario the RAT, CLT and additionally, the MB-CLT methods are trained with five different raw-resolutions $r \in \{7, 14, 28, 56, 112\}$ simultaneously. This is referred to as RAT-M, CLT-M, and MB-CLT-M, respectively. In RAT each batch contains randomly picked downscaled images of equally distributed resolutions r. In CLT the LR batch $\mathcal{B}_{LR}$ contains downscaled images of randomly but equally distributed resolutions r and in MB-CLT each LR branch $\mathcal{B}_r$ contains images of a specific raw-resolution r.

With networks capable of handling arbitrary image resolutions at once, there is a need for a more meaningful evaluation considering multiple resolutions. To make the evaluation reproducible, three evaluation protocols for the utilized benchmark datasets are introduced and published as lists with the specific image resolution for each single image pair: 1) *Low Resolutions Protocol*, which is derived by randomly picking resolutions in the range of 5 px to 10 px for the synthetic downscaling 2) *Intermediate Resolutions Protocol*, which is derived by randomly picking resolutions in the range of 10 px to 40 px and 3) *High Resolutions Protocol*, which is derived by randomly picking resolutions in the range of 40 px to 112 px.

Additionally, an *All Resolutions Protocol* that takes every single resolution in the range of 5 px to 112 px into account is introduced.

For the evaluation, image pairs from the benchmark datasets including LFW [31], CALFW [94], CPLFW [92], CFP-FP [93], and AgeDB [95] undergo synthetic degradation, mirroring the treatment applied to the training data. In these experiments, two variations of the benchmark protocols are crafted for specific image resolutions: One where the first image in each pair experiences resolution

reduction, and the other where the second image in each pair is similarly downgraded. The overall accuracy is then calculated as the average of both versions.

Notably, the inference time and number of parameters is equal for all approaches, making the comparison fair.

4.2.3. Results

In this section, the CR evaluation results for both training scenarios are presented. Similar to Chapter 3 the feature distances are analyzed in more detail. Finally, the results are compared with other relevant works and computational efforts are discussed.

4.2.3.1. Two-Resolution Scenario

Face Verification Accuracy. As introduced in Section 2.3.4, accuracy is a common metric to measure the performance of an FV model. Figure 4.5 depicts the average FV accuracy of the RAT and CLT approach compared to the baseline model across five common datasets: LFW [31], *Cross-Age Labeled Faces in the Wild* (CALFW) [94], *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92], *Celebrities in Frontal-Profile — Frontal-Profile* (CFP-FP) [93], and *Age Database* (AgeDB) [95]. Note that each data point of RAT and CLT represent a distinctive model (RAT-r, CLT-r), which was specifically trained for the resolution r. Due to hardware limitations, models are trained only for depicted resolutions. Both approaches clearly outperform the baseline on lower image resolutions. For very LRs, *i.e.*, 5 px to 8 px, the performance can be increased from about 50% (random guess) up to 70%. That is equivalent

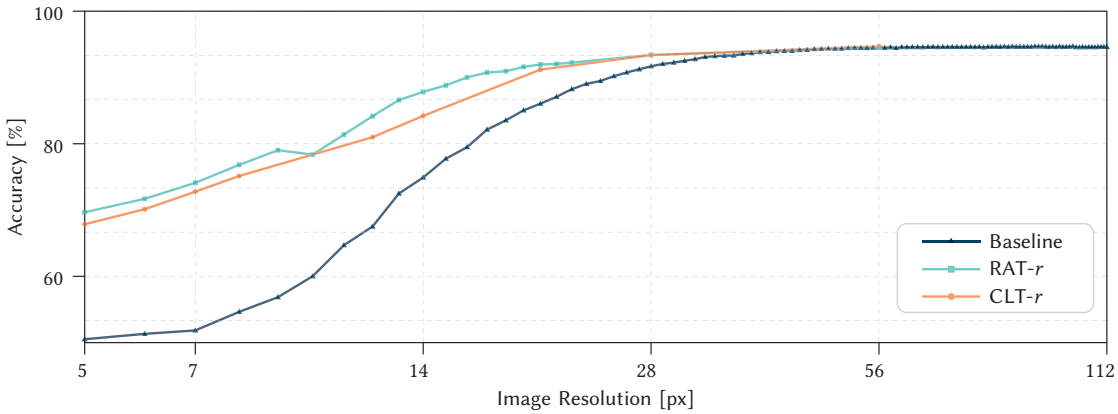


Figure 4.5.: Face verification accuracy of RAT-r and CLT-r models, benchmarked against the baseline performance of an ResNet-50 (ArcFace) model re-implementation adapted from [1†], all trained on the MS1M-V2 dataset. Each data point corresponds to a distinct model, specifically trained and tested on images with identical synthetically downscaled resolutions r, adhering to the cross-resolution protocol.

Table 4.1.: Face verification accuracy on downscaled cross-resolution LFW benchmark [31] protocols of the proposed approaches, trained on MS1M-V2 compared to other works.

	Accuracy [%] for LFW [31]		
Model	16 × 16 px	32 × 32 px	32 × 32 px
RAT-16 [1†]	**98.17**	–	–
CLT-16 [1†]	97.88	–	–
S-16-sc [156]	85.87	–	–
RAT-32 [1†]	–	**99.08**	–
CLT-32 [1†]	–	98.82	–
S-32-sc [156]	–	89.72	–
RAT-64 [1†]	–	–	**99.38**
CLT-64 [1†]	–	–	99.35
S-64-sc [156]	–	–	92.83
Talreja *et al.* [155]	–	91.08	94.92

to a relative improvement of 40%. Above about 40 px image resolution, no significant difference between all approaches is present, which affirms the expectations since the LR images are visually hardly distinguishable from the original images and the absolute pixel difference is very small (see Section 3.2).

Generally, the performance improvement is increasing with decreasing resolutions. The RAT method performs slightly better than the CLT method, which leads to the conclusion that the siamese approach might concentrate too much on projecting the features of the same image in different resolution to the same space than on classifying the correct identity regardless of the resolution.

Moreover, the FV accuracy of the RAT and CLT approach is also compared to the work of Ge *et al.* [156] and Talreja *et al.* [155] on the LFW dataset for three distinct image resolutions $r \in \{16, 32, 64\}$. Table 4.1 reveals that the proposed approaches outperform both competitors significantly. However, the comparison to Ge *et al.*'s approach might not be fair, their baseline model (teacher model) only reaches an accuracy of 97.15%, which is not comparable to our baseline and the state-of-the-art. Moreover, the model's number of parameters also differs in the comparison. As obvious in Figure 4.5 the strength of the proposed approaches is the ability to handle very LRs and it would be interesting to see how the other approaches perform on such very LRs. However, they do not provide any results.

Feature Analysis. To get a deeper understanding of the accuracy results, a closer look to the feature distances for the LFW benchmark is conducted. Figure 4.6 compares the cosine distance distributions for the RAT-r and CLT-r approach trained and tested with specific image resolutions r. In line with the improvements of accuracy (see Figure 4.5), cosine distances of genuine and imposter pairs are much better separable for LRs $r \in \{14, 28, 56\}$ than the baseline model (see Figure 3.4). However, there is a remarkable shift of distances in the very LR scenario (7×7 px). Although the distance values are very small, still distances for imposters are larger than distances for genuines and the distributions are separable, which is consistent with improvement of accuracy (see Figure 4.5). Overall, both methods show only slight differences in the distributions, which is also in line with the accuracy values.

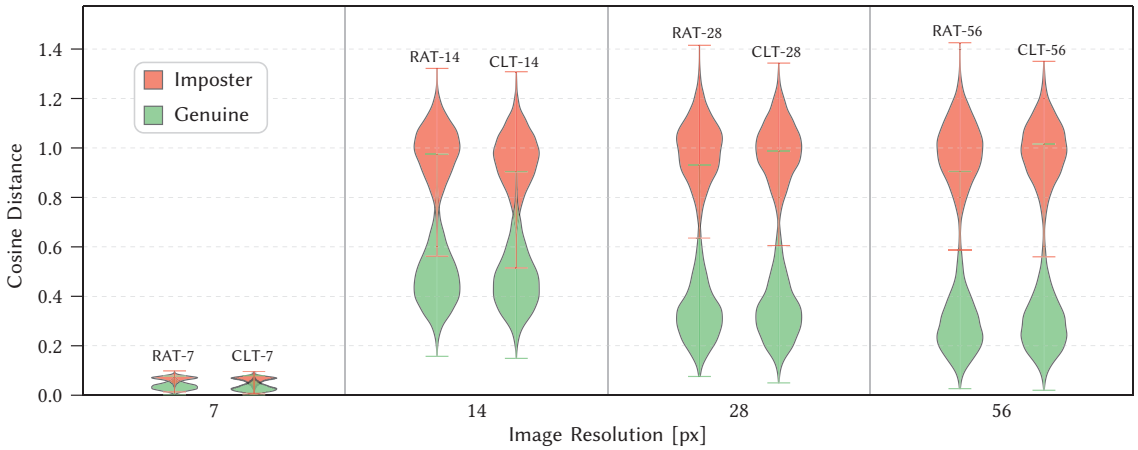


Figure 4.6.: The cosine feature distances, divided into genuine and imposter pairs, for the RAT-*r* and CLT-*r* methods adapted from [1†], trained on MS1M-V2, are calculated for various synthetically downscaled cross-resolution image pairs of the LFW dataset.

A more fine-grained analysis on the best threshold on the LFW dataset is depicted in Figure 4.7. The best threshold is calculated for the accuracy values of RAT-*r* and CLT-*r* models with a 10-fold cross-validation technique (see Section 2.3.4). All models show a slight increase of the threshold value the lower the resolutions. This is in line with the increasing cosine distances for all image pairs in Figure 4.6. The RAT method is then dropping at a resolution of 10×10 px, whereas the CLT method is dropping at a resolution of 14×14 px. Both methods are plateauing from then on. Looking at the accuracy values in Table 4.1, there is a slight kink in the line of the RAT method at a resolution of 10×10 px, which could be connected with the drop in the threshold. However, this is not the case for the CLT method, but could be due to the lack of model data points in both plots.

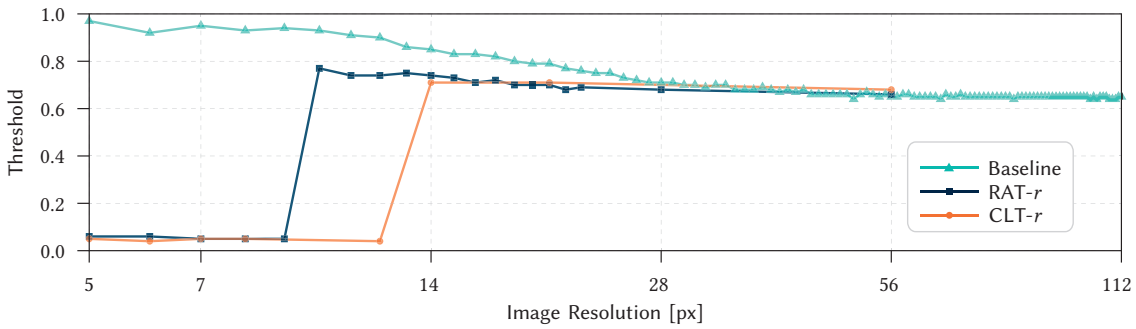


Figure 4.7.: Best face verification evaluation thresholds adapted from [1†] for the LFW dataset derived via 10-fold cross-validation. Models were trained on the MS1M-V2 dataset.

4.2.3.2. Multiple-Resolution Scenario

Face Verification Accuracy. In Figure 4.8, the mean FV accuracy of the proposed methods in the multiple-resolution scenario on five benchmark datasets (LFW, CALFW, CPLFW, CFP-FP, and AgeDB) are presented. Note, that compared to the single-resolution scenario, in this scenario, only one model is used for all testing image resolutions. All approaches show a performance decrease for testing resolutions above 28×28 px. This is the trade-off for the ability to handle arbitrary resolutions at once. However, the RAT-M model performs best for this rather HRs. Below 28×28 px testing image resolution, the RAT-M model's performance is decreasing, but outperforming the baseline except for 15×15 px. The same effect is present for CLT-M and MB-CLT-M, but they only start to fully exploit their potential from a resolution of about 18×18 px, or 16×16 px, respectively.

Focusing on very LRs ($r < 7$), the baseline approach holds an accuracy of approx. 50%, which is equivalent to random guessing and all proposed approaches significantly increase the performance up to 75%. The RAT-M model shows a significant peak at a testing resolution of 14×14 px, which might reason in the included training resolution. However, this effect is not present for the other two methods.

In general, the RAT-M approach outperforms the other two methods for $r > 16$, but vice-versa for $r < 16$. This might be due incorporation of the contrastive loss, which forces the feature distances between HR and LR images to be minimized, and thus the network might concentrate too much on projecting the features of the same image in different resolution to the same space than on classifying the correct identity regardless of the resolution.

Comparing the CLT to the MB-CLT architecture, the CLT method performs better across all tested resolutions. This leads to the conclusion, that the multi-branch approach with largely increased computational effort does not provide any significant benefit in the multiple-resolution scenario.

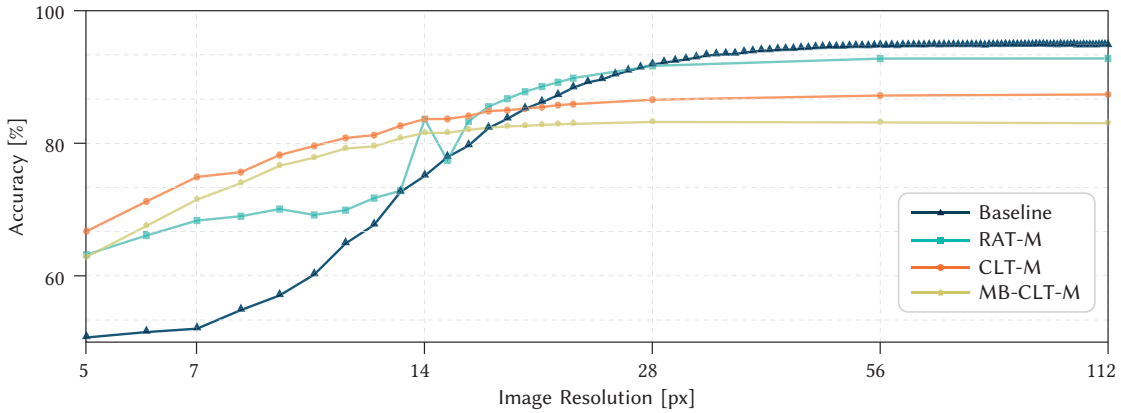


Figure 4.8.: Face verification accuracy of RAT-M and CLT-M models, benchmarked against the baseline performance of an ResNet-50 (ArcFace) model re-implementation adapted from [1†], all trained with multiple-resolutions simultaneously on the MS1M-V2 dataset.

Feature Analysis. Similar to the two-resolution scenario, a closer look to the feature distances for the LFW benchmark is taken. Figure 4.9 shows the cosine distance distributions for the RAT-M model on the left, and the CLT-M and MB-CLT-M approach on the right. The training with multiple resolutions simultaneously does not lead to shrinking feature distances for the RAT approach. In contrast, the feature distances for the CLT and MB-CLT approach are shrinking down to a maximum of 0.1 for all resolutions. Notably, on the very low testing resolution of 7×7 px, the feature distances for the RAT-M model are even larger than for all other resolutions. Interestingly, the genuine and imposter feature distance distributions show a deviation in terms of the Gaussian shaped distributions of distances. Here, the imposter pair distances seem to be a cumulation of three Gaussian distributions and the genuine pair distances seem to be a cumulation of two Gaussian distributions.

Upon a closer examination of the distribution, especially the maximum and minimum values of the three methods, it is observable that for the RAT method, the minimal distances for the imposter pairs do not significantly overlap with the distribution of the genuine pairs. This indicates that if one desires a very low *False Positive Rate* (FPR), or even aims for it to be zero, the RAT-M model is more suitable than the two contrastive loss methods, as the latter exhibit very small distances for the imposter pairs. Similarly, concerning the maximum genuine distances, the RAT-M model also proves to be more appropriate than the other two methods, although the effect is less pronounced. This subtlety could be attributed to the utilization of only a fraction of the 512-dimensional space for the features, which might also explain why the distances do not vary significantly. This also provides an explanation for why the accuracy values for the RAT-M model are superior to those of the other two methods, suggesting that there is inherently more potential within it.

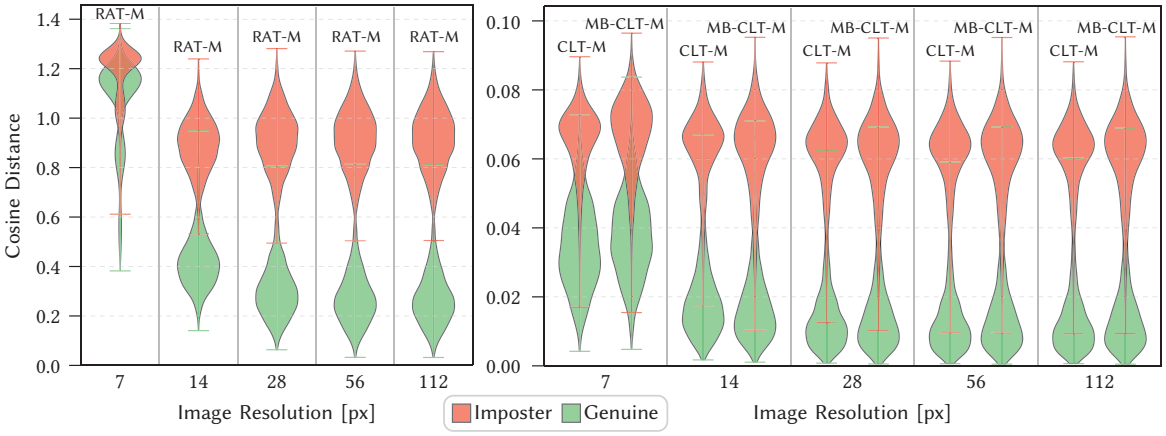


Figure 4.9.: The cosine feature distances, divided into genuine and imposter pairs, for the RAT-M, CLT-M, and MB-CLT-M methods adapted from [1†], trained on MS1M-V2, are calculated for various synthetically downscaled cross-resolution image pairs of the LFW dataset.

4.2.3.3. Comparison of the proposed Methods

This sections draws a comparison of the three proposed methods under both training scenarios. Figure 4.10 shows the accuracy values for the LFW dataset in testing image resolutions the region $r \in [5, 9]$ on the left and in the region $r \in [10, 18]$ on the right. This analysis aims to reveal the strength and weaknesses of the proposed methods in LRs regions and also the effect of neighboring resolutions of the specific resolution trained models (in this case at 7×7 px and 14×14 px).

Looking at the trend of the RAT-14 and CLT-14 models, there is a slight peak visible at $r = 14$ resolution, which is reasonable due to this exact resolution being utilized during training. However, the performance on the neighboring resolutions is not significantly worse. The same effect is visible in the left chart of Figure 4.10 for the RAT-7 and CLT-7 models. This demonstrates that the levels chosen in these experiments are sufficient to represent the resolution spectrum for training, and a finer gradation is not necessarily required. A similar effect is also visible for the multiple-resolution scenarios for the RAT-M and CLT-M model. Another interesting observation is that the accuracy for the RAT-M model is even worse than the baseline for a resolution of 13×13 px. All other methods are clearly outperforming the baseline on the tested image resolutions.

Moreover, the FV accuracy development during training of the baseline, RAT, and CLT ($r \in \{7, 56\}$) is depicted in Figure 4.11. The baseline model achieves about 98% accuracy after the first epoch, followed by almost peak accuracy already after the second epoch. After the third epoch, the performance saturates and no significant changes in accuracy are visible. The RAT-56 starts with equal accuracy after the first epoch and then takes another two epochs to reach almost peak accuracy. The CLT-56 gets only after epoch 4 approximately peak performance. This model needs significantly more training samples than both previously mentioned models to achieve similar accuracy. One reason for that

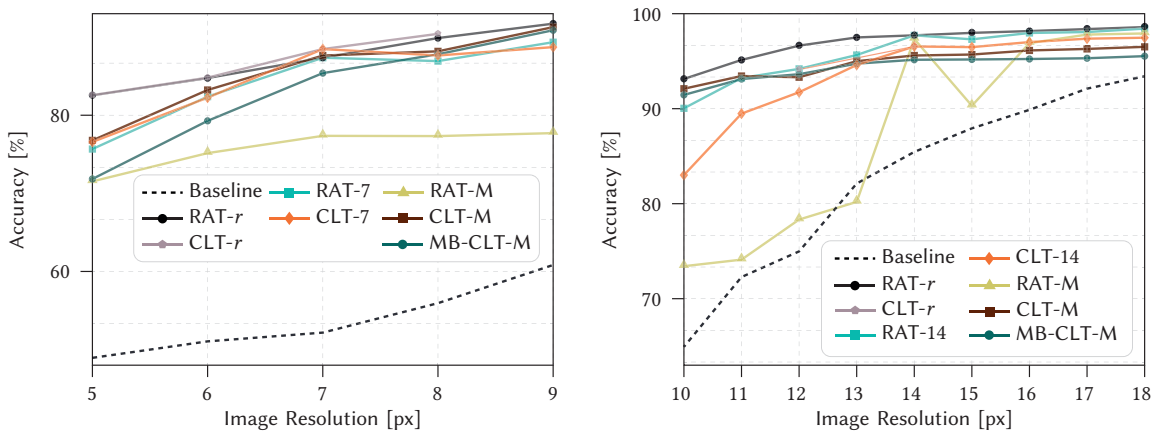


Figure 4.10.: Comparison of face verification accuracy for the proposed models adapted from [1†]. MS1M-V2 is used for training the models and downscaled cross-resolution image pairs for testing are taken from the LFW dataset.

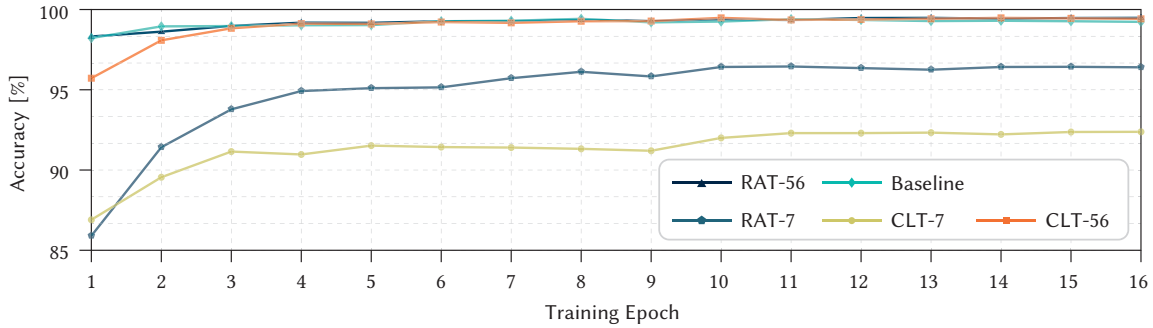


Figure 4.11.: Emergence of face verification accuracy of the proposed models adapted from [1†] on the LFW dataset during training of the proposed models with MS1M-V2.

could be the additional feature distance loss, which forces the network to additionally concentrate on minimizing feature distances and not only classification.

The peak performance for RAT-7 and CLT-7 is significantly lower compared to the other approaches. This is reasonable due to less information in the very low LR images. Moreover, these methods need at least 10 epochs to reach their maximum capability.

Lastly, the training time and FV accuracy on the LFW dataset for different resolution protocols (see Section 4.2.2.3) are depicted in Table 4.2. The resolution protocols confirm that the RAT-M model is the most robust model for arbitrary resolutions including the HR scenario (112×112 px) and also very efficient in terms of training time. However, it lacks in performance for very LR images. The CLT-M model is the most robust model for LR images. It also yields the best performance on specifically 5×5 px CR image pairs. However, it is more time-consuming in training than the RAT-5, RAT-7, and RAT-M approaches. The MB-CLT-M model is the most time-consuming model, which is reasonable due to the additional inference of all LR images in the LR branches. It's performance is worse on all resolution protocols, which is in line with the findings in Section 4.2.3.1. The integration of contrastive

Table 4.2.: Training time and face verification accuracy adapted from [1†] for different cross-resolution protocols of the LFW dataset. The training of the models is conducted using the MS1M-V2 dataset.

		Accuracy [%] for LFW [31]					
			Resolution Protocols				
Model	Training Time [h]	112×112 px	All	High	Intermediate	Low	5×5 px
ArcFace [47]	2	99.23	96.86	99.2	95.89	77.57	54.65
RAT-5 [1†]	2	–	–	–	–	–	69.66
RAT-M [1†]	2	**99.3**	**97.72**	**99.33**	**97.78**	87.17	71.53
RAT-7 [1†]	2	–	–	–	–	–	67.86
CLT-M [1†]	4	97.4	96.76	97.35	96.98	**91.5**	**76.78**
MB-CLT-M [1†]	20	95.62	95.07	95.62	95.51	88.72	71.84

loss, while yielding a performance increase of approximately 14% in the LR protocol for very LR images, necessitates a compromise with a reduction of about 2% in performance for HR scenarios.

In conclusion, the strategy of augmenting images during the training phase yields significant advantages in enhancing the resolution robustness of FV models. However, this method encounters limitations at extremely LR images. The application of contrastive loss, designed to compel the network towards minimizing feature distances across varying resolutions, emerges as a promising methodology. This revelation lays the foundational groundwork for the formulation of an advanced loss function, the details of which will be elucidated in the subsequent section.

4.3. Robustness-Enhancing Fine-Tuning

Elaborating on the findings of the previous section, this section introduces an enhancement of the contrastive loss addition. The key idea is to force also different images of the same identity to be projected close to each other, regardless of the resolution. Moreover, images of different identities shall be far apart from each other. In contrast to the previous section, this approach is a fine-tuning strategy and thus requires pre-trained models. The following section primarily integrates the publication [2†] into this work and adapts it to the context, while largely drawing upon it. After introducing the methodology, the experimental settings are described and results are discussed. Finally, this work discusses several ablation studies.

4.3.1. Methodology

Central in this section is the *Octuplet Loss Training* (OLT) method, a fine-tuning approach designed to build upon the foundational principles of triplet loss. By formulating a sophisticated loss function that integrates both HR and LR images within the same framework, the aim is to refine the ability of FR models to maintain performance consistency across a spectrum of image qualities. This method not only facilitates a direct learning pathway between HR and LR images but also ensures that the network's proficiency with HR images remains uncompromised.

The primary purpose of this approach constitutes improving the robustness of existing FR models by exploiting the triplet loss. Inspired by [159, 167], four different triplet loss terms to combine HR and LR images are formulated. Contrary to the methodology presented in [159], the concept of fine-tuning, rather than beginning training from scratch with a classification loss, is adopted. The introduction of the OLT aims to enable networks to directly understand the connection between HR and LR images while maintaining performance on HR images. The idea of applying triplet loss to features from different image resolutions was also proposed in [160]. However, in their approach, features are computed via two separate branches of the network, thus increasing computational costs. The goal is to project embeddings from images of arbitrary resolutions r into a common feature space directly without modifying existing network architectures. The left part of Figure 4.12 illustrates the OLT strategy and the right part visually explains the concept of octuplet loss.

FV benchmarks and applications typically utilize the distance between facial feature embeddings to distinguish between genuine and imposter identities. Therefore, similar to the CLT approach

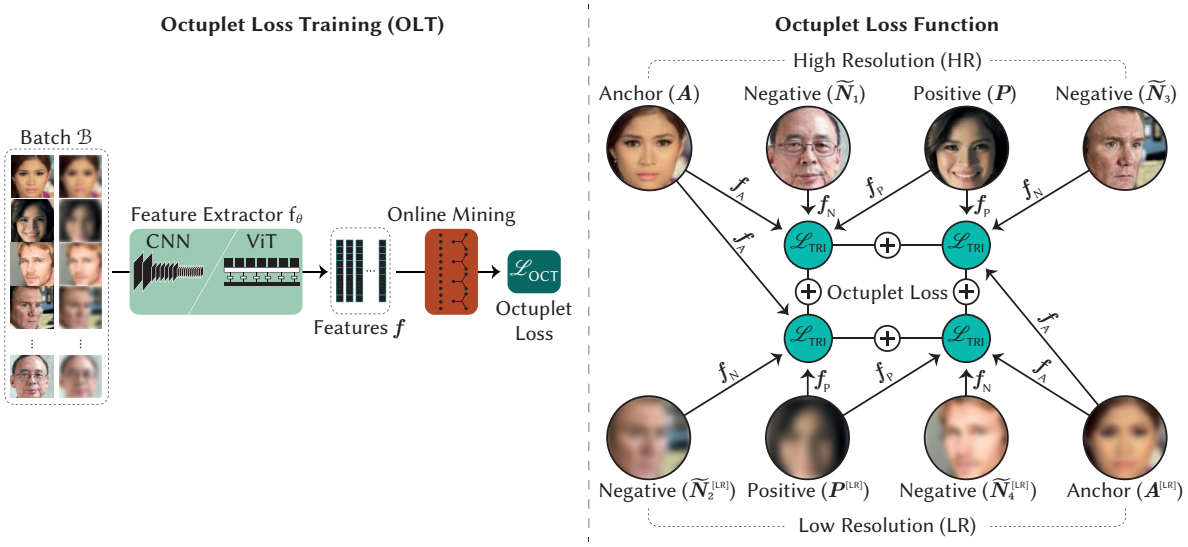


Figure 4.12.: Illustration of the octuplet loss training strategy on the left. Visual description of the octuplet loss function leveraging eight high- and low-resolution images simultaneously on the right. Adapted from [2†].

(see Section 4.2.1) it is reasonable to employ the feature distances directly in the training phase. Due to the lack of sufficiently large FR training datasets containing both LR and HR images, LR images are synthetically deteriorated. They are downscaled to three particular image dimensions (7×8 px, 14×14 px, and 28×28 px) and subsequently upscaled to 112×112 px using scale factors $s \in \{8, 4, 2\}$. The scaling process is similar to the one used in the CLT approach (see Section 4.2.2.2), except for bilinear interpolation including anti-aliasing, and can be formulated for a single mini-batch $\mathcal{B}$ containing B images $\boldsymbol{I}$ as:

$$\mathcal{B}_{\text{LR}} := \left\{ \uparrow_s \left(\downarrow_s (\boldsymbol{I}) \right) \; : \; \boldsymbol{I} \in \mathcal{B} \right\}. \tag{4.4}$$

With this image degradation method, an LR mini-batch $\mathcal{B}_{\text{LR}}$ containing the same images as the HR mini-batch $\mathcal{B}$ is created, but with different image resolutions. This mini-batch then acts as an additional source for selecting the triplets. Triplets are defined as the combination of an anchor image $\boldsymbol{A}$, a positive image $\boldsymbol{P}$, and a negative image $\boldsymbol{N}$ (see also Section 2.2.2.1). To avoid cluttering with subscripting, in contrast to Section 2.2.2.1, the notation of these three distinct image types slightly differs. In this case, anchor, positive, and negative images can be also picked from $\mathcal{B}_{\text{LR}}$, which are then denoted accordingly as $\boldsymbol{A}^{[\text{LR}]}$, $\boldsymbol{P}^{[\text{LR}]}$, $\boldsymbol{N}^{[\text{LR}]}$. The anchor and positive images are of the same identity, while the negative image is of a different identity. For the triplet loss $\mathscr{L}_{\text{TRI}}$ the feature distances between the anchor and positive images are minimized, while the feature distances between the anchor and negative images are maximized. The feature distances are calculated by a distance function $\mathrm{d}(\cdot, \cdot)$, which utilizes extracted features $\boldsymbol{f}$ from an arbitrary feature extractor. Feature for the anchor, positive, and negative images are denoted as $\boldsymbol{f}_{\text{A}}$, $\boldsymbol{f}_{\text{P}}$, and $\boldsymbol{f}_{\text{N}}$, respectively.

Given the constraint that, a mini-batch strictly contains two randomly selected images of the same identity, the number of identities within each mini-batch is $B/2$. For each anchor image $\boldsymbol{A}$, one can find exactly one positive image $\boldsymbol{P}$ and $B-2$ negative images $\boldsymbol{N}$. Hence, the cardinality of the set is calculated as $|\mathcal{T}| = B^2 - 2B$ (see Equation (2.14) in Section 2.2.2.1). With this set of triplets $\mathcal{T}$, the maximum information within each mini-batch is exploited by the triplet loss. However, the majority of triplets within $\mathcal{T}$, according to Equation (2.14), do not contribute towards $\mathscr{L}_{\text{TRI}}$ as they are already correctly classified and thus fulfill $\mathrm{d}\big(\boldsymbol{f}_{\text{A}}, \boldsymbol{f}_{\text{P}}\big) + \alpha < \mathrm{d}\big(\boldsymbol{f}_{\text{A}}, \boldsymbol{f}_{\text{N}}\big)$ (see also Equation (2.15)). According to Hermans *et al.* [46] the training procedure is accelerated by selecting only the most relevant hard negative sample $\widetilde{\boldsymbol{N}}$, which is obtained for a given anchor image $\boldsymbol{A}$ by:

$$\widetilde{\boldsymbol{N}} = \arg\min_{\boldsymbol{N}} \mathrm{d}\big(\boldsymbol{f}_{\text{A}}, \boldsymbol{f}_{\text{N}}\big) \,. \tag{4.5}$$

This additional constraint leads to a more meaningful set $\mathcal{T}$ and hence a less costly minor cardinality $|\mathcal{T}| = B$. However, selecting the most challenging sample is prone to include outliers, *e.g.*, incorrectly labeled data, and thus hinders the feature extractor in learning meaningful associations. Nevertheless, in line with [46], a large number of triplets mitigates this effect within each mini-batch. Thus, the hard sample mining strategy is considered as a valid method for the proposed fine-tuning. The interested reader is referred to the work of Kaya and Bilge [168] for a more detailed overview of sampling selection processes. Together with this hard sample mining strategy, the following four sets of triplets are then defined:

$$\mathcal{T}_{\text{HHH}} := \left\{ \big(\boldsymbol{A}, \boldsymbol{P}, \widetilde{\boldsymbol{N}}_1\big) \in \mathcal{T}\big(\mathcal{B}, \mathcal{B}, \mathcal{B}\big) \;:\; \widetilde{\boldsymbol{N}}_1 = \arg\min_{\boldsymbol{N}} \mathrm{d}\big(\boldsymbol{f}_{\text{A}}, \boldsymbol{f}_{\text{N}}\big) \right\}, \tag{4.6}$$

which exclusively picks HR images from the original mini-batch $\mathcal{B}$. Then,

$$\mathcal{T}_{\text{HLL}} := \left\{ \big(\boldsymbol{A}, \boldsymbol{P}^{[\text{LR}]}, \widetilde{\boldsymbol{N}}_2^{[\text{LR}]}\big) \in \mathcal{T}\big(\mathcal{B}, \mathcal{B}_{\text{LR}}, \mathcal{B}_{\text{LR}}\big) \;:\; \widetilde{\boldsymbol{N}}_2^{[\text{LR}]} = \arg\min_{\boldsymbol{N}^{[\text{LR}]}} \mathrm{d}\big(\boldsymbol{f}_{\text{A}}, \boldsymbol{f}_{\text{N}}^{[\text{LR}]}\big) \right\} \tag{4.7}$$

and

$$\mathcal{T}_{\text{LHH}} := \left\{ \big(\boldsymbol{A}^{[\text{LR}]}, \boldsymbol{P}, \widetilde{\boldsymbol{N}}_3\big) \in \mathcal{T}\big(\mathcal{B}_{\text{LR}}, \mathcal{B}, \mathcal{B}\big) \;:\; \widetilde{\boldsymbol{N}}_3 = \arg\min_{\boldsymbol{N}} \mathrm{d}\big(\boldsymbol{f}_{\text{A}}^{[\text{LR}]}, \boldsymbol{f}_{\text{N}}\big) \right\}, \tag{4.8}$$

which utilize a mix of LR and HR images from the $\mathcal{B}$ and $\mathcal{B}_{\text{LR}}$ mini-batches. Lastly,

$$\mathcal{T}_{\text{LLL}} := \left\{ \big(\boldsymbol{A}^{[\text{LR}]}, \boldsymbol{P}^{[\text{LR}]}, \widetilde{\boldsymbol{N}}_4^{[\text{LR}]}\big) \in \mathcal{T}\big(\mathcal{B}_{\text{LR}}, \mathcal{B}_{\text{LR}}, \mathcal{B}_{\text{LR}}\big) \;:\; \widetilde{\boldsymbol{N}}_4^{[\text{LR}]} = \arg\min_{\boldsymbol{N}^{[\text{LR}]}} \mathrm{d}\big(\boldsymbol{f}_{\text{A}}^{[\text{LR}]}, \boldsymbol{f}_{\text{N}}^{[\text{LR}]}\big) \right\}, \tag{4.9}$$

which comprises solely LR images from $\mathcal{B}_{\text{LR}}$.

Calculating the triplet loss $\mathscr{L}_{\text{TRI}}$ (see Section 2.2.2.1) simultaneously for each set, involves assessing the feature distances among up to eight distinct images for each $\boldsymbol{A} \in \mathcal{B}$. Consequently, the aggregation of all four triplet losses relies on the octuplet $(\boldsymbol{A}, \boldsymbol{A}^{[\text{LR}]}, \boldsymbol{P}, \boldsymbol{P}^{[\text{LR}]}, \widetilde{\boldsymbol{N}}_1, \widetilde{\boldsymbol{N}}_2^{[\text{LR}]}, \widetilde{\boldsymbol{N}}_3, \widetilde{\boldsymbol{N}}_4^{[\text{LR}]})$. Therefore, the combined loss is referred to as the octuplet loss $\mathscr{L}_{\text{OCT}}$, which is computed as follows:

$$\mathscr{L}_{\text{OCT}} = \mathscr{L}_{\text{TRI}}\big(\mathcal{T}_{\text{HHH}}\big) + \mathscr{L}_{\text{TRI}}\big(\mathcal{T}_{\text{HLL}}\big) + \mathscr{L}_{\text{TRI}}\big(\mathcal{T}_{\text{LHH}}\big) + \mathscr{L}_{\text{TRI}}\big(\mathcal{T}_{\text{LLL}}\big) \,. \tag{4.10}$$

This way, Equation (4.10) encompasses all three cases: LR face pairs $\mathscr{L}_{\text{TRI}}\left(\mathcal{T}_{\text{LLL}}\right)$, CR face pairs $\mathscr{L}_{\text{TRI}}\left(\mathcal{T}_{\text{HLL}}\right)$ and $\mathscr{L}_{\text{TRI}}\left(\mathcal{T}_{\text{LHH}}\right)$, and HR face pairs $\mathscr{L}_{\text{TRI}}\left(\mathcal{T}_{\text{HHH}}\right)$. Consequently, the robustness against low- and CR face pairs can be increased and moreover the network is still forced to handle HR face pairs.

4.3.2. Experimental Settings

The experimental settings are designed to evaluate the effectiveness of the proposed OLT method on various pre-trained FR models. Recent state-of-the-art approaches such as MagFace [89], FaceTransformer [91], and *Improved Residual Layer Network* (iResNet)-50 [169] are fine-tuned with the OLT method. Additionally, the MobileNetV2 [170] architecture is included to demonstrate the benefit of the OLT method on lightweight networks. Furthermore, the in the previous section introduced approaches RAT, CLT, and MB-CLT from [1†] are also included in the experiments.

All experiments use the MS1M-V2 [47, 62] database for training and validation, comprising 5.7 M images of 87 k identities. The vast majority ($\sim 99.9\%$) is used for the fine-tuning strategy, and only $\sim 1‰$ is retained for validation. From the latter subset, 3 000 genuine and 3 000 imposter image pairs are generated to validate the FV performance during training. Due to the condition that each identity within a mini-batch must appear exactly twice (see Section 4.3.1), a specific algorithm that creates the mini-batches is employed. Images are picked from the entire dataset according to the number of unpicked images per identity. By updating the underlying probability distribution after every batch, diverse batches at the end of every epoch are ensured.

For training the MagFace [89] model, stochastic gradient descent with a learning rate of 0.001 for one epoch is utilized. The FaceTransformer [91] is fine-tuned one epoch employing the *Adaptive Moment Estimation with Weight Decay* (ADAMW) [171] algorithm ($\epsilon = 10^{-8}$), with a learning rate of 0.0005. Both networks converge already within the first epoch and are not fine-tuned further. The remaining architectures in the experiments are fine-tuned for six epochs with *Adaptive Gradient Algorithm* (ADAGRAD) [52] optimizer ($\epsilon = 1.0$) using a learning rate of 0.01, which is divided by 10 after epochs 2, 4, and 5. Due to hardware restrictions, a mini-batch size $B = 64$ for the FaceTransformer and MagFace is applied, whereas $B = 256$ for the remaining architectures. If not stated otherwise, the Euclidean distance metric is utilized in the triplet loss, while setting the margin α to 25, and using non-normalized features.

Fine-tuning on an NVIDIA RTX 3090 (24 GB) takes approximately 18 hours for the ArcFace approach with ResNet-50 [54] (3 hours per epoch), which is more time-consuming by a factor of two (1.5 hours per epoch) than pre-training with cross-entropy loss. Fine-tuning on MagFace with an iResNet-50 [169] takes 16 hours, 34 hours for FaceTransformer [91], and 2 hours for the MobileNetV2 [170] architecture.

Following [47] in data preprocessing, aligned face crops (112×112 px) with five facial landmarks are extracted with the *Multi-Task Cascaded Convolutional Networks* (MTCNN) [68] for all experiments (see also Section 2.3.2.2). Additionally, besides horizontal flipping, random brightness and saturation variation are applied as data augmentation.

For the evaluation the following datasets are utilized: LFW [31], CALFW [94], CPLFW [92], CFP-FP [93], *Celebrities in Frontal-Profile — Frontal-Frontal* (CFP-FF) [93], AgeDB [95], XQLFW [3†], and *Similar-Looking Labeled Faces in the Wild* (SLLFW) [97]. The image pairs from the benchmark datasets undergo synthetic degradation, mirroring the treatment applied to the training data. In these experiments, the first image in each pair experiences resolution reduction. The cosine distance is employed as the distance metric for all evaluations and similar to Section 4.2.3 10-fold cross-validation is applied to derive the accuracy values.

4.3.3. Results

This section first compares the CR FV accuracy of several state-of-the-art FR models before and after fine-tuning with the OLT method. Then, the performance on specific image resolutions leveraging the fine-tuned models is compared to other works in the field of CR FV. Finally, some ablations on the OLT method, including the triplets utilized in the octuplet loss function, margins, mini-batch size, feature normalization, and distance metrics are discussed.

4.3.3.1. Performance Improvements

To evaluate the effectiveness of the OLT method, the FV accuracy of several state-of-the-art FR models is compared before and after fine-tuning. The results are summarized in Table 4.3. The accuracy values are reported for the LFW dataset [31] and the XQLFW dataset [3†] (see Chapter 5). The XQLFW dataset is particularly suitable for this investigation, as the pairs in the evaluation protocol show a difference in resolution. Moreover, accuracy values are reported for various datasets with synthetically downscaled images for CR evaluation.

Without OLT, the models RAT-M, CLT-M, and MB-CLT-M from [1†] are already trained to be resolution invariant and perform best on XQLFW [3†] and very LR images (7×7 px) of the other datasets. All remaining models are very susceptible to image resolution and show a decrease in accuracy for LR images. However, although the FaceTransformer [91] network tends to be more robust than structures solely based on CNNs, its performance is still worse for very LR images.

After applying OLT, all models perform significantly better on the majority of evaluation scenarios with LR images while maintaining their performance on HR images. Only a few minor deteriorations can be observed for the FaceTransformer [91] and MagFace [89] architecture. The most considerable improvement holds for the ArcFace [47] method (baseline). OLT boosts the accuracy from 74.22% to 93.27% on the most realistic CR dataset XQLFW [3†] while even slightly surpassing the baseline accuracy on LFW [31] with 99.55%. The proposed fine-tuning method further improves the accuracy for RAT-M [1†], CLT-M [1†], and MB-CLT-M [1†] on HR images, *i.e.*, it recovers the prior drop in accuracy reported in Section 4.2.3.2. This behavior shows that the OLT method better exploits the available network capabilities and makes them more robust. With the exception of the 7×7 px resolution, the best overall performance after fine-tuning with OLT is accomplished by the FaceTransformer network. The vast increase in accuracy, which is observed on four different architectures and four unique pre-training loss functions, demonstrates that the OLT approach is universally applicable and works on various network architectures.

Table 4.3.: Improvement of cross-resolution face verification accuracy by applying the OLT strategy. MS1M-V2 is utilized for fine-tuning and evaluation is performed on LFW and XQLFW. The average accuracy is reported for various datasets with synthetical downscaled images for cross-resolution evaluation.

	Accuracy [%] for Datasets							
		mean (LFW, CALFW, CPLFW, SLLFW, CFP-FF, CFP-FP, AgeDB)						
Model	LFW	7×7 px	14×14 px	28×28 px	56×56 px	112×112 px	Mean	XQLFW
ArcFace [47]	99.50	50.82	69.83	91.49	94.67	95.01	80.36	74.22
+ OLT [$2^{\dagger}$]	99.55 (+0.05)	83.07 (+32.25)	90.72 (+20.89)	93.65 (+2.16)	94.46 (-0.21)	94.65 (-0.36)	91.31 (+10.95)	93.27 (+19.05)
RAT-M [$1^{\dagger}$]	99.30	69.93	84.46	92.69	93.89	93.83	86.96	83.60
+ OLT [$2^{\dagger}$]	99.38 (+0.08)	**84.54** (+14.61)	91.59 (+7.13)	93.91 (+1.22)	94.47 (+0.58)	94.56 (+0.73)	91.81 (+4.85)	94.20 (+10.60)
CLT-M [$1^{\dagger}$]	97.30	74.93	84.46	87.10	87.84	87.97	84.46	90.97
+ OLT [$2^{\dagger}$]	98.90 (+1.60)	84.17 (+9.24)	90.34 (+5.88)	92.21 (+5.11)	92.86 (+5.02)	92.74 (+4.77)	90.46 (+6.00)	93.47 (+2.50)
MB-CLT-M [$1^{\dagger}$]	95.87	72.44	82.40	83.89	84.05	83.82	81.32	90.82
+ OLT [$2^{\dagger}$]	98.80 (+2.93)	81.72 (+9.28)	88.41 (+6.01)	90.55 (+6.66)	90.80 (+6.75)	90.71 (+6.89)	88.44 (+7.12)	92.93 (+2.11)
ArcFace [47] (MobileNetV2)	98.85	54.38	70.18	87.57	91.19	91.55	78.97	72.73
+ OLT [$2^{\dagger}$]	98.78 (-0.07)	79.41 (+25.03)	87.03 (+16.85)	90.30 (+2.73)	91.44 (+0.25)	91.35 (-0.20)	87.91 (+8.94)	91.70 (+18.97)
FaceTransformer [91]	99.70	60.53	84.82	96.03	97.21	97.28	87.17	87.88
+ OLT [$2^{\dagger}$]	**99.73** (+0.03)	82.96 (+22.43)	**91.72** (+6.90)	**95.13** (-0.90)	**96.35** (-0.86)	**96.52** (-0.76)	**92.54** (+5.37)	**95.12** (+7.24)
MagFace[89]	99.63	52.82	73.71	94.32	96.71	96.87	82.89	76.95
+ OLT [$2^{\dagger}$]	99.63 (0.00)	81.69 (+28.87)	90.22 (+16.51)	93.84 (-0.48)	94.61 (-2.10)	94.72 (-2.15)	91.01 (+8.12)	92.92 (+15.97)

Additionally to the CR evaluation, Table 4.4 reports FV accuracy for pairs of images with the same image resolution. The accuracy values are averaged across several datasets for each image resolution. The results indicate that the baseline model is slightly worse in *Equal Resolution* (ER) FV than in the CR scenario (see Table 4.3). This discrepancy is understandable due to the reduced information content of both LR images. However, the OLT approach substantially increases the performance from 77.57% to 89.74% on average across all image resolutions. These outcomes show that the proposed OLT technique is not limited to CR scenarios and can also be applied in ER scenarios.

To get more insights across the different evaluation benchmarks and resolutions, the FV results are visualized in Figure 4.13. The chart shows the accuracy for the ResNet-50 [54] architecture pre-trained

Table 4.4.: Comparison of ResNet-50 architecture pre-trained with the ArcFace on MS1M-V2 and a fine-tuning with the OLT [$2^{\dagger}$] strategy. Equal-resolution face verification accuracy is reported on average over several synthetically downscaled datasets for specific image resolutions.

	Accuracy [%] for mean (LFW, CALFW, CPLFW, SLLFW, CFP-FF, CFP-FP, AgeDB)					
Model	7×7 px	14×14 px	28×28 px	56×56 px	112×112 px	Mean
ArcFace [47]	51.08	57.74	89.43	94.59	95.01	77.57
+ OLT [$2^{\dagger}$]	78.20 (+27.12)	88.31 (+30.57)	92.97 (+3.54)	94.56 (-0.03)	94.65 (-0.36)	89.74 (+12.17)

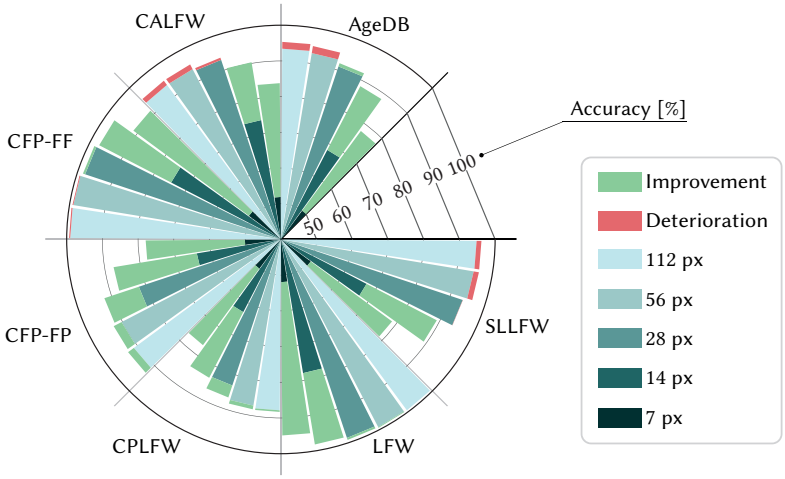


Figure 4.13.: Cross-resolution face verification accuracy comparison of ResNet-50 architecture pre-trained with the ArcFace utilizing the MS1M-V2 dataset and the fine-tuning with the OLT strategy as in [2†]. Results are reported on various datasets with synthetically downscaled image resolutions.

with the ArcFace [47] (dashed line) and the accuracy after fine-tuning with the OLT method (solid line). Improvements are visualized with green areas while deterioration is shown in red.

In general, a significant performance decrease of the baseline model on challenging datasets like AgeDB, CALFW, CPLFW, and CFP-FP that focus on age, pose, or person similarity is observed. The proposed OLT method accomplishes the best accuracy for LFW and CFP-FF with over 90% at all resolutions, which are the easiest benchmarks. In contrast, CPLFW seems to be the most challenging dataset, with a performance below 90% at all scales.

The chart also reveals that for large pose variations datasets such as CPLFW and CFP-FP, there is still a moderate increase of accuracy at 28×28 px image resolution, whereas the boost at that scale is marginal for all other datasets. Only on data with a large age gap (AgeDB and CALFW) and similar faces (SLLFW), the OLT approach marginally reduces the accuracy on 56×56 px and 112×112 px image resolution.

For a deeper analysis, Figure 4.14 analyzes the *Receiver Operating Characteristic* (ROC) (see also Section 2.3.4) of the baseline model and the OLT approach on the XQLFW dataset and specific CR scenarios of the LFW benchmark. Primarily at very low *False Acceptance Rate*s (FARs), the performance gain utilizing OLT is tremendous. While the baseline model fails for challenging situations, the fine-tuned version achieves superior results. On the XQLFW dataset, OLT increases the *True Acceptance Rate* (TAR) for very low FARs from 0% to over 65%. This improvement is similar to the behavior on the LFW dataset at 7×7 px. The effect vanishes the higher the resolution until, at 112×112 px the rates remain nearly equal. Overall, this improvement shows the benefit of OLT, especially in security applications, *e.g.*, manhunts via surveillance cameras, which require very low FARs.

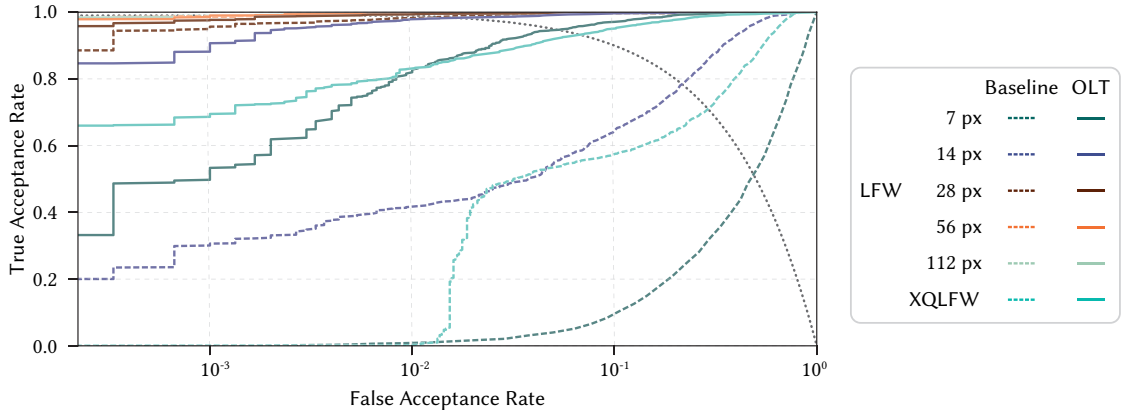


Figure 4.14.: Cross-resolution receiver operating characteristic curve comparison of the baseline model (dashed) with the OLT fine-tuning approach (solid) on the XQLFW dataset and LFW dataset with selected downscale image resolutions as in [2†]. The equal error rate is indicated by a dotted line.

Furthermore, the exact values for the TAR at specific FARs for the baseline model and the OLT approach are reported in Table 4.5. The TAR increases on the XQLFW benchmark and LR images (7×7 px and 14×14 px), while still improving in all other configurations. An improvement of up to 69.47% is achieved in the most realistic CR scenario simulated with the XQLFW database for a FAR or 0.001, *i.e.*, accepting only one *False Positive* (FP) in 1 000 attempts. This underlines the robustness of the OLT method in challenging CR scenarios and shows that the method is particularly beneficial for various real-world applications which require a high robustness against CR face pairs.

In conclusion, this analysis uncovers that OLT improves the robustness of various FR models in CR scenarios while maintaining their performance on HR images. The method is not limited to relatively simple datasets like LFW but also offers benefits to more challenging datasets, which involve, *e.g.*, large age gaps or head pose variances.

Table 4.5.: Comparison of the TAR at specific FARs for the baseline and the OLT approach on the XQLFW dataset and specific synthetically downscaled CR scenarios of the LFW benchmark.

		TAR@FAR [%] for Datasets					
		LFW [31]					
Model	FAR	7×7 px	14×14 px	28×28 px	56×56 px	112×112 px	XQLFW
ArcFace [47]	0.001	0.07	30.63	95.67	98.70	98.97	0.03
+ OLT [2†]		53.33 (+53.26)	90.70 (+60.07)	97.70 (+2.03)	99.00 (+0.30)	99.03 (+0.06)	69.50 (+69.47)
ArcFace [47]	0.01	0.80	41.77	98.30	99.47	99.53	0.27
+ OLT [2†]		82.80 (+82.00)	97.90 (+56.13)	99.40 (+1.10)	99.53 (+0.06)	99.57 (+0.04)	83.10 (+82.83)
ArcFace [47]	0.1	9.37	64.43	99.60	99.73	99.73	57.40
+ OLT [2†]		97.10 (+87.73)	99.67 (+35.24)	99.73 (+0.13)	99.77 (+0.04)	99.77 (+0.04)	95.07 (+37.67)

Table 4.6.: Cross-resolution face verification accuracy evaluated on the LFW dataset for particular image resolutions. The OLT fine-tuning is performed with MS1M-V2 dataset.

	Accuracy [%]				
	LFW				
Model	8 × 8 px	12 × 12 px	16 × 16 px	32 × 32 px	High Resolution
Lai and Lam [159]	**94.8**	97.6	98.2	–	99.10 (128 × 128 px)
Sun *et al.* [157]	90.0	94.9	97.2	–	99.10 (112 × 96 px)
DCR [152]	93.6	95.3	96.6	–	98.70 (112 × 96 px)
TCN [160]	90.5	94.7	97.2	–	98.80 (112 × 96 px)
Ge *et al.* [156]	–	–	85.87	89.72	97.15 (224 × 224 px)
ArcFace [47] + OLT [2†]	90.38	96.88	98.28	99.48	94.59 (112 × 112 px)
FaceTransformer [91] + OLT [2†]	94.02	**98.17**	**99.08**	**99.57**	99.63 (112 × 112 px)

4.3.3.2. Comparison with other Approaches

After demonstrating that the OLT improves the robustness of various FR models in CR scenarios, a comparison with other state-of-the-art CR methods is drawn. For this purpose, the two best performing approaches (FaceTransformer + OLT and MagFace + OLT) on LFW are evaluated with particular resolutions to match the evaluation conditions of other approaches and enable a direct comparison.

The results are reported in Table 4.6 and show that the proposed approaches outperform all other methods except for 8 × 8 px image resolution, where Lai and Lam [159] achieve a higher accuracy. However, a notable drawback of their approach is the weak performance for HR images. We must interpret the results of Ge *et al.* [156] carefully as their approach is based on a teacher model that performs worse on HR images (only 97.15%) and they report numbers of specific models for each image resolution. Moreover, the training resolution is inconsistent across the compared methods and can lead to slight deviations. Concluding, these results provide a reasonable classification of OLT as the state-of-the-art and underline its advantages.

In addition, the OLT fine-tuning strategy is compared to the approach of Terhörst *et al.* [163]. While they perform worse on XQLFW (83.95%), they report a slightly higher accuracy on LFW (99.83%) and much better results on AgeDB (98.50%) and CFP-FP (98.7%). However, they aim at general quality-robust FR encompassing resolution, age, and pose. In contrast, this work focuses exclusively on the images' resolution; hence, this is not a fair comparison and should be considered with caution.

4.3.3.3. Ablation Studies

In the following, multiple ablations are conducted to understand the influence of the loss terms, distance metric, feature normalization, margins, and mini-batch size in the OLT approach.

Loss Terms. As described in Section 4.3.1 the octuplet loss function consists of four different triplet loss functions. Each term affects the overall performance, and hence, experiments are conducted to obtain the contribution of each term. In this study, the Euclidean distance and no feature normalization

Table 4.7.: Cross-resolution face verification accuracy for different configurations of the triplet loss terms in the octuplet loss function. The pre-trained baseline model with OLT fine-tuning (both utilizing MS1M-V2) is evaluated on several datasets and different image resolutions.

Triplets utilized in Octuplet Loss				Accuracy [%] for Datasets							
					mean (LFW, CALFW, CPLFW, SLLFW, CFP-FF, CFP-FP, AgeDB)						
$\mathcal{T}_{HHH}$	$\mathcal{T}_{HLL}$	$\mathcal{T}_{LHH}$	$\mathcal{T}_{LLL}$	LFW	7×7 px	14×14 px	28×28 px	56×56 px	112×112 px	Mean	XQLFW
✓	✓	✓	✓	99.55	83.07	**90.72**	93.65	94.46	94.65	**91.31**	93.27
✓				99.42	61.17	78.77	92.21	94.77	95.11	84.40	80.98
	✓			99.52	79.93	89.39	93.57	94.57	94.64	90.42	92.15
		✓		99.48	80.10	89.48	93.53	94.56	94.62	90.46	92.32
			✓	98.95	82.49	88.80	91.40	92.21	92.06	89.39	93.20
✓	✓			99.57	79.60	89.50	93.93	**95.19**	**95.18**	90.68	91.93
✓		✓		**99.58**	79.42	89.75	93.97	95.11	95.14	90.68	92.18
✓			✓	99.57	81.66	90.05	93.66	94.65	94.84	90.97	92.77
	✓	✓		99.45	80.51	89.90	93.52	94.48	94.54	90.59	92.55
✓	✓	✓		**99.58**	80.60	90.11	**93.94**	94.97	94.94	90.91	92.55
✓	✓		✓	99.50	82.92	**90.72**	93.70	94.52	94.60	91.29	93.30
✓		✓	✓	99.37	82.49	90.35	93.58	94.47	94.56	91.09	**93.48**
	✓	✓	✓	99.20	**83.09**	90.62	93.18	93.93	93.95	90.96	93.42

is used. As depicted in Table 4.7 the reference point is the best mean accuracy across all datasets and image resolutions, which is obtained by including all triplet loss terms: $\mathscr{L}_{TRI}(\mathcal{T}_{HHH})$, $\mathscr{L}_{TRI}(\mathcal{T}_{HLL})$, $\mathscr{L}_{TRI}(\mathcal{T}_{LHH})$, and $\mathscr{L}_{TRI}(\mathcal{T}_{LLL})$.

As expected, utilizing only $\mathscr{L}_{TRI}(\mathcal{T}_{HHH})$ leads to the worst results on the XQLFW dataset and images at image resolutions (7×7 px, 14×14 px, and 28×28 px), which is reasonable since no LR images are included in the loss. Interestingly, this configuration does not improve the accuracy on the LFW dataset, which can be considered as an HR benchmark, whereas it does improve the accuracy on average across all datasets. Two plausible explanations could be: 1) The performance is already saturating for LFW. 2) There might be a few LR images in the LFW dataset, although exclusively HR are expected in that database. However, not utilizing $\mathscr{L}_{TRI}(\mathcal{T}_{HHH})$ leads in all cases to a decrease in performance on the LFW dataset. Consequently, this term is essential to constrain the network and not focus entirely on LR. In contrast, utilizing solely the $\mathscr{L}_{TRI}(\mathcal{T}_{LLL})$ significantly improves the performance on LR images. But this comes with a decrease of the accuracy on HR images and thus is either not considered preferable.

Considering exclusively the $\mathscr{L}_{TRI}(\mathcal{T}_{HLL})$ and $\mathscr{L}_{TRI}(\mathcal{T}_{LHH})$, or the inclusion of both terms, leads to a moderate increase of robustness to image resolution but also comes with the trade-off of reducing the accuracy on HR images. A similar effect occurs for the combination of $\mathscr{L}_{TRI}(\mathcal{T}_{HHH})$ and $\mathscr{L}_{TRI}(\mathcal{T}_{LLL})$. Interestingly, the $\mathscr{L}_{TRI}(\mathcal{T}_{HHH}) + \mathscr{L}_{TRI}(\mathcal{T}_{HLL})$ or $\mathscr{L}_{TRI}(\mathcal{T}_{HHH}) + \mathscr{L}_{TRI}(\mathcal{T}_{LHH})$ configuration yields the best performance on intermediate image resolutions (28×28 px and 56×56 px). Furthermore, experiments with three triplet loss terms reveal that removing $\mathscr{L}_{TRI}(\mathcal{T}_{HHH})$ or $\mathscr{L}_{TRI}(\mathcal{T}_{LLL})$ leads to a marginal decline in performance.

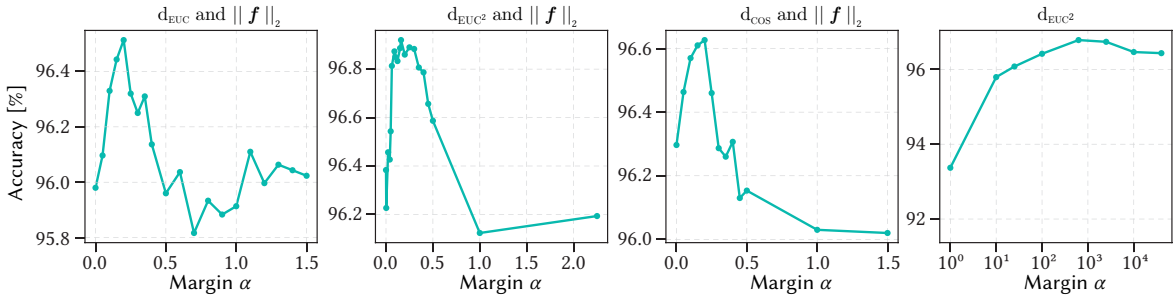


Figure 4.15.: Face verification accuracy utilizing particular margins for different distance metric configurations in the octuplet loss function as in [2†]. The pre-trained baseline model with OLT fine-tuning (both utilizing MS1M-V2) is evaluated.

This analysis of individual loss terms demonstrates that each term is crucial to the overall performance. The advantage of incorporating both $\mathscr{L}_{TRI}(\mathcal{T}_{HLL})$ and $\mathscr{L}_{TRI}(\mathcal{T}_{LHH})$ is modest, as they both link HR images with LR images. Nevertheless, this benefit should not be underestimated; including both terms is essential for achieving total improvement.

Distance Metric and Feature Normalization As described in Section 4.3.1, the OLT approach follows the work of Hermans *et al.* [46] and utilizes the Euclidean distance metric d_{EUC} without feature normalization for the octuplet loss function. However, as proposed in other works [44, 77, 172, 173], experiments with a cosine d_{COS} and Euclidean-squared distance metric d_{EUC^2} are conducted. Additionally, feature normalization $\|\cdot\|_2$ is applied to the extracted features $\boldsymbol{f}$ before calculating the distance. The results are reported in Table 4.8.

Utilizing d_{EUC} with no feature normalization turns out to be superior on the average across all datasets and image resolutions, and also on very LRs (7×7 px and 14×14 px), which is the reference point for this study. However, utilizing d_{EUC^2} together with feature normalization leads to the best performance on the LFW and XQLFW dataset and also for intermediate and HRs (28×28 px, 56×56 px, and 112×112 px) at mean across several datasets. Depending on the application, this configuration

Table 4.8.: Cross-resolution face verification accuracy for different distances metrics and with or without feature normalization. The pre-trained baseline model with OLT fine-tuning (both utilizing MS1M-V2) is evaluated on several datasets and different image resolutions.

Distance Metric and Normalization				Accuracy [%] for Datasets							
					mean (LFW, CALFW, CPLFW, SLLFW, CFP-FF, CFP-FP, AgeDB)						
d_{EUC}	d_{EUC^2}	d_{COS}	$\|\boldsymbol{f}\|_2$	LFW	7×7 px	14×14 px	28×28 px	56×56 px	112×112 px	Mean	XQLFW
✓				99.55	**83.07**	**90.72**	93.65	94.46	94.65	**91.31**	93.27
✓			✓	99.60	79.53	88.64	93.59	94.80	94.95	90.30	93.27
	✓		✓	**99.63**	80.08	89.63	**93.83**	**95.11**	**95.24**	90.78	**93.58**
	✓			99.53	81.36	89.75	93.55	94.61	94.71	90.80	93.23
		✓	✓	99.58	79.57	88.85	93.70	94.92	95.03	90.41	93.27

might be preferred. Due to the utilization of d_{COS} in our evaluation protocols, one would expect that utilizing this metric in our octuplet loss fine-tuning strategy leads to the best results. However, this is not the case, as seen in the bottom row of Table 4.8. The performance on LFW and XQLFW is similar, but for lower image resolutions, fine-tuning with d_{COS} leads to minor accuracy improvements.

Since the configurations in Table 4.8 highly affect the magnitude of the margin, the best margins are empirically determined for each configuration. The achieved accuracy dependent on the margin is reported in Figure 4.15. Note that all axes of the charts are scaled differently. To be precise, the best margin for the d_{EUC} and $\|\boldsymbol{f}\|_2$ configuration is 0.2, for the d_{EUC^2} and $\|\boldsymbol{f}\|_2$ configuration 0.175, and 0.2 for the d_{COS} with $\|\boldsymbol{f}\|_2$ configuration. For the d_{EUC^2} and no feature normalization configuration, the margin is due to the square term very high and 600 turned out to be the best value.

Margin and Mini-Batch Size As a final ablation study, the influence of the mini-batch size B for different margins α in the range between 1 and 500 is investigated to understand the impact of the hard sample mining algorithm. The results are reported in Figure 4.16.

The best accuracy is achieved for a margin of 25 across all tested mini-batch sizes, which indicates that the margin is independent of the mini-batch size. The performance decreases for smaller mini-batch sizes and lower or larger margins. This effect is unsurprising since a larger mini-batch size increases the probability of the hard sample mining algorithm finding even more challenging samples. Batches containing less than 64 samples lead to even worse accuracy, and hence, they are not further investigated in this work. Due to hardware limitations, experiments with larger mini-batch size than 256 could not be conducted. Nevertheless, one would expect this trend to continue, with an increasing computational cost for the mining algorithm.

The ablation studies conducted on the OLT approach highlight the intricate balance between loss terms, distance metrics, feature normalization, margins, and mini-batch size in optimizing FV

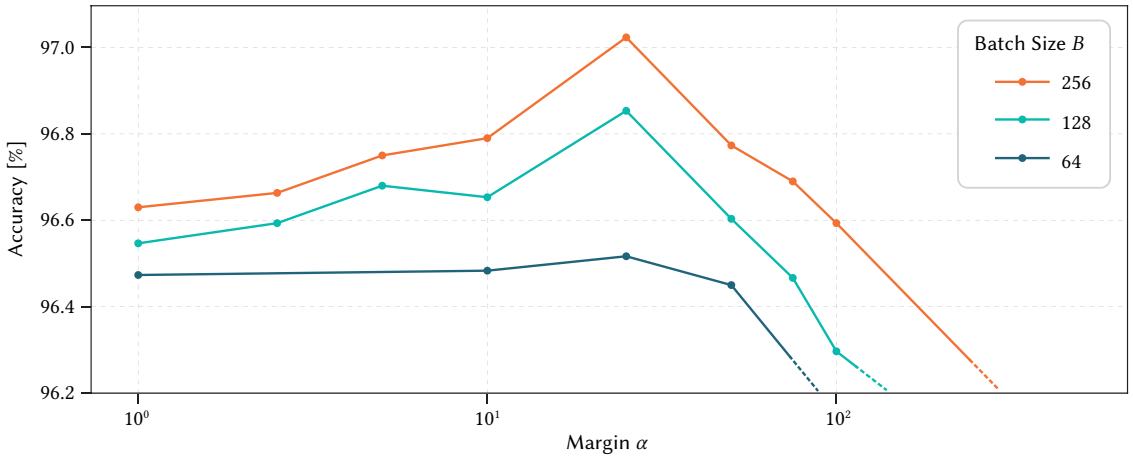


Figure 4.16.: Average face verification accuracy utilizing the OLT method on synthetically downscaled images from the LFW dataset at resolutions $r \in \{7, 14, 28, 56, 112\}$, tested on the validation subset of MS1M-V2. Adapted from [2†]

accuracy across different resolutions. Each parameter plays a critical role in the performance of the system, indicating a complex interplay that requires careful adjustment to achieve optimal results. Specifically, the experimentation reveals that while some configurations excel in LR scenarios, others are more suited for higher resolutions, underscoring the necessity for a nuanced approach in tuning these parameters. Ultimately, these findings underscore the importance of a holistic understanding of the OLT approach's components in enhancing CR FV accuracy, offering valuable insights for future research and application in biometric security systems.

4.4. Conclusion

The exploration of the impact of varying image resolutions on FV performance, as delineated in this study, underscores a critical challenge within the domain of classical FR networks: The lack of scale invariance. It has been elucidated that a deviation from the image resolution, as encountered in the corresponding training dataset of a network, precipitates a notable decrement in performance, especially at lower image resolutions. This revelation posits a significant impediment to the efficacy of conventional FR methodologies, thereby necessitating the inception of novel strategies aimed at ameliorating this deficiency.

In response to this exigency, the investigation introduces three innovative training strategies, published in [1†], designed to engender the direct learning of scale-invariant features. The first method, RAT, involves the incorporation of both LR and HR images within training mini-batches in equivalent proportions. The second method, CLT, employs a siamese network architecture derived from other state-of-the-art networks, further augmented by an additional contrastive loss function targeting the minimization of feature distances between LR and HR variants of identical images. The third method MB-CLT extends the CLT approach for multiple-resolutions by adding specific branches for distinct image resolutions, and respectively adding several contrastive loss functions to the total loss. These approaches collectively manifest a significant uplift in CR FV accuracy across a spectrum of standard datasets, thereby illuminating the potential for a single model to adeptly navigate images across arbitrary scales.

Moreover, the study proposes a fine-tuning strategy leveraging the octuplet loss function (OLT), published in [2†], aimed at bolstering the robustness of existing models to fluctuations in image resolution. This strategy, through the simultaneous application of four distinct triplet loss terms to both LR and HR images, exploits the nuanced relationships both within and across identities and resolutions. It emerges that this approach not only circumvents the necessity for exhaustive re-training but also exhibits a marked enhancement in robustness against resolution variance without compromising the accuracy on HR images. As of writing, the OLT approach is the state-of-the-art for CR FV on the XQLFW dataset, achieving an accuracy of 95.12%.

The ablations reveal, however, that the performance of the OLT approach is highly dependent on the choice of triplet loss terms, distance metrics, feature normalization, margins, and mini-batch size, underscoring the necessity for a nuanced understanding of these parameters to optimize FV accuracy across different resolutions.

The culmination of these efforts portends a significant advancement towards the development of universal, resolution-independent FR systems. The findings from this study not only contribute valuable insights into the mechanisms underpinning scale variance in FR networks but also lay the groundwork for future explorations into the domain of masked FR, amongst others. Through the dissemination of the methodologies and findings under an open license, the study fosters a foundation for the broader community to build upon, thereby facilitating the evolution of more robust FR systems capable of navigating the complexities presented by varying image resolutions.

"I think it's much more interesting to live not knowing than to have answers which might be wrong."

— Richard Feynman

Cross-Resolution Face Verification Benchmark Dataset

This chapter introduces a *Face Verification* (FV) benchmark protocol constructed from the well-known *Labeled Faces in the Wild* (LFW) database: *Cross-Quality Labeled Faces in the Wild* (XQLFW), which was published at the IEEE conference series on Automatic Face and Gesture Recognition (2021) [3†]. This dataset focuses on significant image quality variations and thus, evaluates *Face Recognition* (FR) systems on their robustness against image quality. After discussing relevant related work, the methodology of the construction of the XQLFW dataset and verification protocol is presented. The results section compares the new dataset to the original LFW and benchmarks several state-of-the-art approaches on the new dataset. The chapter concludes with a discussion of the results and the implications of the new dataset.

5.1. Related Work

A majority of works related to *Cross Resolution* (CR) FV [1†, 145, 146, 154, 155, 156, 158, 174, 175] focus on a simulation of *Low Resolution* (LR) images for the evaluation of their approaches. They simulate a lower image resolution by down- and up-scaling with bicubic or bilinear kernels. However, several studies [176, 177, 178, 179] on image *Super Resolution* (SR) showed that native LR images differ from synthetically generated images. Moreover, scaling kernels, used for synthetic downscaling vary across software packages and make a fair comparison impossible.

The SCFace [180] and the COX-S2V [181] databases are the most relevant databases for CR FV evaluations. The SCFace database contains 4 160 images from 130 unique identities, and the images are captured with different cameras, resolutions and at different distances. They also include infra-red images, which are taken under low lighting conditions. Moreover they provide images with different poses. The COX-S2V database contains LR videos in an uncontrolled environment from 1 000 unique identities and *High Resolution* (HR) still images in a controlled environments. The images are captured with a surveillance camera and contain a variety of poses, expressions, and occlusions. However, they aim on the still-to-video problem and do not provide any FV protocol. The ChokePoint [182] database consists of 48 video sequences and 64 204 of 54 persons walking through a portal. Faces have variations in terms of illumination conditions, pose, and sharpness. However, the small amount

of identities and the controlled scenario make it unsuitable for a general evaluation of the robustness of FV models towards images resolution.

Considering LR FV in general, there are few databases which contain solely LR images. The TinyFace [183] database includes 169 403 native LR images of 5 139 subjects and contains an identification evaluation protocol designed to test FR approaches on LR images. The *Unconstrained College Students* (UCCS) [184] database contains more than 70 k images of 1 732 students and staff members from the University of Colorado at Colorado Springs. The images are captured with a surveillance camera at the campus and contain a variety of poses, expressions, and occlusions. However, this database has been discussed controversial due to privacy concerns and has been removed from the public domain in 2019.

In connection to the LFW database, exhaustive work was done by Deng *et al.*, which released five datasets based on the LFW database: *Similar-Looking Labeled Faces in the Wild* (SLLFW) [97], *Cross-Age Labeled Faces in the Wild* (CALFW) [94], *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92], *Masked Labeled Faces in the Wild* (MLFW) [99], and *Transferable Adversarial Labeled Faces in the Wild* (TALFW) [98] (see also Section 2.3.3). These datasets focus on specific properties like age, pose, or transferable adversarial attacks and enhance the difficulty of the original LFW database. Those datasets are frequently used to evaluate FR systems in more challenging scenarios and re-use the original evaluation protocol of LFW. However, they do not lay a focus on CR evaluations. Another work based on the LFW database is the work of Hörmann *et al.* [10†], which modified LFW to contain occlusions and evaluate FR systems on their robustness against occlusions.

5.2. Methodology

Since the inherent quality or raw-resolution difference within the pairs of the LFW database is tiny, the idea is to synthetically degrade images and create a modified XQLFW database to enlarge the raw-resolution variance and thus being able to test the robustness of FR systems against cross-image resolution. In the publication of XQLFW [3†], the term quality is used to describe the raw-resolution of an image. However, this work introduced the term 'raw-resolution' to describe the difference of image quality in terms of their original image resolution and avoid confusion with the more general term 'image quality' which is used in the context of image quality assessment and includes a variety of factors like sharpness, contrast, and noise. Hence, in the following sections, the term low raw-resolution or LR is used instead of 'low quality' to be more precise. The following sections describe the assessment of image raw-resolution the synthetic image quality degradation, and the evaluation protocol for the construction of the XQLFW dataset.

5.2.1. Image Quality Assessment

Besides a variety of existing approaches for referenceless image quality assessment and specific facial image quality assessment methods, this work utilizes the *Blind Referenceless Image Spatial Quality Evaluator* (BRISQUE) [36] and *Stochastic Embedding Robustness Face Image Quality* (SER-FIQ) [38] metrics to assess the quality of the images. The interested reader is referred to review of Kamble and

Bhurchandi [37] for a comprehensive overview of no-reference image quality assessment methods in general and specifically on faces to the work of Khryashchev *et al.* [185] and Yang *et al.* [186].

First, the BRISQUE and SER-FIQ scores for every single image of the LFW database are calculated. To mitigate effects from the background around the face, the scores are calculated after pre-processing, *i.e.*, cropping and aligning images with *Multi-Task Cascaded Convolutional Networks* (MTCNN) [68] as in [47] (see also Section 2.3.2.2). While BRISQUE measures the visual quality of an image, SER-FIQ evaluates the quality of the face itself via facial feature assessment (*i.e.*, occlusions or extreme head poses) which result in a meaningless identity feature and thus also reflect poor quality. A calculated correlation coefficient close to 0 (-0.021) between both metrics in the entire LFW database proves the independence of both metrics and makes a combination reasonable. To allow a meaningful combination of both scores, each score is normalized to the same value range $[0, 1]$.

The normalized BRISQUE score $Q_{\mathrm{B}}(\cdot)$ for a given image $\boldsymbol{I}$ is calculated as follows:

$$Q_{\mathrm{B}}(\boldsymbol{I}) = \frac{\left[\left[0, \tilde{Q}_{\mathrm{B}}(\boldsymbol{I})\right]_{+}, 100\right]_{-}}{100}, \tag{5.1}$$

where $\tilde{Q}_{\mathrm{B}}(\cdot)$ is the original BRISQUE score. According to [36] the lowest possible value of BRISQUE is 0, and the highest possible value is 100.

Following the same procedure, the normalized SER-FIQ score $Q_{\mathrm{S}}(\cdot)$ for a given image $\boldsymbol{I}$ is calculated as follows:

$$Q_{\mathrm{S}}(\boldsymbol{I}) = \frac{\left[\left[0.78, \tilde{Q}_{\mathrm{S}}(\boldsymbol{I})\right]_{+}, 0.91\right]_{-}}{0.13}, \tag{5.2}$$

where $\tilde{Q}_{\mathrm{S}}(\cdot)$ is the original SER-FIQ score. Experiments show that the lowest possible value of SER-FIQ is 0.78, and the highest possible value is 0.91 in the LFW dataset.

The normalized combined image quality score $\tilde{Q}(\cdot)$ for a given image $\boldsymbol{I}$ is then calculated using both normalized scores as follows:

$$Q(\boldsymbol{I}) = \frac{1 - Q_{\mathrm{B}}(\boldsymbol{I}) + Q_{\mathrm{S}}(\boldsymbol{I})}{2}. \tag{5.3}$$

5.2.2. Synthetic Image Quality Degradation

Several works [176, 177, 178, 179] argue that simple downscaling images with, *e.g.*, bicubic or bilinear kernels, is not sufficiently reflecting native low image resolution. In contrast to previous works, this work applies the proposed method from Bell *et al.* [179]. They first blur images with random variations of 21×21 px an-isotropic Gaussian kernels and then downscale the images with Nearest Neighbor downscaling. Scaling factors are randomly chosen from a list of $\{3, 4, 5, 6, 7, 8, 10, 12, 14, 16\}$.

To generate the XQLFW pool of images with various quality scores Q, this synthetic image degradation is only applied to images with a quality score Q already below a certain threshold of 0.85, to assure that only images with a rather low quality are degraded further. Additionally, the number of

degraded images within an identity does not exceed half the number of images of that identity. This procedure assures that a certain amount of high quality images remains in the pool for each identity.

This generated pool of images is then further utilized to construct the XQLFW evaluation protocol.

5.2.3. Construction of Evaluation Protocols

To make the benchmark easy-integrable, the characteristics of the original LFW evaluation protocol by Huang *et al.* [31] are maintained, *i.e.*, the same protocol is used. A *Semi Cross-Quality Labeled Faces in the Wild* (SXQLFW) evaluation protocol is constructed from the original images of the LFW whereas the XQLFW protocol is constructed utilizing the synthetically degraded images (see Section 5.2.2). The image pairs for testing FV is described in the following:

Genuine Pairs. 3 000 pairs are formed by iteratively picking randomly an identity which has at least two images in the database and then randomly selecting a second image from the given identity. During this process, the criterion of $\Delta Q > 0.15$ is assured by using the quality scores Q for each image calculated in Section 5.2.1. Additionally, if the same image pair is selected again, the process is repeated until a unique pair is found.

Imposter Pairs. 3 000 pairs are formed by iteratively picking randomly two different identities out of all identities and then randomly selecting one image from each identity. During this process, it is assured that every pair is unique and meets the criterion of $\Delta Q > 0.15$. Moreover, similar to [92] and [94], gender and race are forced to be equal by using the attributes provided in LFW.

Since the LFW database contains 13 233 images from 5 749 unique identities and the number of images per identity varies from 1 up to 530, this process results in a considerable number of unique

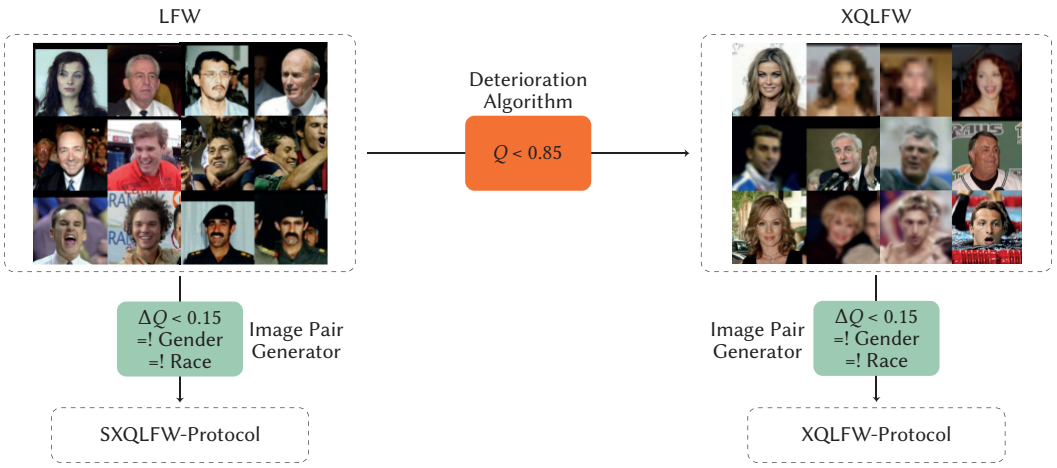


Figure 5.1.: Flowchart for constructing the XQLFW pool and the proposed XQLFW and SXQLFW evaluation protocols as in [3†].

identities and images for the XQLFW evaluation protocol. The overall pipeline for the construction is depicted in Figure 5.1.

5.3. Results

This section first analyzes the quality scores Q of the LFW image pool and the synthetically degraded XQLFW image pool. It further compares the intra-pair quality score differences ΔQ of the various datasets. The section then benchmarks several state-of-the-art FR approaches on the XQLFW and SXQLFW evaluation protocols and compares the results to the original LFW database.

5.3.1. Quality Scores

The first experiment in this section aims to analyze the correlation between the quality scores Q and the applied downscaling factor for the synthetic degradation of images. Therefore, the entire LFW dataset is synthetically degraded with the method described in Section 5.2.2. The quality score distributions at each downscaling factor are then analyzed in Figure 5.2. Q is constantly decreasing with the downscaling factor, which underlines the effectiveness of the synthetic degradation.

First, the quality scores Q of the LFW image pool and the synthetically degraded XQLFW image pool are considered. Figure 5.3 depicts the distributions of the quality scores calculated for each image in the pool. The LFW database mainly contains images with scores Q in the range of 0.7 to 0.9. In contrast, XQLFW principally consists of two groups of images with Q in the range of 0.7 to 0.9 corresponding to the non-degraded images and additionally another group between 0.25 and 0.65,

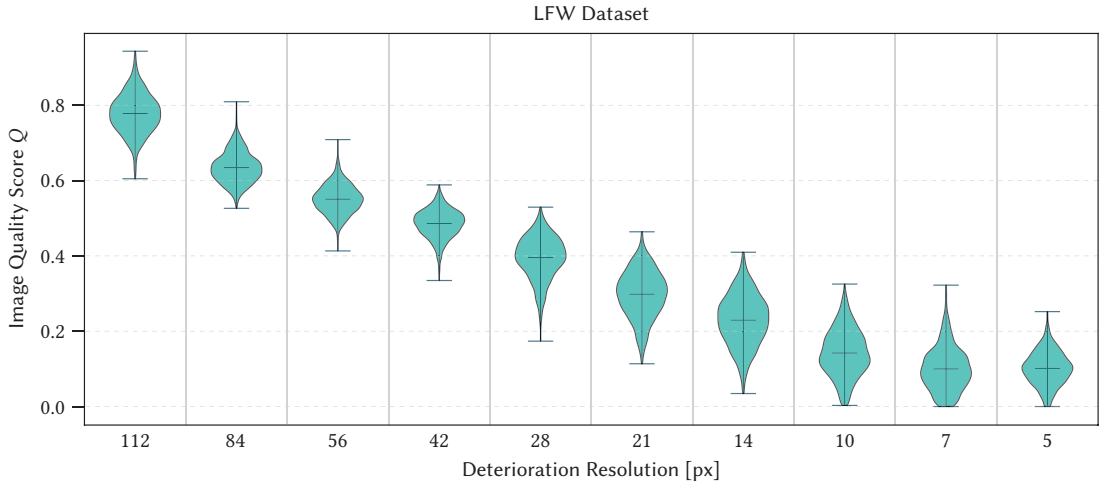


Figure 5.2.: Quality score distribution of the image from LFW dataset degraded to particular image resolutions. Adapted from supplementary material of [3†].

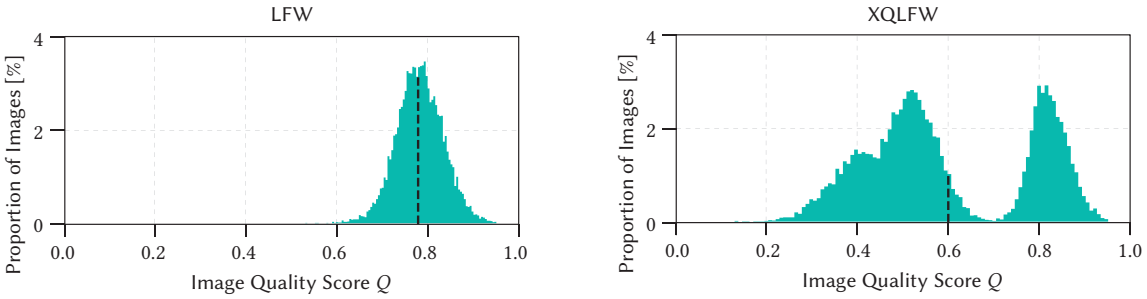


Figure 5.3.: Quality score distributions of the images in LFW on the left, and the images in the degraded XQLFW dataset on the right. Adapted from [3†].

necessarily synthetically degraded images. This pool enables the formation of imposter and genuine image pairs with a more extensive quality score difference.

The average quality score $\overline{Q}$ of the LFW image pool is 0.85, and 0.75 for the XQLFW image pool. This significant difference underlines that the degradation of the images was successful in terms of the quality score.

After employing the evaluation protocol construction, *i.e.*, defining the genuine and imposter pairs, the average quality score difference $\Delta\overline{Q}$ can be calculated for the SXQLFW and XQLFW evaluation protocols. Table 5.1 compares the new evaluation protocols to the original LFW and various other datasets that focus on specific properties like age, pose, or similarity. The original LFW database holds a relatively low average $\Delta\overline{Q}$ of 0.056. In contrast, SXQLFW and XQLFW contain a significantly larger average $\Delta\overline{Q}$ of 0.177 and 0.327, respectively. Interestingly, $\Delta\overline{Q}$ of CPLFW is slightly larger than the one of LFW, which underlines the susceptibility of the SER-FIQ metric against extreme head pose variations. CALFW and SLLFW contain a relatively low average $\Delta\overline{Q}$ of 0.078 and 0.056, respectively.

Moreover, the number of unique identities and images for the resulting evaluation protocols are compared to LFW, CALFW, CPLFW, and SLLFW (see Table 5.1). Due to the relatively small image quality variations within the LFW database, the construction of SXQLFW leads to excessive use of particular identities and images. *E.g.*, mainly the rare identities with large quality variation within the images are preferably chosen for genuine pairs. Consequently, the SXQLFW evaluation protocol contains only 2 450 identities and 4 395 unique images, thus lacking generality. CALFW, CPLFW, and

Table 5.1.: Number of unique identities, images, and intra-pair quality score differences across various datasets.

	Accuracy [%] for Datasets					
Metric	LFW [31]	CPLFW [92]	CALFW [94]	SLLFW [97]	SXQLFW [3†]	XQLFW [3†]
#Identities	4 281	2 296	2 997	2 810	2 450	3 743
#Images	7 701	5 984	7 167	6 091	4 395	7 263
$\Delta\overline{Q}$	0.056	0.078	0.046	0.054	0.177	0.327

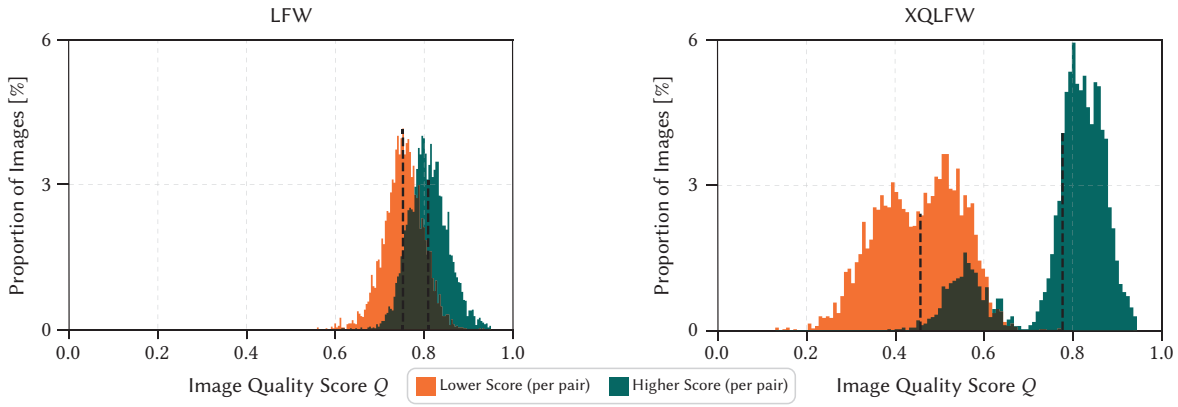


Figure 5.4.: Intra-pair quality score distribution Q for LFW and XQLFW database as in [3†]. The lower score per pair is highlighted in orange and the higher score in cyan.

SLLFW similarly have fewer unique identities and images compared to LFW. However, our proposed evaluation protocol (XQLFW), derived from the degraded database, contains 3 743 individual identities and 7 263 images, topmost among other LFW derivatives mentioned in Table 5.1.

To get a deeper insight into the quality score differences, Figure 5.4 depicts the image quality distribution for LFW and XQLFW again, but calculated for each image in the evaluation protocol, *i.e.*, the score for each image in the defined genuine and imposter pairs. In contrast to LFW, one can see the widening gap of scores in XQLFW, which implicates a significantly larger quality score difference. From the left histogram, one can see that image pairs existent in the LFW protocol, which both have a relatively low or high-quality score. By not strictly picking non-degraded images from the XQLFW image pool to form a pair, it is also guaranteed that a certain amount of CR pairs with both images having relatively low quality are included. But still, a minimum quality score difference of 0.15 is assured.

Another view of the quality score differences is given in Figure 5.5, which shows the distribution of the absolute difference in image quality per pair. The left histogram depicts the distribution of the absolute difference in image quality for the LFW database, and the right histogram depicts the distribution for the XQLFW database. The right histogram shows a significantly larger number of pairs with a larger quality score difference, which underlines the effectiveness of the synthetic degradation. Moreover, two example pairs for each database are depicted, which show the difference in image quality and the corresponding quality scores Q of the images.

5.3.2. Face Verification Benchmark

This section employs the XQLFW and SXQLFW evaluation protocols to benchmark a range of state-of-the-art FV approaches. The analysis aims to highlight the distinctions between these protocols and the performance metrics obtained using the LFW database. Table 5.2 reveals that although most of the tested models show a high accuracy on the LFW database, they significantly differ in their

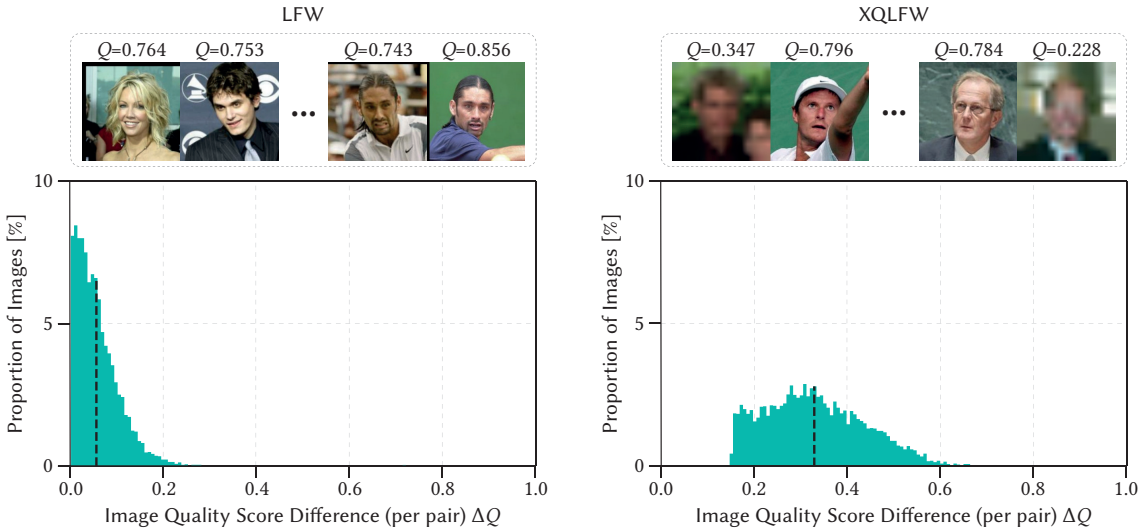


Figure 5.5.: Two example image pairs of the LFW and XQLFW database with corresponding quality scores on the top. The bottom section shows the intra-pair quality score differences of both databases. Adapted from [3†]

performance on the XQLFW and SXQLFW evaluation protocols. The small decrease in performance for SXQLFW underlines the requirement of further degradation of images to measure the susceptibility of FR systems to image quality.

While the performance on XQLFW drops substantially for ArcFace [47] and MagFace [89], the decrease of accuracy for *Resolution Augmentation Training* (RAT)-M, *Contrastive Loss Training* (CLT)-M, and *Multi-Branch Contrastive Loss Training* (MB-CLT)-M from [1†] is moderate. This is reasonable,

Table 5.2.: Face verification accuracy for several state-of-the-art approaches on LFW and the proposed SXQLFW and XQLFW evaluation protocols. All models are trained with MS1M-V2 [47, 62]. The absolute decrease in relation to the LFW database is denoted in parentheses.

	Accuracy [%] for Datasets		
Model	LFW [31]	SXQLFW [3†]	XQLFW [3†]
ArcFace [47]	99.50	99.13 (-0.37)	74.22 (-25.28)
MagFace [89]	99.63	**99.35** (-0.28)	76.95 (-22.68)
FaceTransformer [91]	99.70	**99.35** (-0.35)	87.90 (-11.80)
ProdPoly [107]	99.80	–	86.90 (-12.90)
RAT-M [1†]	99.30	99.10 (-0.20)	83.60 (-15.70)
CLT-M [1†]	97.30	96.50 (-0.80)	90.97 (-6.33)
MB-CLT-M [1†]	95.87	94.77 (-1.10)	90.82 (-5.05)
ArcFace + OLT [2†]	99.55	–	93.27 (-2.28)
FaceTransformer + OLT [2†]	**99.73**	–	**95.12** (-4.61)

due to the fact, that those models are specifically designed to be robust against varying image resolutions in the testing set. However, their performance on the LFW database is considerably low, which underlines the need of further improvements in designing resolution-robust models, while maintaining high accuracy on HR images. As introduced in Section 4.3, the *Octuplet Loss Training* (OLT) approach [2†] is able to achieve a high accuracy on the XQLFW database, and still is performing competitively on the LFW database as can be seen in the bottom two lines in Table 5.2. Interestingly, the performances of the FaceTransformer approach [91] and ProdPoly [107] model with no fine-tuning are remarkably high on the challenging XQLFW protocol. This could reveal, that the concept of Transformer [187] architectures, introduced in the speech recognition domain, is less susceptible to image resolution or quality than classical *Convolutional Neural Network* (CNN) architectures.

These observations support the conclusion that models exhibit varying levels of robustness to image quality variations. Consequently, the proposed datasets serve as a valuable resource for assessing the CR performance of FV systems.

Finally, to analyze the accuracy more in detail, Figure 5.6 shows the *Receiver Operating Characteristic* (ROC) curve for several state-of-the-art models on LFW and the proposed XQLFW evaluation protocol. The chart reveals that superior FR performance on standard datasets like LFW is not necessarily correlated to the challenging and more realistic XQLFW. The FaceTransformer approach outperforms all other models, which are not specifically designed to be robust against varying image resolutions, on XQLFW. Focusing on the RAT-M model, the horizontal line behavior at a *True Acceptance Rate* (TAR) of approximately 0.7 on the XQLFW dataset is remarkable. To interpret this behavior, the results in Section 4.2.3.2 are considered, which show that the feature distances of both genuine and imposter CR image pairs from the RAT-M model are significantly larger for very low image resolutions,

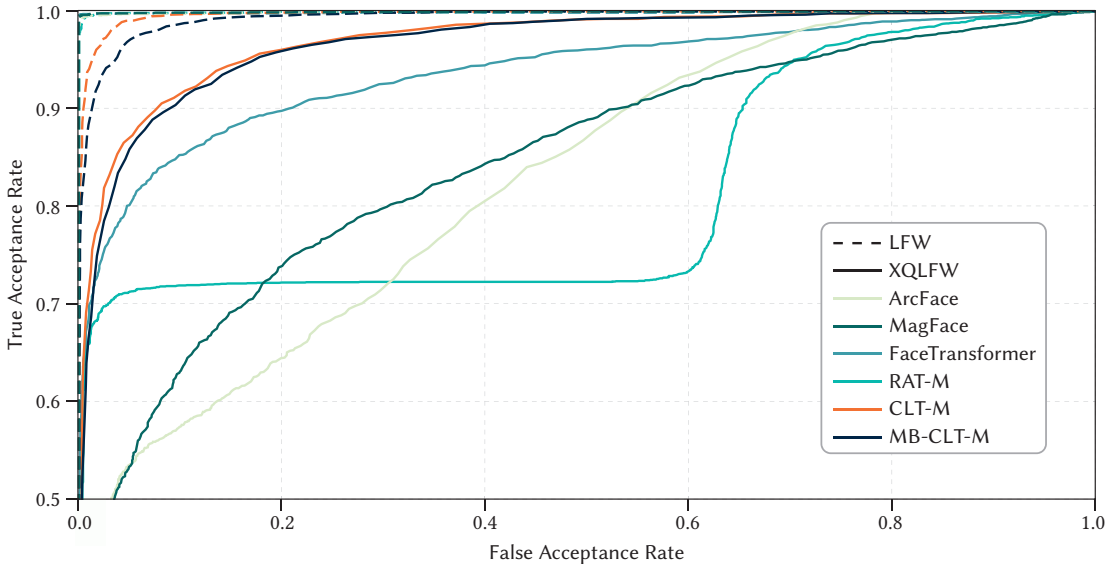


Figure 5.6.: Cross-resolution receiver operating characteristic curve comparison on the LFW dataset (dashed) and XQLFW dataset (solid) of various face verification approaches as in [3†].

compared to higher resolution image pairs (see Figure 4.9). This jump in feature distances for very LR genuine image pairs leads to a higher *False Acceptance Rate* (FAR) but almost no change in TAR, which results in the horizontal line behavior. When the thresholds reaches again the feature distances of the very LR imposter image pairs, the TAR starts to rise back again, while the FAR is still increasing.

In conclusion, the analysis of the ROC curve further helps to understand the behavior of the models in CR scenarios and underlines the importance of the proposed XQLFW evaluation protocol.

5.4. Conclusion

In summary, this chapter introduced the generation and evaluation of the XQLFW dataset published in [3†], which focuses on significant image quality variations and thus, evaluates FV systems on their robustness against image quality or resolution. By utilizing two independent image quality metrics, the quality scores of the images in the well-known LFW database are calculated. The synthetical degradation of rather low-quality images not only leads to a significant increase in the quality score variance, but also to a larger number of unique identities and images in the resulting evaluation protocol, in contrast to simply forming new image pairs with larger quality variations from the original LFW database. The XQLFW dataset still contains a competitively large number of identities and images, making it more suitable for a general evaluation. By maintaining the original LFW evaluation protocol, the XQLFW dataset is easy-integrable into existing frameworks and encourages the comparison of new methodologies with established ones.

The presented results showcased the susceptibility of several FV systems to image quality variations and hence the importance of the proposed XQLFW dataset.

In general, the creation of the XQLFW dataset, subjected to a singular application of more realistic synthetic downscaling addresses the issue of diverse downscaling kernels and methods, employed across numerous studies by establishing a common baseline. This approach markedly enhances the comparability of outcomes, providing a more uniform framework for evaluating the efficacy of different methodologies in the field of FV.

"Great things are not done by impulse, but by a series of small things brought together."

– Vincent van Gogh

Enhancing Explainability in Face Verification

This chapter ventures slightly beyond the principal focus of this dissertation, which centers on *Cross Resolution* (CR) *Face Verification* (FV), to explore *Explainable Artificial Intelligence* (XAI) within a broader context. The methodologies presented here, though originally proposed for FV in a broader sense, are also applicable to CR contexts and can be used to further understand the models behavior in CR scenarios. This chapter details two techniques for enhancing the transparency and interpretability of FV systems, as well as an interactive web platform, all of which have been published at the 3rd Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics (xAI4 Biometrics) at the Winter Conference on Applications of Computer Vision (WACV) in 2023 [4†]. First, the *Confidence Score* (C-Score) is introduced, quantifying the reliability of a model's prediction. Second, a visualization approach is established, which generates *Explanation Map*s (X-Maps), to obtain more meaningful predictions from an FV system. After discussing the most relevant related work in the field of explainable FV, the methodology of both techniques is presented. Then, the experimental results are shown, and the interactive web platform is presented. Finally, the chapter concludes with a short discussion.

6.1. Related Work

Research on interpretable and explainable *Face Recognition* (FR) can be categorized into two groups: 1) Explanation maps, which include the visualization of the influence of different areas of an image on the prediction of an FV model. 2) Confidence scores that aim to quantify the reliability of an FV model's prediction. In the following two subsections, an overview of the most relevant related work is given.

6.1.1. Explanation Maps

One of the earliest approaches for explaining the decisions in a classification problem is the *Local Interpretable Model-Agnostic Explanations* (LIME) technique introduced by Ribeiro *et al.* [188]. In their work, they proposed a method for faithfully explaining any classifier's predictions by learning an interpretable model locally around the prediction.

One can generally distinguish between model-agnostic and model-specific XAI techniques:

Model-agnostic techniques do not require any knowledge of the model's internal structure and can be applied to any model. Mery and Morris [189] introduced six different heat maps that can be used to explain any FV algorithm without manipulating the model. The key idea of their method is to define a matching score of two facial images, which changes when one image is perturbed. In addition, they experimented with XAI saliency maps based on contours. In [190], Mery introduced an XAI method based on how the probability of recognition of a given image alters when it is perturbed. His algorithm removes and aggregates different parts of the image and then measures the contributions of those parts individually and in-collaboration as well. The generated saliency maps highlight the most relevant areas for the recognition process. The work from Lin *et al.* [191] provided a learnable module that can be integrated into most FV models. The module generates meaningful explanations with the help of a patched cosine and an attention map. These maps represent similarities instead of saliency.

Model-specific techniques require knowledge of the model's internal structure to observe or manipulate the outputs of hidden model layers. Early approaches include the work of Cao *et al.* [192] who modified networks with a feedback loop to infer the activations of hidden layers according to the corresponding targets. The most popular approach is the *Gradient-weighted Class Activation Mapping* (Grad-CAM) algorithm [193] that utilizes the gradient of the class signal with respect to the input image. Recently, many other XAI techniques based on Grad-CAM, like Grad-CAM++ [194], HiRes-CAM [195], Ablation-CAM [196], Score-CAM [197], or XGrad-CAM [198], have been introduced. In [199] and [200], the authors trained separate models to predict saliency explanation maps. Pruning a neural network for a given single input to keep only neurons that highly contribute to the prediction was introduced in the work of Khakzar *et al.* [201].

6.1.2. Confidence Scores

The first work that included uncertainty information in the FR process was done by Shi and Jain [202]. They proposed *Probabilistic Face Embedding*s (PFEs), which estimate the uncertainty information of a single face image by a stage-wise learning process on top of a pre-trained FR model to incorporate the uncertainty of the facial features into the embedding. Chang *et al.* [203] extended the idea of PFEs and proposed a method to learn the feature and the uncertainty of the facial features simultaneously. Another approach, described by Schlett *et al.* [204], estimates the quality of face images for FR purposes, termed *Face Image Quality* (FIQ). They also presents an exhaustive review of works on FIQ in computer vision. However, these works do not directly provide a confidence score for the decision of the FV system and rather use the confidence or quality estimation of the input images to improve the general performance of a model.

Huber *et al.* [205] introduced a more closely related approach. They exploited the approximation of model uncertainty through dropout and proposed an uncertainty score for the comparison of two images. The score is utilized to additionally calculate a decision confidence to make the decisions for FV more transparent without any training effort.

A similar approach is presented by Neto *et al.* [206]. Their Probabilistic Interpretable Comparison (PIC)-Score is derived by learning probability density functions of the distances between genuine and imposter pairs using kernel density estimation.

6.2. Methodology

This chapter introduces two techniques from [4†] to make FV systems more transparent and interpretable: 1) The C-Score is a confidence score that quantifies the reliability of a model's prediction. 2) The X-Maps are visualizations of the influence of different areas of an image on the prediction of an FV model. The following two sections describe the methodologies more in detail.

6.2.1. Confidence Score

Typical FV systems use feature extractors and make their predictions based on the distance between two feature vectors (see also Section 2.3.2.4). Those feature vectors are derived from an FR $f_\theta(\cdot)$, which extracts facial features $f_\theta(\boldsymbol{I}) = \boldsymbol{f}$ from an aligned facial image $\boldsymbol{I} \in \mathbb{R}^{112\times112\times3}$. Many approaches [1†, 2†, 47, 56, 57, 89, 91, 105, 107] utilize the cosine distance metric d_{COS} (see Equation (2.11)) during testing for calculating the distance between two facial feature vectors $\boldsymbol{f}_1, \boldsymbol{f}_2$ from given images $\boldsymbol{I}_1, \boldsymbol{I}_2$. The cosine distance is derived by calculating the cosine similarity, measuring the angle between two features in a high dimensional space (see also Section 2.2.2.1). A cosine distance of 0 indicates that the two feature vectors are identical in terms of direction. A distance of 1 indicates that the two feature vectors are orthogonal and a distance of 2 indicates that the two feature vectors are opposite. The decision whether two given features belong to the same identity is typically driven by setting a threshold t for the cosine distance. For common FV benchmark datasets (*e.g.*, *Labeled Faces in the Wild* (LFW) [31], *Cross-Age Labeled Faces in the Wild* (CALFW) [94], *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92], *Similar-Looking Labeled Faces in the Wild* (SLLFW) [97], *Cross-Quality Labeled Faces in the Wild* (XQLFW) [3†]), the threshold t is derived by applying 10-fold cross-validation on the test set. A certain threshold is found for each fold by maximizing the verification accuracy on the remaining folds. A prediction from an FV model with a distance d close to the threshold t can be interpreted as uncertain. In contrast, a large distance close to 2 or a small distance close to 0 indicates high confidence in the model's prediction. However, there is no clear rule on interpreting the absolute distance to the threshold t in terms of prediction confidence. For instance the feature distance can be in various ranges for different models, as can be seen in the feature analysis of Section 4.2.

Therefore, this chapter introduces the Confidence Score (C-Score), a more expressive metric, which takes not only the imbalance of the feature distance distribution into account, but also exploits information from the distribution of correct and wrong predictions of the model for each dataset. The calculation of the C-Score is described step-by-step in the following:

Step-1 Histogram: Given the cosine distance distribution derived from an arbitrary FV model for genuine and imposter image pairs, a histogram with 400 bins is constructed for both, the genuine and imposter pairs. The histogram illustrates the cosine distances for the first fold of the LFW [31] dataset on the left of Figure 6.1.

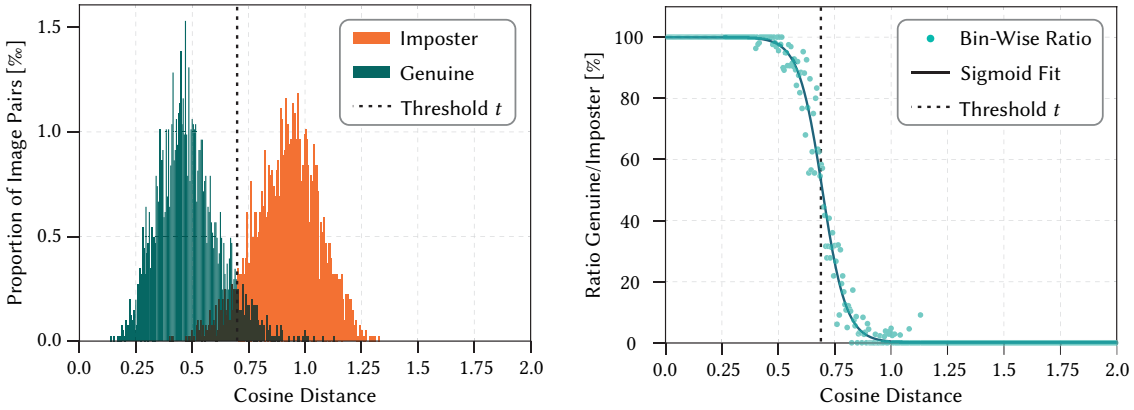


Figure 6.1.: Histogram of cosine distances for the first fold of the LFW dataset and the bin-wise ratio between genuine and imposter distance counts. The distances are derived from an ArcFace model fine-tuned with OLT [2†].

Step-2: Bin-wise Ratios: For each bin, the number of genuine pairs is divided by the total number of distances within that bin. If a bin contains no pair distances, to prevent division by zero, the ratio is set to 1 for bins less than the threshold t, and 0 for bins greater than t. This approach allows for the calculation of the ratio between genuine and imposter pairs for each distance bin. A ratio of 1 indicates that only distances from genuine pairs are present in the bin, suggesting a high confidence level in predictions of genuineness as no low-distance imposter pairs are observed. Conversely, a ratio of 0 implies only distances from imposter pairs in the bin, indicating a high confidence level in predictions of imposters, due to the absence of high-distance genuine pairs. As the distance bins approach the threshold t, there is an increased likelihood of encountering genuine pairs that are distantly placed and erroneously classified as imposters. Similarly, closer imposter pairs might be incorrectly classified as genuine. In the right chart of Figure 6.1, the LFW dataset displays these bin-wise ratios: The ratio for genuine pairs begins at 1 and declines towards the threshold t, while the ratio for imposter pairs starts at 0 and rises towards t.

Step-3: Sigmoid Fitting: Next, the calculated bin-wise ratios are utilized to fit a logistic sigmoid curve SIGMOID(d) to the distribution of ratio values. The sigmoid curve, dependent on the feature distance d is defined as:

$$\text{SIGMOID}(d) = \frac{v}{1 + e^{-k \cdot (d-e)}} + b \,, \tag{6.1}$$

with v as the curve's maximum value, k as the steepness of the curve, e as the x-value of the sigmoid's midpoint, and b as the curve's minimum value. These parameters of the sigmoid curve are derived using the dogbox [207] algorithm and characterize the distribution based on the data. In a final step, the fitted sigmoid curve is clipped to a range $[0, 1]$. The right chart of Figure 6.1 shows the fitted sigmoid curve for the LFW dataset.

Step-4: Inversion: Finally, to enhance the interpretability of the score, values beyond the threshold t are modified by applying an inversion formula:

$$C(d,t) = \begin{cases} \text{SIGMOID}(d) & \forall \quad d \leq t \\ 1 - \text{SIGMOID}(d) & \forall \quad d > t\,. \end{cases} \tag{6.2}$$

As a result, the final C-Score C is introduced, where the values are located in the range of $[0.5, 1]$ for either genuine or imposter predictions. These values represent an intuitive measurement of confidence for correctness. A confidence of 0.5 indicates that the model is uncertain about the prediction and can be treated as random guessing. On the other hand, a confidence of 1 indicates that the model is very confident about the prediction. With C, an additional value is established to the binary output prediction of an FV system and thus makes the prediction more meaningful. However, it is important to note that for calculating the C, the ground-truth information of a dataset is required. Hence, for the application of the C-Score to field data, the parameters for the sigmoid function need to be derived from other data, *e.g.*, a validation dataset.

6.2.2. Model-Agnostic Explanation Maps

The fundamental principle of this method is to visualize the extent of change in feature distance between the two input facial images when at least one of the images is modified with occlusions. This variation in distance is subsequently utilized to assign weights to the mask that was employed for the image's occlusion. By adopting a systematic occluding, *i.e.*, methodically applying occlusions across all parts of the image, it becomes possible to illustrate the distance deviation across all regions of the

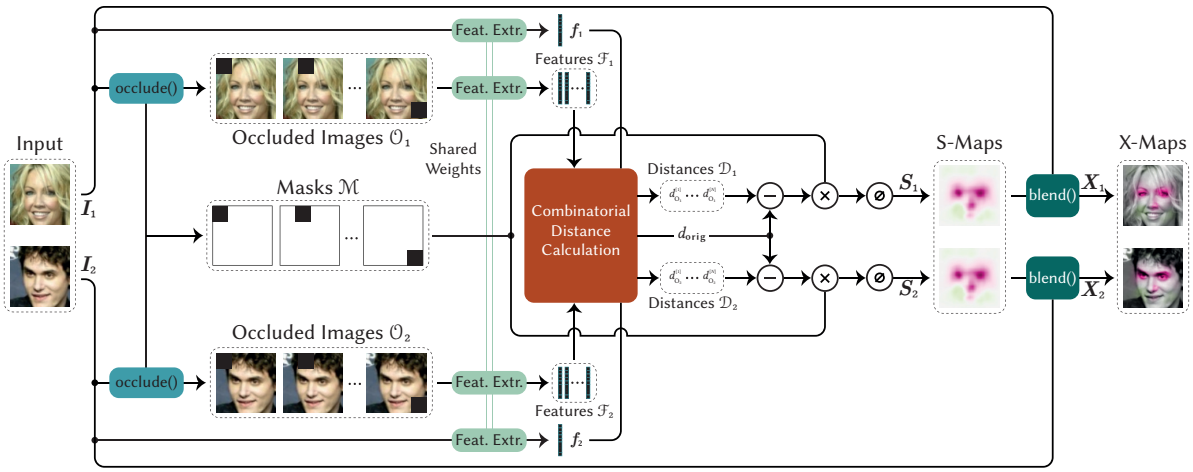


Figure 6.2.: Flowchart of generating two independent X-Maps as proposed in [4†] for both images of a given example input image pair from LFW database. After systematic image occlusion and combinatorial distance calculation, the masks are weighted by the distances and averaged to form *Similarity Maps* (S-Maps) and then blend to the final X-Maps

entire image through these weighted masks. Additionally, various patch sizes are used to ensure that the distance deviation is visualized not just for smaller areas but also for larger sections of the image.

The overall approach presented in Figure 6.2 takes in an 2-tuple of facial images $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ as input. Both images are opposed to be verified by an FV model f_θ generating a 2-tuple of X-Maps $(\boldsymbol{X}_1^{[\mathrm{m}]}, \boldsymbol{X}_2^{[\mathrm{m}]})$ given the input images, where $m \in \{A, B, C\}$ denotes the applied method for the combinatorial feature distance calculation. The X-Maps are then used as an explanation for the prediction of the FV model.

In general, the method can be divided into five steps: 1) Systematic Image Occluding, 2) Inference, 3) Combinatorial Feature Distance Calculation, 4) Generation of the S-Map, and 5) Blending to an X-Map.

Step-1: Systematic Image Occluding: First, each input image $\boldsymbol{I} \in \mathbb{R}^{112\times112\times3}$ is systematically occluded with different patch sizes γ and strides β which generate a set of occluded images $\mathcal{O} := \left\{\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_N\right\} | \boldsymbol{O} \in \mathbb{R}^{112\times112\times3}$ and a set of masks $\mathcal{M} := \left\{\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_N\right\}$, with $\boldsymbol{M} \in \mathbb{R}^{112\times112}$. This is achieved with Algorithm 1 as:

$$\texttt{occlude}(\boldsymbol{I}): \quad \boldsymbol{I} \mapsto \mathcal{O}, \mathcal{M}. \tag{6.3}$$

The generated masks $\boldsymbol{M}$ are sparsely populated; only the occluded areas contain the values 1. The generated occluded images $\boldsymbol{O}$ contain the original pixel information from the input image, except for the occluded areas, which are simply filled with a constant value of 0, if not stated otherwise. Applying this process to the 2-tuple of input images $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ results in a 2-tuple of occluded images $(\mathcal{O}_1, \mathcal{O}_2)$ and masks $(\mathcal{M}_1, \mathcal{M}_2)$, respectively. Note that the generated masks are actually the same for both input images as the occlusions are applied in the same manner. The number W of generated occluded images $\boldsymbol{O}$ and masks $\boldsymbol{M}$ by Algorithm 1 is dependent on the patch size γ and the stride β, and can be calculated as:

$$W = \left\lfloor \left(\frac{112 - \gamma}{\beta} \right) \right\rfloor^2. \tag{6.4}$$

Note, that if not stated otherwise, a stride of $\gamma = 5$ is applied in the experiments and three different patch sizes are used, $\gamma \in \mathcal{P} := \{7, 14, 28\}$. The following three steps refer to a specific patch size γ.

Algorithm 1: Systematic Image Occluding $\texttt{occlude}(\cdot)$

Input: image $\boldsymbol{I}$
$\beta \longleftarrow$ stride
$\gamma \longleftarrow$ size of patch
Start at top left corner of $\boldsymbol{I}$
while *within* $\boldsymbol{I}$ **do** move right β pixels
  while *within* $\boldsymbol{I}$ **do** move down β pixels
    $\boldsymbol{M} \longleftarrow$ draw a patch with size γ at that location
    $\boldsymbol{O} \longleftarrow$ occlude $\boldsymbol{I}$ with patch of size γ at that location
  end
end
Output: occluded images $\mathcal{O}$, masks $\mathcal{M}$

Step-2 Feature Extraction: To explain the decisions of an FV model f_θ, this model is also used to extract features $\boldsymbol{f} \in \mathbb{R}^{512}$ in the inference step of the proposed approach. Inferring the original image $\boldsymbol{I}$ and occluded set of images $\mathcal{O}$ can written as:

$$\begin{aligned} \boldsymbol{f} &= f_\theta(\boldsymbol{I}), \\ \mathcal{F} &:= \left\{ f_\theta\left(\mathcal{O}^{(i)}\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}, \end{aligned} \tag{6.5}$$

with $\mathcal{O}^{(i)}$ denoting the i-th occluded image in $\mathcal{O}$. Applying inference to the 2-tuple of input images $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ and their corresponding 2-tuple of occluded images $(\mathcal{O}_1, \mathcal{O}_2)$ consequently results in a 2-tuple of feature vectors $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ and a 2-tuple of feature vector sets $(\mathcal{F}_1, \mathcal{F}_2)$, respectively.

Step-3: Combinatorial Feature Distance Calculation: The cosine distance $d_{\cos}$ between the feature vectors of the original images $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ is calculated and acts as a reference feature distance d_{orig}. Then, the cosine distances $\mathcal{D}_1, \mathcal{D}_2$ are calculated based on the feature vectors of the original images $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ and the set of feature vectors of the occluded images $\left(\mathcal{F}_1, \mathcal{F}_2\right)$. The combinatorial feature distance calculation is applied using three different rules A, B, and C, which are described in the following:

Method-A selects the cosine distances for $\mathcal{D}_1$ and $\mathcal{D}_2$ such that the average distance from each individual occluded image feature $\mathcal{F}_1^{(i)}$ corresponding to $\boldsymbol{I}_1$, to all occluded images features $\mathcal{F}_2$ corresponding from $\boldsymbol{I}_2$, and vice-versa is calculated. The selection can be formulated mathematically as:

$$\begin{aligned} \mathcal{D}_1 &:= \left\{ \frac{1}{W} \sum_{j=1}^{W} d_{\cos}\left(\mathcal{F}_1^{(i)}, \mathcal{F}_2^{(j)}\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}, \\ \mathcal{D}_2 &:= \left\{ \frac{1}{W} \sum_{j=1}^{W} d_{\cos}\left(\mathcal{F}_2^{(i)}, \mathcal{F}_1^{(j)}\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}. \end{aligned} \tag{6.6}$$

Method-B selects the cosine distances for $\mathcal{D}_1$ and $\mathcal{D}_2$ such that the cosine distance between the feature vector $\boldsymbol{f}_1$ of the original image $\boldsymbol{I}_1$ and the occluded versions feature vectors $\mathcal{F}_2$ of the other original image and vice-versa is calculated. By occluding only one of the input images, this method measures the effect on the feature distance for the occlusions specifically for the image which is then occluded. This distance sets are defined as:

$$\begin{aligned} \mathcal{D}_1 &:= \left\{ d_{\cos}\left(\mathcal{F}_1^{(i)}, \boldsymbol{f}_2\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}, \\ \mathcal{D}_2 &:= \left\{ d_{\cos}\left(\mathcal{F}_2^{(i)}, \boldsymbol{f}_1\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}. \end{aligned} \tag{6.7}$$

Method-C selects the cosine distances $\mathcal{D}_1$ and $\mathcal{D}_2$ according to the mathematical expression:

$$\mathcal{D}_1 = \mathcal{D}_2 := \left\{ d_{\cos}\left(\mathcal{F}_1^{(i)}, \mathcal{F}_2^{(i)}\right) \,:\, \forall\, i \in [1, 2, \ldots, W] \right\}. \tag{6.8}$$

In this method, the cosine distances between the feature vectors of the occluded images are calculated, with the constraint that the occlusions are at the same position in both input images. Hence, the similarity maps for Method-C are identical.

Algorithm 2: Color Blending `blend(·, ·)`

Input: image $\boldsymbol{I}$, S-Map $\boldsymbol{S}$
$l \longleftarrow$ get luminance from: RGBtoHLS($\boldsymbol{I}$)
$h \longleftarrow$ get hue from: RGBtoHLS($\boldsymbol{S}$)
$s \longleftarrow$ get saturation from: RGBtoHSV($\boldsymbol{S}$)
$\boldsymbol{B} \longleftarrow$ HLStoRGB(h, l, s)
Output: blended image $\boldsymbol{B}$

Step-4: Generation of the S-Map: To visualize this change in distance for every part of an image, the deviation of the given distance set $\mathcal{D}$ is compared with the original distance d_{orig} of both non-occluded input images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. Therefore, the set of masks $\mathcal{M}$, which is related to the distances in $\mathcal{D}$, is weighted by the difference in distance compared with d_{orig}. These weighted masks are then averaged to visualize the deviation caused by the occlusion on the entire image. The process can be formulated as:

$$\boldsymbol{S}_\gamma = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{D}^{(i)} - d_{\text{orig}}) \cdot \mathcal{M}^{(i)} \,. \tag{6.9}$$

Given that occlusions are applied using a patch size of γ, the resulting S-Map for the input image $\boldsymbol{I}$ is denoted as $\boldsymbol{S}_\gamma$. Note that this procedure is performed separately for every particular occlusion patch size $\gamma \in \mathcal{P}$ and the final S-Maps are averaged based on the size of the patch area according to:

$$\boldsymbol{S} = \sum_{i=1}^{|\mathcal{P}|} \frac{\boldsymbol{S}_\gamma}{\gamma_i^2 \cdot |\mathcal{P}|} \,. \tag{6.10}$$

Performing this procedure on the 2-tuple of input images $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ and their corresponding 2-tuple of distance sets $(\mathcal{D}_1, \mathcal{D}_2)$ results in a 2-tuple of S-Maps $\left(\boldsymbol{S}_1^{[\text{m}]}, \boldsymbol{S}_2^{[\text{m}]}\right)$ for the input images, where $m \in \{A, B, C\}$ denotes the applied method for the combinatorial feature distance calculation.

If the feature distance between those two images is decreasing, the occluded area is interpreted as dissimilar and vice-versa similar for a greater distance. A Gaussian-Blur algorithm is applied to highlight the resulting effect and compensate the raster artifacts, caused by using a stride β instead of shifting the occlusion patch pixel by pixel. The algorithm is parametrized with an $\beta \times \beta$ kernel and $\sigma = \beta$ is applied to each resulting S-Map $\boldsymbol{S}$ followed by normalization to the range $[-1, 1]$. Furthermore, the values of each S-Map are mapped to a (red-to-white-to-green) color map to further enhance the visualization.

Step-5: Generation of the final X-Map: Finally, the X-Map $\boldsymbol{X}$ is generated by color blending the original image $\boldsymbol{I}$ with the S-Map $\boldsymbol{S}$ using the algorithm Algorithm 2. The algorithm utilizes a color-space transformation and separation to combine the original image and with the S-Map in a meaningful way. The procedure is done for both S-Maps $\boldsymbol{S}_1^{[\text{m}]}$ and $\boldsymbol{S}_2^{[\text{m}]}$ and results in the final X-Maps $\boldsymbol{X}_1^{[\text{m}]}$ and $\boldsymbol{X}_2^{[\text{m}]}$ for the input images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, respectively. Note that $m \in \{\text{A}, \text{B}, \text{C}\}$ is denoting the applied method for the combinatorial feature distance calculation.

The proposed approach is able to generate an image-specific X-Map for both input images. The X-Map is a visualization of the influence of different parts of an image on the prediction of a model. It is important to note that the X-Map is generated without any knowledge of the model's internal structure and can be applied to any FV model, assuming the extracted facial feature can be accessed.

6.3. Results

The result section can be divided into two categories: 1) Quantitative results evaluate the C-Scores generated for several datasets. 2) Qualitative results visually evaluate the effectiveness of the proposed X-Maps method.

6.3.1. Quantitative Results

The proposed C-Score is calculated for various datasets utilizing two FV models (FaceTransformer [91] and ArcFace [47]) fine-tuned with the *Octuplet Loss Training* (OLT) method (see Section 4.3). Table 6.1 presents the C-Score and accuracy for several datasets and highlights the differences between genuine and imposter pairs for both metrics. Overall, it can be observed that deviations for LFW are minimal in comparison to CPLFW, where deviations are significant. This effect could be attributed to the complexity of the image pairs being classified where deviations are less significant, with the relatively simpler pairs from LFW compared to CPLFW.

It is noteworthy that the accuracy for pairs classified as imposters is consistently higher across all datasets than that for genuine pairs. This suggests that FV models more effectively identify imposter pairs than genuine pairs, indicating a lower error rate in the classification of imposter pairs. Consequently, the decision threshold has been adjusted to favor the correct identification of imposter pairs, albeit at the expense of reducing accuracy for genuine pairs. Conversely, the C-Score presents a different scenario except for the LFW benchmark: Confidence levels for pairs classified as genuine are higher than those for pairs classified as imposters, as setting a lower threshold leads to more accurate classification of imposter pairs despite a lower confidence rating. However, such thresholds

Table 6.1.: Mean accuracy and C-Score for all pairs of various datasets, alongside columns detailing the deviation between all pairs and only genuine ones, as well as the deviation for imposter pairs. Two models which are trained on the MS1M-V2 database [47, 62] are compared.

	FaceTransformer + OLT [$2^{\dagger}$]						ArcFace + OLT [$2^{\dagger}$]					
	C-Score [%]			Accuracy [%]			C-Score [%]			Accuracy [%]		
Dataset	Genuine	All	Imposter	Genuine	All	Imposter	Genuine	All	Imposter	Genuine	All	Imposter
LFW	-0.32	99.58	+0.32	-0.13	99.73	+0.13	-0.23	99.20	+0.23	-0.18	99.55	+0.18
CALFW	+0.64	95.72	-0.64	-2.53	94.93	+2.53	+0.61	94.92	-0.61	-2.88	93.85	+2.88
CPLFW	+1.49	92.82	-1.49	-4.72	91.58	+4.72	+1.56	89.87	-1.56	-7.10	88.37	+7.10
SLLFW	+0.45	98.52	-0.45	-2.62	96.78	+2.62	+1.64	95.51	-1.64	-2.73	94.90	+2.73
XQLFW	+0.34	95.19	-0.34	-0.55	95.12	+0.55	+0.96	93.32	-0.96	-1.33	93.27	+1.33

also result in a higher rate of genuine pairs being mistakenly classified as imposter. An exception is achieved for the LFW due to the fitting of the sigmoid function, which merely approximates the data without perfect replication.

6.3.2. Qualitative Results

To qualitatively evaluate the efficacy of the generated X-Maps, this section first presents a selection of example pairs from the LFW [31] dataset. Then, the different techniques, applied for the combinatorial feature distance calculation, are compared. Followed by experiments with meticulous splicing of facial features from one image and their subsequent integration into another image. Finally, the sensitivity of the patch size, shape, fill, and edge quality is analyzed.

6.3.2.1. Visual Evaluation

To visually evaluate the effectiveness of the proposed X-Map method, a selection of images pairs from the LFW [31] dataset is shown in Figure 6.3. The X-Maps are generated utilizing a FaceTransformer [91] model fine-tuned with OLT [2†] (see also Chapter 4). The X-Maps are generated using the Method-C (see also Equation (6.8)) for the combinatorial feature distance calculation. Three genuine and three imposter example pairs are shown. As depicted in Figure 6.3, the X-Maps are able to highlight the facial regions that are most influential for the model's decision. The green-colored facial regions indicate similarity, while the red-colored facial regions indicate dissimilarity. In all images the deviation is only pronounced in the facial area, as the background shall not influence the model's decision.

The X-Maps of the genuine pairs are dominated areas highlighted as similar. The X-Map of the left example pair reveals that the eyes and mouth of the person seem to not play an essential role in the model's decision. The cosine distance will get even slightly smaller for occlusions on those parts of the face. Similarly, in the center pair, the nose also seems to be less critical for the model's

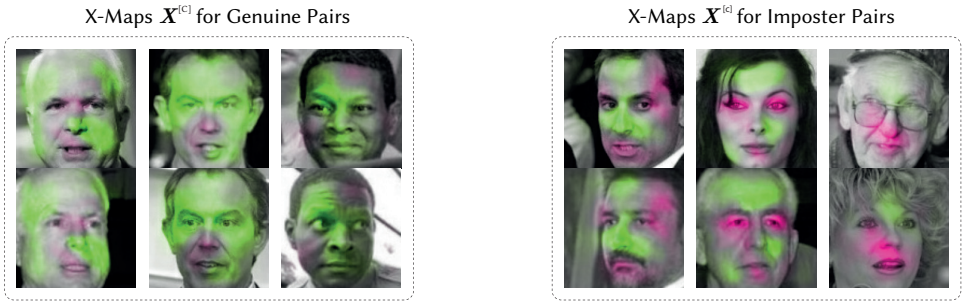


Figure 6.3.: X-Maps of Method-C for three genuine and three imposter example pairs from the LFW dataset. Green colors indicate similar facial regions and red highlights dissimilar regions. All X-Maps are generated utilizing the FaceTransformer model fine-tuned with OLT on the MS1M-V2 database. Adopted from [4†].

prediction. In contrast, the X-Map of the right example pair indicates that the eyes, nose, and mouth are highlighted, *i.e.*, those regions push the distance closer to 0.

The X-Maps of the imposter pairs indicate more dissimilar facial regions than the genuine pairs' X-Maps. In all three pairs, the forehead region is highlighted rather as similar compared to the more distinctive facial parts such as the eyes, nose, and mouth. The most dissimilar facial region in the imposter pairs seems to be the eyes of the center example pair. Furthermore, the nose region of the right example pair is also highlighted as dissimilar. In contrast, the nose of the left example pair is specified as a very similar facial region. Interestingly, the glasses of one subject in the right example does not seem to affect the models decision, as these area is not highlighted in the X-Map at all.

However, the interpretation of the X-Maps is not always obvious and should be seen as further contribution towards a more comprehensive understanding of the model's decision or makes it challenging to interpret them in a proper way.

6.3.2.2. Method Analsysis

As mentioned in Section 6.2.2, Method-A and Method-B focus on image specific deviations of cosine distances between the features of the image pair. In contrast Method-C solely considers co-located occlusions in both images, resulting in identical X-Maps for both images. The aforementioned method is thus best suited for image pairs where the facial features are located at the same positions. While the alignment process (see Section 2.3.2.2) ensures that faces are aligned — projecting landmarks of the eyes, nose, and mouth corners to a normalized position — this is not always achievable with significant changes in pose, as observed in real world scenarios or in datasets like CPLFW and *Celebrities in Frontal-Profile — Frontal-Profile* (CFP-FP). Therefore, it is advisable to consider Method-A or Method-B in such instances.

Figure 6.4 illustrates the different effects of applying Method-A, Method-B, and Method-C for the combinatorial feature distance calculation in the X-Maps generation process for a genuine and imposter sample image pair. Additionally to the observations in Figure 6.3, the Method-A and Method-B also highlight only the facial regions, hence the background is not influencing the model's decision. The fact that the X-Maps are unique for each image complicates their interpretation, as the previous interpretation of distinguishing between similar and dissimilar areas no longer applies, at least not in terms of co-located similarity. It's more accurate to assume that certain parts of one image are similar in terms of it's identity compared to different areas of another image. For instance, in the imposter pair using Method-B, only the left eye in the top image is markedly reddish, whereas in the bottom image, it's the eyes and the nose. This could suggest that the left eye more likely belongs to a different identity than the right eye. Conversely, covering the nose in the top image increases the distance, indicating that the nose could originate from the person in the lower image, yet it is not similar to the nose in the lower image.

In the genuine case, the right cheek appears to play a minimal role in the decision-making process during the co-located comparison (Method-C), unlike in Method-A and Method-B, where it is colored green. This result suggests that the cheek does not look very similar in the two images compared. However, if it is covered in just one of the images, it plays a significant role in the model's decision.

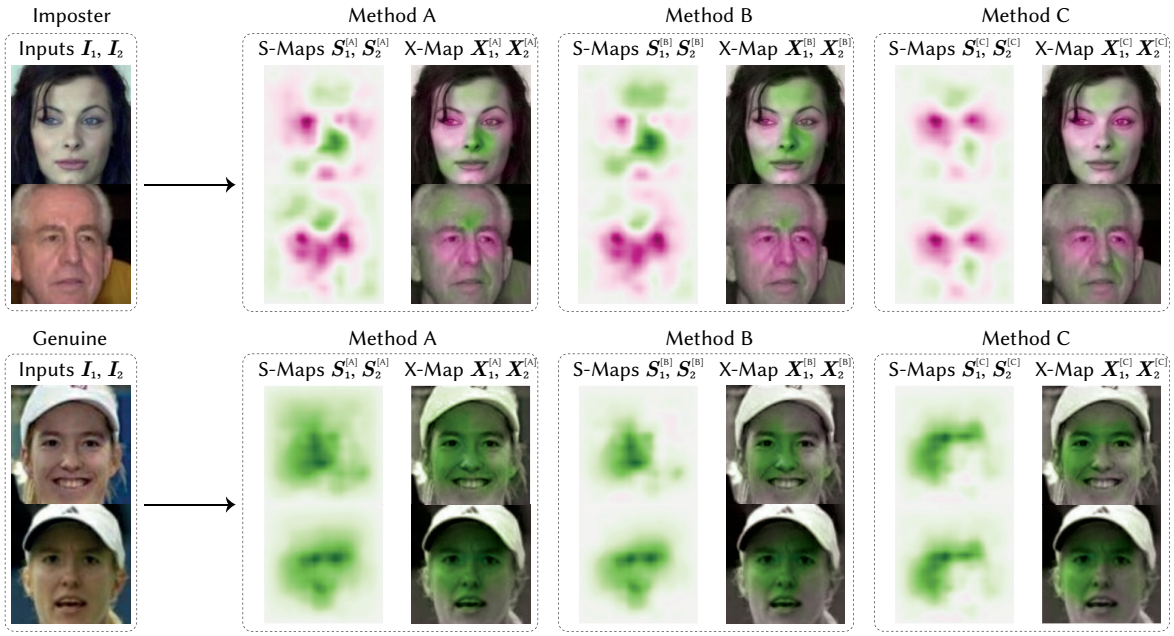


Figure 6.4.: Comparison of Method-A, Method-B, and Method-C for the combinatorial feature distance calculation in the X-Maps generation process. One genuine and one imposter sample pair is taken from LFW. All X-Maps are generated utilizing a FaceTransformer model, fine-tuned with OLT. Adapted from [4†].

These examples illustrate the challenge to derive a clear interpretation and diminish the conclusiveness of Method-A and Method-B. Nevertheless, they provide a form of explanation and offer another way to interpret the model's decision.

6.3.2.3. Facial Feature Splicing Test

To validate the efficacy of the explanation generation method — specifically, its capability to accurately classify regions of similarity and dissimilarity within images — another experimental approach was adopted. It involved the meticulous splicing of facial features from one image and their subsequent integration into another image. This process was designed to rigorously test the method's capability in distinguishing between similar and dissimilar facial regions. In Figure 6.5 three examples of facial replacements are depicted. In the left example, the eye region of the lower image is replaced by the eye region of the upper image. The with Method-C generated X-Map highlights the replaced eye region as very similar and the remaining facial area as dissimilar, which understates the interpretation of the X-Map as a kind of similarities and dissimilarities. In the center example, the half side of the face of the lower image is replaced by the half side of the face of the upper image. The X-Map generated by the proposed method highlights the replaced half side of the face as similar and the remaining facial area as dissimilar. In the right example, the eye and mouth of the lower image are replaced by

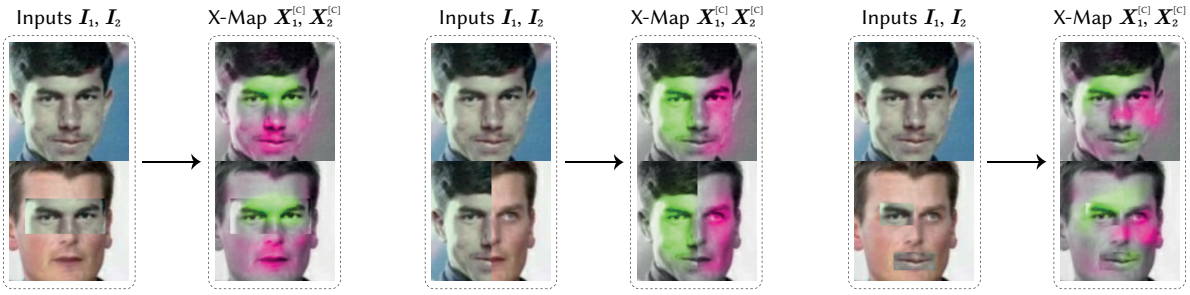


Figure 6.5.: Three example pairs from LFW containing facial replacements and the corresponding X-Map generated by a FaceTransformer model fine-tuned on MS1M-V2, with the proposed process. Adapted from [4†].

the eye and mouth of the upper image. The generated X-Map reveals the strong similarity of the left eye and mouth in contrast to the remaining facial parts, which are categorized as dissimilar. However, this effect is weakly pronounced in the right example.

6.3.2.4. Sensitivity Studies

To get a better understanding of the influence of the shape used for the systematic image occluding, the foundation of the X-Map generation process, multiple sensitivity studies are conducted. The shape, size, edge quality, and coloring of the patches used in the systematic image occlusion algorithm Algorithm 1 are varied and the S-Map are generated with Method-C for the combinatorial feature distance calculation. The results of these sensitivity studies are depicted in Figure 6.6. As a reference, the S-Maps with the default settings, *i.e.*, all patches are black colored, square shaped and the edges are sharp, are shown in the top center of Figure 6.6.

First, three different patch sizes 7×7 px, 14×14 px, and 28×28 px are used for the systematic image occlusion algorithm and shown in the bottom left of Figure 6.6. The smallest patch generates a the most fine-grained S-Map and highlights small areas differently compared to using the largest patch. The behavior is linear between the patch sizes. In order to obtain one generalized S-Map with information from different levels of granularity, the maps for all three patch sizes are merged and weighted based on the area of the patches (see Equation (6.10)).

Second, the coloring of the patches is varied and displayed in Figure 6.6 at the bottom center. In contrast to the reference X-Map, the gray, white, and noisy occlusions generate X-Maps with higher intensities. However, the difference between the coloring is only marginal.

Third, in the reference method, the occlusions are sharp, *i.e.*, the masks only contain the values 0 and 1. To analyze the effect of the edge quality, the occlusions are Gaussian blurred with different kernel sizes and sigmas ($\{7, 14, 56\}$). Similar to the patch size itself, this heavily affects the S-Maps regarding visual granularity observable in bottom right of Figure 6.6. The larger the kernel size and sigma, the more blurred the S-Map becomes. Interestingly, using a large kernel size and sigma, the dissimilar areas are getting more prominent in the S-Maps. The gradual transition from white to black

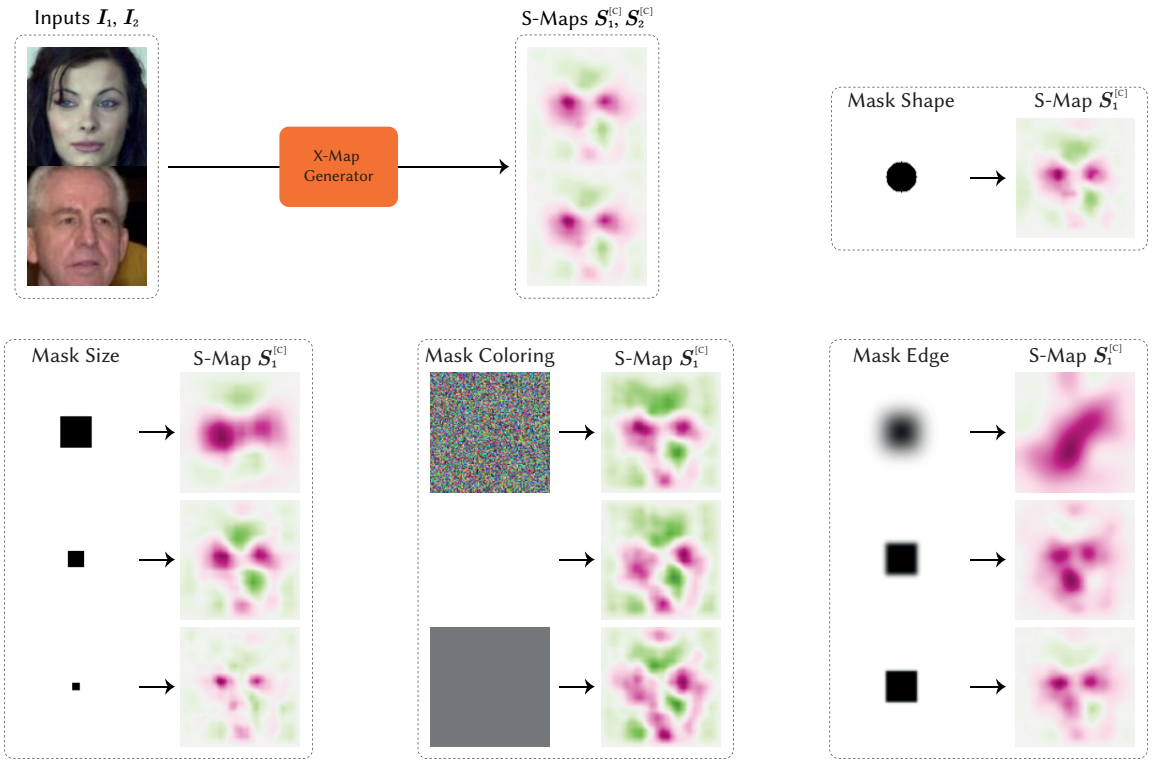


Figure 6.6.: Sensitivity studies on different patch size, edge quality, coloring, and shape of the occluded areas. A sample image pair from LFW is taken as the sample and reference S-Map. All S-Maps are generated utilizing a FaceTransformer model fine-tuned with OLT on MS1M-V2. Adapted from [4†].

inside the occlusion masks is likely causing certain facial features to be altered in such a way that they are no longer correctly recognized, leading to a change of the identity. Since there are naturally no sharp transitions from one facial region to another in a human face, this interpretation is quite reasonable.

Lastly, the S-Map generation process is done with circle-shaped patches instead of square shape. As obvious in top right of Figure 6.6 the shape does not have a significant influence on the resulting S-Map.

In conclusion, these sensitivity studies reveal, that the S-Map content depends controversy on the patch characteristics. Whereas the patch size and edge quality have the most significant impact, the coloring and shape have only a marginal effect. Therefore, some patch characteristics should be adjusted carefully to the purpose of the X-Map.

6.4. Web Platform

This section briefly explains the *eXplainable Face Verification* platform[i], developed besides the X-Map and C-Score algorithms. The platform is designed to present the qualitative results of the approach and to help the community familiarizing with several test datasets and different model behaviors. Therefore, X-Maps and C-Scores are calculated for two FV models and several popular datasets. The datasets include LFW [31], CALFW [94], CPLFW [92], SLLFW [97], and XQLFW [3†]. The models are the FaceTransformer [91] and the ArcFace [47], each fine-tuned with the OLT [2†] method. The interactive platform has two main functionalities as illustrated in Figure 6.7:

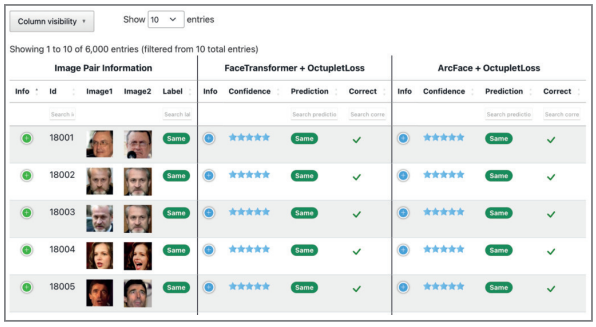
The "Explorer" page, which contains an interactive table where users can filter and sort the data containing the image pairs, the corresponding metadata, the C-Score values, cosine distance, decision threshold, label, and image quality score as proposed in [3†].

The "Viewer" page, which presents the generated X-Maps in an interactive, adjustable way for every image pair of the datasets. This includes the possibility to select different models, methods, and maps for each pair of images. Additionally, the image quality scores, labels, predictions and C-Scores are displayed for each pair. In total, the platform includes 30 k image pairs from five different datasets and the results of two different models.

The platform is hosted entirely on the Google Cloud Platform and runs a flask [208] framework inside an App Engine and the image data is stored in a Cloud Storage Bucket. To be able to interact with the data, a SQLite[ii] database stores the metadata of the image pairs. The backend does all the sorting, filtering, and accessing to improve the user experience.

[i] Available at `https://explainable-face-verification.ey.r.appspot.com/` (accessed on March 20, 2024).
[ii] SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. Source: `https://www.sqlite.org` (accessed on March 20, 2024).

Screenshot of the „Explorer"-Module



Screenshot of the „Viewer"-Module

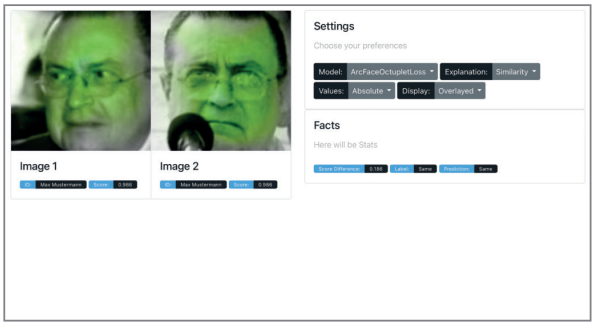


Figure 6.7.: Screenshots of the "Explorer" and "Viewer" module on the *eXplainable Face Verification* web platform. Sample images are taken from LFW database. Both screenshots were taken on March 20, 2024.

Although the platform is designed to be user-friendly, it is important to note that the platform is not optimized for mobile devices and the user experience may be limited on smaller screens. Moreover, the platform contains insights for only two FV models, thus make it difficult to generalize the results.

6.5. Conclusion

This chapter presented the in [4†] introduced methods to further explain the decision-making process of FV models. The C-Score is a confidence score for an FV model's prediction and C-Score is a measurement of the model's confidence in its prediction. It utilizes the distribution of other cosine distances produced from the model to get a level of confidence for a particular prediction. The proposed X-Map algorithm generates a visualization of the influence of different parts of an image on the prediction of a model, based on systematic occluding the input images and measure the change in cosine distance. The experimental results revealed that the interpretation of X-Maps is not as straightforward as it seems and is only able to highlight locally appearing similarities. Hence, more global identity features, *e.g.*, skin color or face shape, are not revealed. Both methods share a model-agnostic characteristic, *i.e.*, the C-Score and X-Map can be applied to any FV model (if the extracted facial features can be accessed). Moreover, a web platform is developed to present the qualitative results of the approach and to further enlighten the characteristics of various FV benchmarks and models. Although the focus of the proposed algorithm is explicitly on faces, the approach is not limited to the domain of faces and can potentially be applied to other binary decision problems.

“One never notices what has been done; one can only see what remains to be done.”

— Marie Curie

Human Performance and Fusion Strategies in Face Verification

This chapter slightly expands beyond the main focus of the dissertation, which is primarily on *Cross Resolution* (CR) *Face Verification* (FV), to examine the combination of machine and human operators in the FV task. The study presented here, including its methods and results, does not solely focus on the CR domain but also involves experiments with various other challenging datasets. This chapter primarily encompasses the contributions published in [5†] at the 18th International Conference on Machine Vision Applications (MVA) in 2023. The study investigates the effect of a combination of machine and human operators in the FV task. After a brief review of the related work, a look closer at the edge cases for several state-of-the-art models to discover common datasets' challenging settings is presented. Then, the study design is outlined, and a fusion strategy to combine human and machine prediction for the FV task is introduced. The result section thoroughly analyses the human tasks and finally demonstrates that combining machine and human decisions can further improve the performance of state-of-the-art FV systems on various benchmark datasets. The chapter concludes with a discussion of the results and limitations of the study.

7.1. Related Work

Human performance in FV or face matching has been extensively studied. However, only a few works investigated human FV accuracy on popular datasets commonly used to evaluate automatic *Face Recognition* (FR) systems, hence drawing a comparison is not straightforward here. In [110], the authors conducted a study using Amazon Mechanical Turk[i] workers and evaluated the accuracy on the *Labeled Faces in the Wild* (LFW) [31] database, achieving 99.2%. To the best of our knowledge, no studies have yet been conducted on measuring human FV accuracy for other significantly large benchmark datasets.

From a psychological perspective, the characteristic of the study design to measure FV ability of humans has attracted several researchers. Howard *et al.* [209] investigated the influence of prior face identity decisions on subsequent human judgments about face similarity, *e.g.*, users get feedback of their performance on the latest task. Their study revealed that prior identity decision labels alter

[i] A crowdsourcing platform from Amazon, where humans perform tasks that computers are currently unable to do efficiently.

volunteers' internal criteria for judging face pairs. Fysh and Bindemann [210] examined the impact of onscreen trial labels with consistent, inconsistent, or unresolved information on face-matching accuracy. Their experiments demonstrated that human face-matching decisions are influenced by onscreen identifications, leading to increased accuracy when information is correct and increased errors when information is misleading. In [211], the authors explored the potential improvement in accuracy resulting from using multiple-image arrays in unfamiliar face-matching tasks. The study found that simultaneous viewing of multiple images of the same person improved matching accuracy.

Other works have focused on developing standardized tests for human analysis in face-matching tasks. *E.g.*, [212] introduced the *Glasgow Face Matching Test* (GFMT), which presents two images (genuine or imposter) taken in the same pose, minutes apart, with high-quality cameras to the subjects. Despite these apparently optimal conditions, the authors state that this task is not trivially easy, and they have demonstrated that there is large inter-individual variation in FV performance. After more then a decade, White *et al.* [213] improved the GFMT and introduced the GFMT-2, an enhanced version of a widely used measure for unfamiliar face matching ability, featuring four key improvements. These improvements include increased variation in test items for more realistic face identification challenges, equal difficulty in short and long test versions for repeated testing, elimination of repeating face identities, and separate short versions for exceptionally high or low performers. In the same year, the *Oxford Face Matching Test* (OFMT) [214] was published. It is a test designed to identify individual differences in face processing abilities across the full spectrum, from prosopagnosia[ii] to super recognizers. By using FR algorithms to obtain unbiased face pair similarity ratings, the test addresses the challenge of establishing difficulty for atypical groups who may use different strategies for face processing. Across five studies, the OFMT demonstrated sensitivity to individual differences in various populations, with test-retest reliability comparable to the Cambridge Face Memory Test [215] and the GFMT.

More closely related to the use of deep *Convolutional Neural Network* (CNN) for FV, Abudarham *et al.* [216] investigated the representation of face identity in deep CNNs and whether these networks rely on the same facial features used by humans for FR. The findings reveal that deep CNNs optimized for face identification are tuned to the same facial features as humans, with sensitivity to these features and view-invariant face representation emerging at higher layers in deep CNNs optimized for FR. The results validate human perceptual models of FR, support the use of deep CNNs for testing predictions about human face and object recognition, and contribute to the interpretability of deep CNNs.

7.2. Preliminary Analysis of Edge Cases

To better understand the challenges faced by state-of-the-art FV models on various benchmarks, a preliminary analysis with the in Chapter 6 proposed *Confidence Score* (C-Score) method was conducted. Therefore, three predictions and corresponding C-Scores of ArcFace [47], FaceTransformer [91], both

[ii] Prosopagnosia: Also known as face blindness, is a cognitive disorder of face perception in which the ability to recognize familiar faces, including one's own face (self-recognition), is impaired, while other aspects of visual processing (*e.g.*, object discrimination) and intellectual functioning (*e.g.*, decision-making) remain intact.

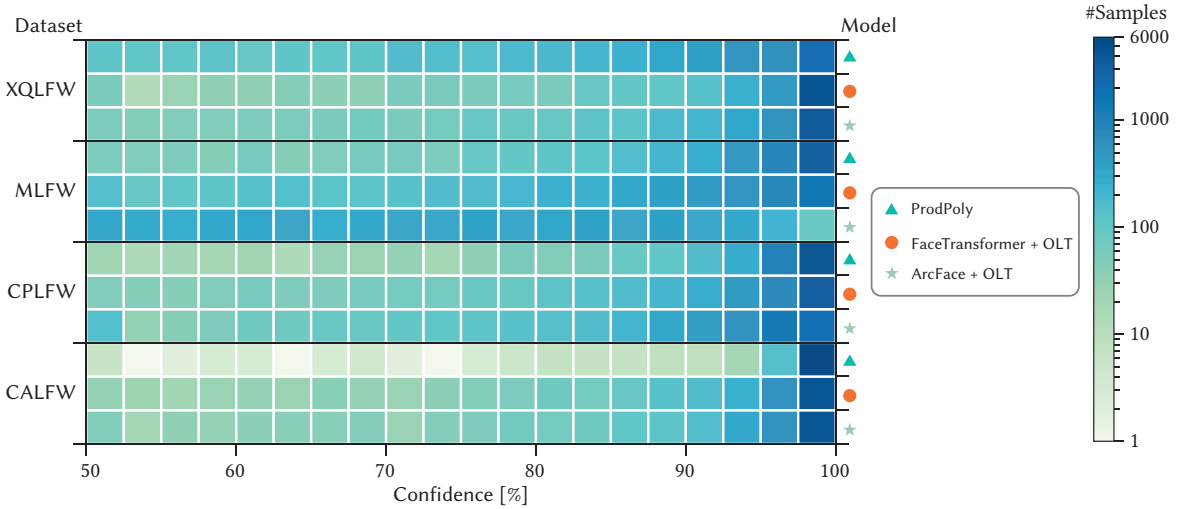


Figure 7.1.: C-Score [4†] distributions of three state-of-the-art face verification models trained on MS1M-V2 [47, 62] on several face verification benchmark datasets as in [5†].

fine-tuned with the *Octuplet Loss Training* (OLT) [2†] method (see also Chapter 5) as well as the ProdPoly network [107] are analyzed. The following datasets are utilized in the experiment: *Cross-Age Labeled Faces in the Wild* (CALFW) [94], *Cross-Pose Labeled Faces in the Wild* (CPLFW) [92], *Masked Labeled Faces in the Wild* (MLFW) [99], and *Cross-Quality Labeled Faces in the Wild* (XQLFW) [3†]. The C-Score distributions are presented in Figure 7.1. The colors represent the amount of classified FV image pairs within the respective C-Score range. In general, there is no clear pattern recognizable across the datasets or models. However, the most challenging dataset for the models appears to be the masked dataset MLFW, which is in line with the low accuracy (see also Table 7.3). Interestingly, the ProdPoly model takes it's decisions on the CALFW database with significantly higher C-Score than other models. Moreover, the ArcFace + OLT model is significantly less confident in it's decisions on the MLFW database. This preliminary analysis motivates further investigation into what are these edge cases where the model show low C-Score and potentially wrong predictions, and could human operators potentially perform better in these cases.

7.3. Methodology

Inspired by the recent work of Zheng *et al.* [217], which introduced a fusion of an *Artificial Neural Network* (ANN) with pathologists for predicting EBVaGC[iii] from histopathology, the aim of the work in this chapter is to propose and analyze a similar fusion of state-of-the-art *Deep Learning* (DL) networks with human decisions for FV. In contrast to the work of Carragher *et al.* [218] human operators are used to evaluate only machine predictions with low confidence. This section first

[iii] Epstein-Barr Virus (EBV)-associated Gastric Cancer

outlines the survey construction, followed by the user interface design. Finally, the algorithm to fuse human operator's and machine's decisions is introduced.

7.3.1. Survey Construction

Due to the large number of image pairs (6 000) per dataset (4), it's not feasible to use all pairs for each participant in the survey as this would overwhelm the users. The survey was divided into five parts, with specific image pairs from various datasets (see Section 7.2) being selected to manage this issue effectively.

The aim for *Survey 1* is to first construct a reference dataset, which is identical for all participants and includes image pairs of different modalities and difficulty. Therefore, based on the mean C-Scores across the three FV models (see Section 7.2) the image pairs are divided into four bins for each database. From each bin, one genuine and one imposter image pair are randomly selected, resulting in a total of 40 image pairs. This dataset is denoted as the BASE benchmark and serves as the reference for the subsequent surveys.

Following up, *Survey 2–5* are constructed to investigate the performance of human operators in comparison to the FV models on the different modalities, *i.e.*, masked, low-quality, cross-pose and cross-age images. Each of the survey is constructed according to the following procedure: First, the image pairs are divided into ten equally sized bins based on the mean C-Score of all models. From each bin, 30 image pairs are randomly selected. This is done to represent the whole range of C-Score levels of the models. Additionally, all image pairs where the minimum C-Score of all three models falls below a threshold of 50% are included in the pool, to ensure that all low confident predictions are evaluated by human operators. This procedure results in four image pair pools containing in total 7 445 tasks. Table 7.1 summarizes the image pair statistics for all constructed surveys. The number of image pairs for the CALFW dataset is significantly lower than for the other datasets, as the mean C-Score of the models is much higher on this dataset. In contrast, the most workload for the participants in terms of the number of tasks is required for the MLFW dataset, as the models have the lowest mean C-Score on this dataset. The ratio between genuine and imposter pairs is slightly imbalanced, with more imposter pairs than genuine pairs. This is due to the fact that the models are

Table 7.1.: Statistics for the five conducted surveys (*Survey 1–5*) constructed from various face verification benchmark datasets.

	Datasets				
	BASE	CALFW [94]	CPLFW [92]	MLFW [99]	XQLFW [3†]
Survey	1	2	3	4	5
# Pairs in Pool	40	684	1 385	3 561	1 815
# Pairs per User	40	12	24	60	31
% of Total Pairs	–	5.37	10.28	28.58	13.58
% Genuine	50.00	47.08	44.55	48.16	44.90

generally more confident in their predictions for genuine pairs, resulting in less confident predictions for imposter pairs.

In contrast to *Survey 1*, the image pairs for *Survey 2–5* are unique for each participant, *i.e.*, no image pair is displayed twice. The ratio of genuine and imposter pairs varies for each user. The survey was constructed to target 60 participants, which results in a total of exactly 167 tasks per participant, with the exception of the last three participants that receive fewer image pairs as the respective image pair pools were depleted. This amount of tasks was chosen, based on the findings of a prior user study, which is not covered in this work, to ensure that the participants are not overwhelmed by the workload and can complete the survey in a reasonable amount of time. Overall, 9 845 tasks are displayed to the participants. The tasks of *Survey 1* are displayed in a random order to the participants, to avoid unintended information exchange between users. For *Survey 2–5*, the image pairs are randomly selected from the corresponding pool of images for each participant.

7.3.2. Survey Design

The survey is distributed with a web-link to the participants. To participate, users must first register on the survey website, where they are asked to voluntarily provide their age, ethnicity, and gender. After registration, general information about FV and the survey procedure is displayed. Participants can complete the surveys at their convenience and the surveys are designed to be completed in multiple sessions, with the progress saved for each participant.

When the user starts a survey, the tasks are displayed one after the other. For each image pair, the user must decide whether the two images show the same or different identities and rate their C-Score regarding the answer (see left part of Figure 7.2). The user can adjust a slider, which ranges from "uncertain" to "certain" without any numerical indicators or additional information. The default position of the slider is at the midpoint, corresponding to a value of 0.5, with the internal range of values extending from 0 (uncertain) to 1 (certain). Note that internally the human operator's certainty C^* is converted to the C-Score range of $[0.5, 1]$, which is done via the formula $C = 0.5 + 0.5 \cdot C^*$. The duration of each task is also recorded in the background, by measuring the time elapsing from the page load until the user clicks the "next" button. It is important to note that this approach may not be entirely accurate, as some users might load the question, leave their computer for an extended period, and then return to answer the question. To account for this, outliers ≥ 60 s from the evaluation of the survey duration are mapped to a fixed value of 60 s. To force the user to make a decision, the "next" button is disabled until the user has moved the slider. After completing at least one question in the survey, participants can leave the survey by clicking on the exit button and review the accuracies of their answers thus far, along with the corresponding FV model's performance in a dashboard (see right part of Figure 7.2). When the user returns to the survey, they are redirected to the last question they answered. To show the user's progress, the number of completed tasks is shown at each task and in the dashboard including additionally the status of all survey.

Participants are not informed of the ratio of genuine and imposter pairs or the labels of the image pairs. To prevent any influence on the participants' performance, the models' predictions are not displayed to the participants as proposed in [210]. Moreover, there is not feedback on the correctness

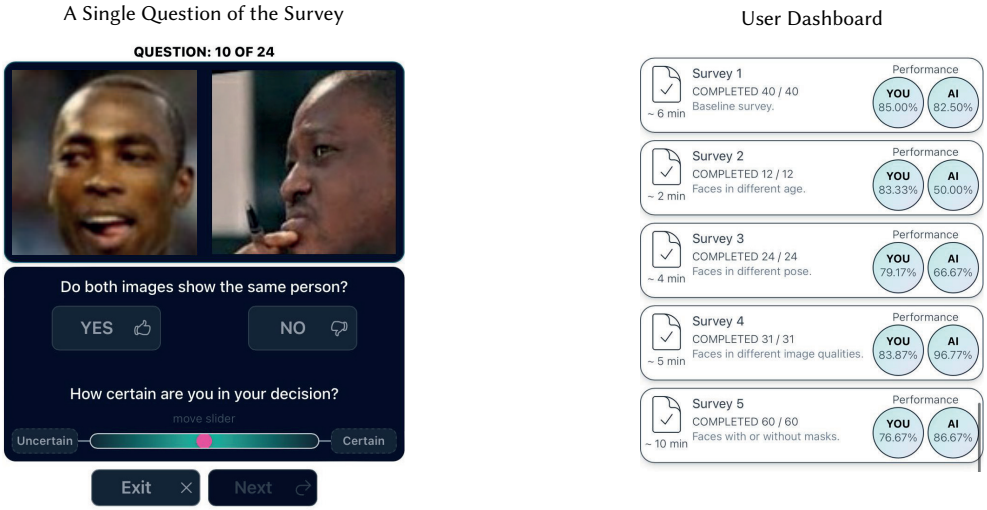


Figure 7.2.: Left: The design of the face verification task with an example image pair from CPLFW [92] database. Right: The dashboard for the users to see their results and survey completion status.

of the answers, as this could influence the participants' decisions on subsequent questions, particularly by revealing whether the previous answer was correct or incorrect [209].

More technically speaking, the survey was hosted on the Google Cloud Platform and running inside an App Engine connected to an SQL database, which allows to complete the survey in multiple parallel sessions. The code for the survey platform is written in Python using the Flask [208] framework.

The survey was advertised to colleagues, undergraduates, friends, and family to encourage participation. To motivate the research community to follow up on this survey, the code and data for the developed survey tool was published on GitHub[iv].

Data for this study was gathered online over a 14-day period in March/April 2023.

7.3.3. Human-Machine Fusion

To examine the potential of combining human and machine decisions in the FV task, a simple fusion algorithm is proposed. This algorithm is designed to decide based on the C-Score of human operators and machine models whether to accept the human decision or the machine decision. The algorithm is illustrated in Figure 7.3 can be outlined as follows: First, a minimum C-Score threshold t_M of 68.5 % is set for machine decisions to be directly accepted. Any machine decisions falling below this threshold are considered for comparison with the human confidence. If the human C-Score for a particular question exceeds a specified threshold t_H of 55.5 % and surpasses the machine C-Score by at least

[iv] A widely-used platform that facilitates version control and collaboration, allowing developers to host and review code, manage projects, and build software together.

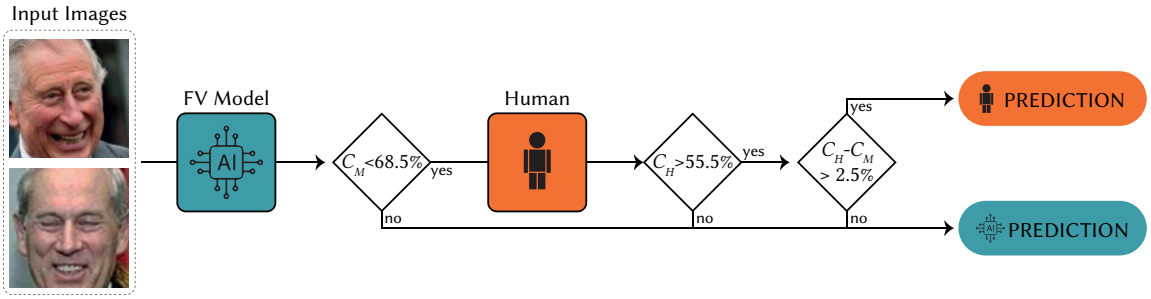


Figure 7.3.: Flowchart of the proposed human-machine fusion algorithm as introduced in [5†].

2.5 %, the human decision preferred over the machine. After collecting the human feedback in the surveys on the pool of image pairs for all datasets, the proposed fusion algorithm is applied to all 6 000 pairs in each dataset.

7.4. Results

First, a few challenging edge case scenarios are presented. Then, the characteristics of the participants in the survey are discussed, followed by a detailed analysis of *Survey 1*. After reviewing the results of the remaining *Survey 2–5*, the human-machine fusion strategy is evaluated.

7.4.1. Edge Cases in Face Verification

As depicted in Figure 7.1 of Section 7.2, there are many image pairs in the datasets where the machine models have a very low C-Score in their predictions.

After conducting the surveys, there exists a subset of these image pairs where the human operators perform significantly better than the machine models. Figure 7.4 presents four genuine and four

Figure 7.4.: Example image pairs from CALFW, CPLFW, MLFW, and XQLFW where human decisions are correct and with higher C-Score than face verification models.

imposter image pairs holding this characteristics. The upper left row shows genuine examples from the CALFW and CPLFW datasets, which are of a good image quality with no occlusions. Hence, they seem to be easy for humans to solve. The lower left row displays genuine examples from the MLFW and XQLFW database. These examples are more challenging due to the low image quality and occlusions, but still solvable for humans. On the right side, imposter examples from the CALFW, XQLFW, and MLFW datasets are shown and although they are more challenging, they are still solvable for humans. This variance showcases humans' ability to outperform machines in FV, particularly in conditions that are traditionally challenging for machine models, thus highlighting the nuanced capabilities of human perception over current machine models in specific scenarios.

7.4.2. Characteristics of Study Participants

In the registration process, participants were asked to provide their age, ethnic background and gender on a voluntary basis. The distribution of these characteristics is shown in Figure 7.5. The mean age of the participants is 37.22 (±17.41) years within a range from 20 to 72. The majority of the participants being male (60.00 %), followed by 33.33 % female and 6.67 % preferring not to state their gender or marked other. Considering the ethnic background, 53 participants stated as 'Caucasian', one participant as 'Mixed', one as 'Other', two as Middle Easter and three provided no ethnic background information. In conclusion the participants are predominantly Caucasian, which is somewhat closely related to the identities inside the datasets used in this study as reported by [219]. The results of the following surveys are hence to be interpreted with caution, as the results may not be generalizable to the entire population.

7.4.3. Survey 1

The first survey *Survey 1* was conducted to establish a reference dataset for the subsequent surveys. The results of this survey are presented in this section. First the overall feedback collected from the 60 participants on the 40 image pairs in the BASE dataset is illustrated in the left of Figure 7.6. The answers are horizontally categorized depending on the ground truth label of the image pairs and the

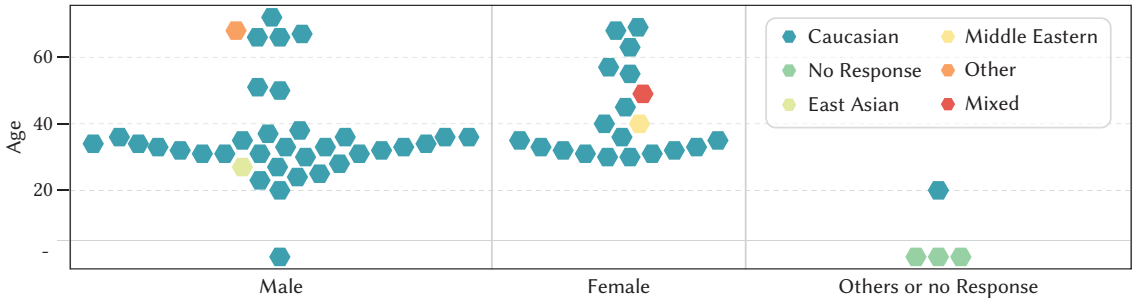


Figure 7.5.: Characteristics of the participants regarding age, gender and ethnicity on a voluntary basis.

color indicated a wrong (red) or false (green) classification of the task. Note that the ratio between genuine and imposter pairs is equal in this BASE dataset. On the vertical axis, the recorded C-Score values from the participants is displayed. In both distributions it is visible that the minimum and maximum possible C-Score values are very frequent, *i.e.*, the human operator puts the slider entirely to the left (uncertain) or right (certain). In between this range, the distribution is more sparse, with a slight tendency towards higher C-Score values. The C-Score levels of the human operators are distributed according to two approximately Gaussian curves, ranging from 50 to 75, and from 75 to 100, respectively. There is a noticeable dip in C-Score levels around the 75 mark, likely a consequence of the slider's design characteristics (see Section 7.3.2). The C-Score value distribution for the genuine pairs show higher density of wrong classified tasks compared to the imposter pairs.

A histogram for the duration, which each participant took to answer a task, is shown in the right of Figure 7.6. The overall mean duration per task is 10.71 s (±8.38 s). When analyzed individually between the participants, *i.e.*, taking the mean on all tasks per human operator, the mean duration is 10.84 s (±5.25 s), and when considered in between the task, *i.e.*, taking the mean on all operator for each specific task, it is 10.73 s (±2.50 s). This indicates that there is a bigger variation in between the human operators, than in between the different questions. In other words, the difficulty is relatively consistent across all tasks, but some participants are consequently faster in answering the questions than others. Differentiating between the time taken for the genuine and imposter image pairs, it is visible that participants took on average 11.44 s (±8.72 s) for genuine image pairs and 9.98 s (±7.96 s) for imposter image pairs. From this, one can interpret, that if there are two different identities displayed to a human operator, this impacts the decision making process and takes shorter than if the same identity is displayed. In other words, human operators think longer about their decision if the same identity is displayed. This finding is in line with the results on the mean durations depending

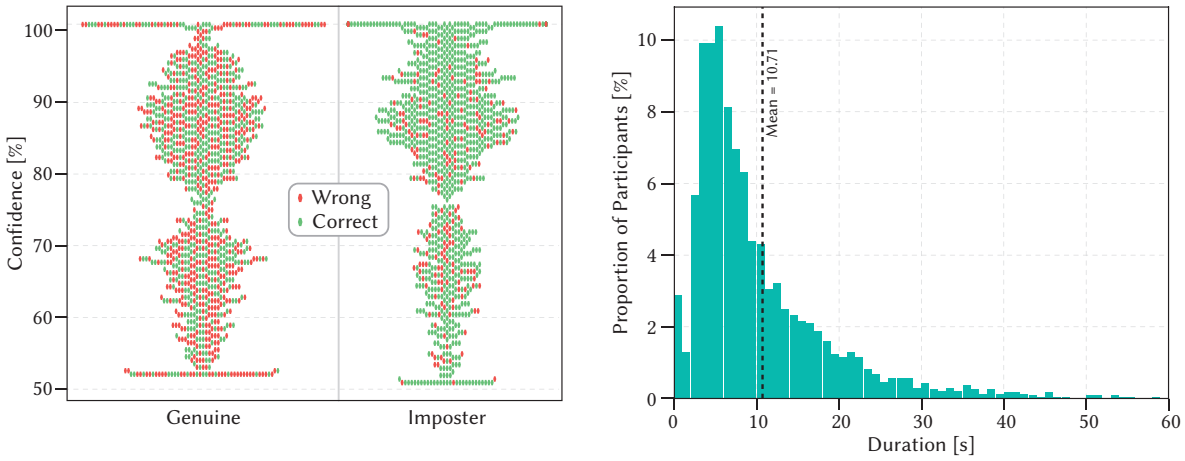


Figure 7.6.: On the left, distribution of participant's answers with respect to their rated C-Score on genuine and imposter image pair tasks of the BASE dataset used in *Survey 1*, adapted from [5†]. On the right, histogram of the durations for the tasks in *Survey 1*.

on the answer, where the participants took on average 11.66 s (±8.78 s) for giving a "same" answer and 10.22 s (±8.12 s) for giving a "different" answer.

Although no correlation could be detected between age and accuracy or confidence, in terms of duration, older participants took longer to answer the questions.

On the left in Figure 7.7 the FV accuracy is put into relation with the C-Scores of the participants. Each data point pair corresponds to all responses from a unique participant in the survey. The responses are divided into genuine (green) and imposter (red), with the mean C-Score and mean accuracy of the responses compared. On average, both the C-Score and accuracy of participants were lower for genuine pairs than for imposter pairs. This can also be partially identified by the trend of the connections between data points, which mostly run from bottom left to top right. Looking at the absolute values, one can determine that the accuracy for genuine pairs often falls below 50 % percent, which is worse than random guessing. In general, participants much more frequently gave the prediction "different" (1 586) than "same" (814). Since both pairs occurred equally often in the survey, higher accuracy for imposter pairs is also expected. Considering all tasks to be answered, the average accuracy among the participants is 67.25 % (±8.53 %) and the average C-Score is 79.83 % (±7.31 %). The correlation coefficient between accuracy and C-Score is 0.44 ($p < 0.001$), indicating a moderate correlation.

The right chart in Figure 7.7 now relates the answers of the participants to the 'machine' predictions on the same task. In only one task out of the 40 tasks in the BASE dataset human C-Score is higher than the machine C-Score and the human prediction is correct, while the machine prediction is incorrect. In contrast, in six tasks the human prediction is incorrect, while the machine prediction

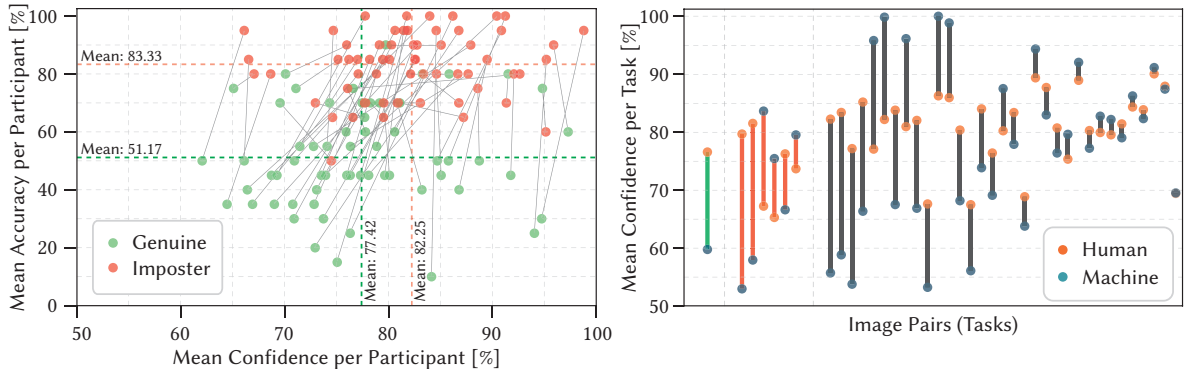


Figure 7.7.: Left: Accuracy versus C-Scores by participant for BASE dataset tasks, adapted from [5†]. Right: Mean C-Score of human and machine predictions (FaceTransformer + OLT, ArcFace + OLT, and ProdPoly, trained on MS1M-V2) per task. The green bar represents tasks where human predictions are correct but machine predictions are incorrect, the red bars indicate scenarios where machine predictions are correct but human predictions are incorrect, and the black bars show cases where both human and machine predictions are either correct or incorrect for the same task.

is correct. In the remaining 33 tasks, both the human and machine are either correct or incorrect. This indicates that there is a potential for human operators to outperform the machine models in some cases, but the machine models are generally more reliable. However, in this database, it is preferred to use the machine models' predictions over the human operators' predictions, as the human operators' decision would lead to six more incorrect predictions, while only one more correct prediction. Considering only human decisions with higher C-Score than the machine decisions, one more correct prediction would be made, but still also two more incorrect predictions. In contrast to the human operators' accuracy of 67.25 %, the mean machine accuracy on the BASE dataset is 85.00 %. The mean C-Score of the machine models is 76.20 % (±14.15 %), which is lower than the human operators' C-Score of 79.83 % (±6.49 %) among the tasks. Note, that the standard deviation in this case is calculated among the tasks, in contrast to the standard deviation among the participants in the previous paragraph.

7.4.4. Survey 2–5

To evaluate the human performance and the relation to machine predictions on more challenging datasets, this section presents the results on the *Survey 2–5*. Note, that in this scenario each task is answered by exactly one participant. As in the previous section, the human and mean machine C-Scores are reported for *Survey 2–5* (CALFW, CPLFW, XQLFW, and MLFW database) in Figure 7.8. In contrast to the chart for the BASE dataset (see Figure 7.7) the tasks on which human and machine predicted equally are skipped and only tasks where human C-Score is higher than the machine is selected to ensure a better visibility. Here also the mean prediction across the three machine models is taken into consideration. The results on *Survey 1* show that taking the decision with higher confidence,

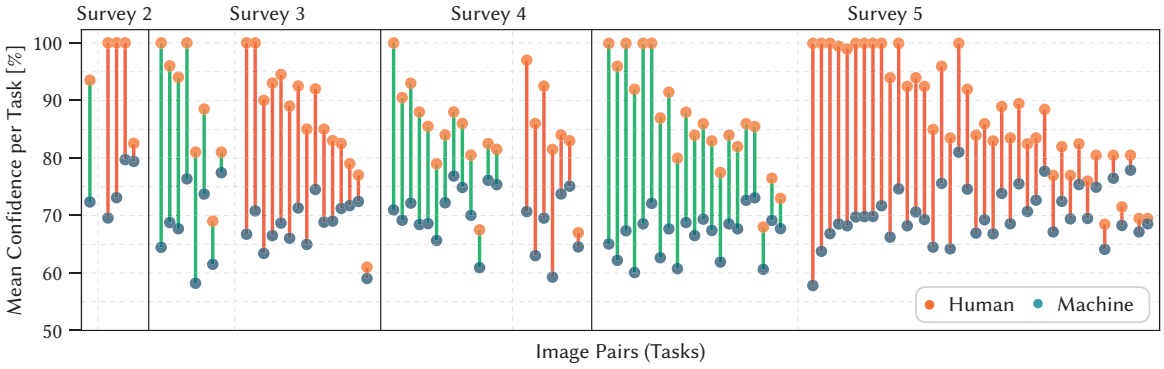


Figure 7.8.: Mean C-Score of human and machine predictions (FaceTransformer + OLT, ArcFace + OLT, and ProdPoly, trained on MS1M-V2) per task for *Survey 2–5*. Green bars show correct human but incorrect machine predictions, red bars for incorrect human but correct machine predictions, and black bars where both human and machine predictions are either correct or incorrect. Cases where the machine C-Score was higher than the human confidence, as well as cases where both gave the same predictions, were disregarded.

Table 7.2.: Accuracy and C-Score of machine models (fine-tuned on MS1M-V2 database) and human operators in *Survey 2–5*.

	Accuracy [%]				Confidence [%]			
Model	Survey 2	Survey 3	Survey 4	Survey 5	Survey 2	Survey 3	Survey 4	Survey 5
ArcFace + OLT [2†]	93.37	86.92	91.87	68.51	94.06 (± 10.53)	88.62 (± 13.36)	92.23 (± 12.28)	71.14 (± 13.16)
FaceTransformer + OLT [2†]	94.67	90.83	94.12	83.33	95.02 (± 9.70)	91.76 (± 12.04)	94.37 (± 10.84)	83.53 (± 14.57)
ProdPoly [107]	95.95	92.34	84.40	90.30	98.30 (± 3.26)	95.42 (± 8.47)	85.59 (± 14.89)	91.22 (± 11.99)
🧍	66.23	68.38	70.19	64.48	82.93 (± 13.13)	80.54 (± 14.43)	75.06 (± 16.30)	80.18 (± 13.81)

with human operators only one misclassified task of the machine can be corrected. However, at the same time four tasks are misclassified by human operators. The same behavior, while in a different magnitude of tasks is present for *Survey 3* and *Survey 5*. In contrast, for *Survey 4* the human operators are able to correct more misclassified tasks of the machine models. This indicates that the human operators are more capable of solving tasks on the XQLFW dataset than on the other datasets and a simple fusion strategy (just decide upon higher confidence) on this dataset could be beneficial to improve the overall performance.

In addition to the results in Figure 7.8, the accuracy and C-Score of the machine models and human operators on *Survey 2–5* are summarized in Table 7.2. Note, that only the subsets of image pairs of the origin datasets (see Section 7.3.1) are considered in this table. The accuracy of the three selected models is significantly less on the surveys compared to the overall dataset (see Table 2.3), which is reasonable due to the selection of the image pairs with low confidence. In contrast to the machine models accuracy, the human operators' accuracy is remarkably low in all surveys. The same holds for the C-Score values, except for the ArcFace + OLT model on the MLFW dataset. This indicates that the human operators are not very confident in their decisions, which is also reflected in the low accuracy. However, as the Figure 7.8 revealed, there are some tasks where the human operators are more confident and correct than the machine models.

7.4.5. Human-Machine Fusion

The previous results showed that human operators are able to outperform machine models in some cases, particularly in scenarios, where machine models have a low C-Score for their decision. This leads to the question of whether a fusion of human and machine is beneficial to improve the overall performance. In this section, the simple fusion strategy as introduced in Section 7.3.3 is applied to the all image pairs of the four datasets. This implies that only a selective segment of the full datasets is taken into account for fusion, particularly those chosen based on the machine models' C-Score levels as outlined in Section 7.3.1. Evaluations of all other pairs of images are conducted exclusively by machine models. This strategy offers two primary benefits: It necessitates human evaluation for only a specific subset of image pairs, thereby saving considerable time given the large volume of pairs within the datasets. Furthermore, it guarantees that human evaluators only review those image pairs that pose difficulties for the machine models.

Table 7.3.: Face verification accuracy for the machine models and the fusion approach. The icon 🧍 is denoting human decisions. The values in parentheses indicate the improvement over the machine model. The number of human decisions considered in the fusion approach is also denoted. All models are trained on the MS1M-V2 database.

	Accuracy [%] for Datasets			
Model	CALFW [94]	CPLFW [92]	MLFW [99]	XQLFW [3†]
ArcFace + OLT [2†]	93.85	88.37	73.53	93.27
ArcFace + OLT [2†] + 🧍	94.03 (+0.18)	88.83 (+0.47)	75.88 (+2.35)	93.65 (+0.38)
# 🧍 Decisions	241 (4.02 %)	470 (7.83 %)	1871 (31.18 %)	257 (4.28 %)
FaceTransformer + OLT [2†]	94.93	91.58	85.63	95.12
FaceTransformer + OLT [2†] + 🧍	94.93 (0)	91.73 (+0.15)	85.70 (+0.07)	95.28 (+0.17)
# 🧍 Decisions	187 (3.12 %)	311 (5.18 %)	778 (12.97 %)	172 (2.87 %)
ProdPoly [107]	96.03	92.75	91.30	86.90
ProdPoly [107] + 🧍	96.13 (+0.10)	92.82 (+0.07)	91.35 (+0.05)	88.05 (+1.15)
# 🧍 Decisions	17 (0.28 %)	125 (2.08 %)	325 (5.42 %)	605 (10.08 %)

Table 7.3 displays the outcomes of fusing human and machine decisions across four benchmark datasets. The values in parentheses indicate the improvement over the purely machine-based model. It is evident that the fusion surpasses the performance of machine models alone on all datasets. Moreover, it becomes clear that the lower the accuracy of the model, the more human decisions are required to enhance the overall performance of the fusion approach. Generally, it can be stated that the fusion approach offers minimal benefits for the CALFW dataset. For all other datasets, it provides added value; however, it is not distinctly clear whether this depends on the model or the dataset, as no clear pattern emerges.

7.5. Conclusion

This chapter presents a study on human-machine fusion in FV tasks. A set of surveys was conducted to evaluate the performance of human operators in comparison to machine models on various benchmark datasets. The surveys are constructed from image pairs of several well known FV benchmark datasets and are designed to investigate the performance of human operators in comparison to the machine models on different modalities, *i.e.*, masked, low-quality, cross-pose and cross-age image pairs. A survey on the BASE dataset was conducted to establish a reference dataset for the subsequent surveys, containing image pairs of different modalities and difficulty. The key findings on this BASE dataset are that human operators perform better on imposter image pairs than on genuine pairs and at the same time have a higher C-Score in their decisions. The subsequent surveys on subsets of CALFW, CPLFW, XQLFW, and MLFW reveal that human operators are capable of outperforming machine models in very specific scenarios. However, the machine models are generally more reliable.

To investigate the potential of combining human and machine decisions in the FV task, a simple fusion algorithm was proposed. This algorithm is designed to decide based on the C-Score of human operators and machine models whether to accept the human decision or the machine decision. The

results of the fusion approach show that the fusion surpasses the performance of machine models alone on all datasets. The lower the accuracy of the model, the more human decisions are required to enhance the overall performance of the fusion approach.

However, the results of the study are limited by the low number of participants and the very one-sided ethnic background of the participants (*Cf*., [220]). Moreover, since the identities in the dataset are predominantly public figures, the study encompasses a combination of familiar and unfamiliar face matching, which could also influence generalizability (*Cf*., [221]).

"I was taught that the way of progress was neither swift nor easy."

– Marie Curie

Conclusion and Outlook

8.1. Conclusion

Having introduced the main body of this work in Chapter 1 and laying the groundwork in Chapter 2, now the time has come to provide answers to the research questions posed in the introduction and point towards interesting directions for future work:

Research Question 1: *How does image resolution impact the performance of face recognition systems?*

Chapter 3 presents the exhaustive analysis, published in [1†], to understand the impact of image resolution on the performance of *Face Recognition* (FR) systems. By simply downscaling and upscaling images to different raw-resolutions the performance of several state-of-the-art FR models is benchmarked on various challenging datasets. The *Face Verification* (FV) accuracy is measured for *Cross Resolution* (CR) and *Equal Resolution* (ER) scenarios and reveals the impact of image resolution on the performance of FV systems. Moreover, the analysis is extended by looking at the feature distances of the FV models to understand the impact of image resolution on the feature space. This work provides a fundamental understanding of the impact of image resolution on FR systems. It highlights the importance of considering image resolution in evaluating FR systems.

To further enhance the validity of this analysis, the degradation method could be further improved by utilizing a more realistic downscaling approach, as *e.g.*, the generation of *Low Resolution* (LR) images applied in Chapter 5. Furthermore, the impact of image resolution could be investigated in more detail within the models, *e.g.*, one could examine the individual layer maps within a *Convolutional Neural Network* (CNN), or the effects on the activations within a CNN.

Research Question 2: *What strategies can be developed to enhance the robustness of face recognition systems against variations in image resolution?*

This question is addressed by the development of three strategies for making FV systems more robust to image resolution, which are described in Chapter 4 and published in [1†] and [2†]. The *Resolution Augmentation Training* (RAT) strategy is a simple approach to enhance the robustness of FV models by augmenting the images towards their raw-resolution during training. A more sophisticated approach is the *Contrastive Loss Training* (CLT) technique. The contrastive loss principle is applied additionally for training an FR network. By adding the feature distance between *High Resolution* (HR) and LR images, the network is forced to learn a more robust feature representation. The third method

incorporates the triplet loss principle and is called *Octuplet Loss Training* (OLT). By utilizing four triplet loss terms, each building a relation between HR and LR images, the network is furthermore forced to not only learn the relation between LR and HR version of a single image, but also the relation between different images and different identities. Combined with a hard-negative mining strategy, the OLT method is used to fine-tune existing FV models and further enhance their robustness against variations in image resolution. The results show that all methods can indeed improve the overall performance on a variety of FV benchmark datasets. While the RAT and CLT methods moderately improve the performance, the OLT method shows the most promising results. Another advantage is the adaptability of the OLT to arbitrary network architectures. In conclusion, the research conducted in Chapter 4 highlights the importance of considering image resolution in the training of FR systems. It provides strategies to enhance the robustness of FV models against variations in image resolution.

For future work, a promising direction is the exploration of the OLT method in other challenging scenarios, such as masked FR, which shows many analogies to CR scenarios. Another interesting direction could be to further optimize the fine-tuning process, particularly the online hard negative mining strategy, which is computationally expensive. Moreover, the amount of data used for the fine-tuning process is quite large, making the training time long. A prior offline mining strategy could be developed to reduce the amount of data needed for the fine-tuning process while achieving similar results.

Research Question 3: *How can a new benchmark be designed to measure the performance of face recognition systems on images with different resolutions more accurately and precisely than current methods?*

To address this question, the *Cross-Quality Labeled Faces in the Wild* (XQLFW) [3†] benchmark dataset creation was introduced in Chapter 5. After having accessed the image quality of the *Labeled Faces in the Wild* (LFW) dataset, the XQLFW dataset was created by synthetically downscaling a subset of the images via subsequent blurring and nearest-neighbor downscaling to different resolutions. The final dataset was then created by an algorithm that selects different quality level images to meet the same evaluation protocol characteristics as the original LFW dataset. The dataset was then used to evaluate the performance of FR systems on images with different resolutions. It reveals that the robustness of several state-of-the-art FV models to image resolution varies. The increasingly frequent use of the XQLFW dataset in the field of FV research highlights the significance and impact of this work. Establishing a dedicated benchmark focusing on the raw-resolution of images is a promising direction for future work in the field of FR, as it standardizes the evaluation of FR systems on images with different resolutions.

In the future, the dataset could be further improved by the evolvement of ever more realistic synthetic image downscaling to meet the natural characteristics of LR images even closer. The methodology of the XQLFW dataset creation could be extended to video datasets, where the resolution of the video frames is also a crucial factor for the performance of FR systems. Focusing on face identification, the methodology to create the evaluation protocol could be extended to generate face identification evaluation protocols additionally. In general, to make a comparison of results in the field of FR more reliable, the establishment of dedicated benchmarks focusing on the challenges of FR is a promising direction for future work in the field of FR.

Research Question 4: *What methods can be established to provide clearer explanations for the decisions made by face recognition systems?*

To answer this question, Chapter 6 introduced the *Confidence Score* (C-Score) and *Explanation Map* (X-Map) model-agnostic methods [4†] and showed the potential to provide meaningful explanations for the decisions made by FV model. The C-Score, in other words, a level of confidence of any given model decision, can be calculated by applying the model on a bunch of labeled image pairs, from which a sigmoid function is estimated to map the feature distance to a confidence value. The X-Map method is a visualization technique derived by systematically occluding parts of the face images and measuring the deviation of the feature distance. For a given face image, the X-Map highlights the regions of the face image that are most important for the decision made by the model. The results showed that the C-Score is more meaningful than the feature distance value itself, as this value is not directly interpretable and varies between different FV models. The qualitative results of the X-Map method showed that the method can highlight the regions of the face image that are most important for the decision made by the model. To showcase the contributions, an interactive web platform was developed and published to visualize the X-Map results for five popular FV benchmarks. Both methods have been applied beyond the scope of CR scenarios to further contribute to the explainability of FV models and provide a deeper understanding of the decisions made by FR models.

Future work in the field of explainable FR could elaborate on taking not only the feature distance itself but also an image quality or raw-resolution estimate, or other factors *e.g.*, age, pose, gender, race, or occlusions into account. As the results in Chapter 6 reveal, the interpretation of the X-Map is not straightforward, and further research could be conducted on improving the interpretability of the X-Map by adjusting the algorithm or finding a more suitable visualization technique. Especially in CR scenarios, the X-Map itself might not explain sufficiently the model's decision, as for human observers, the face is hardly recognizable in very LR images. Therefore, combining super-resolution techniques with the X-Map approach might be interesting.

Research Question 5: *How do humans perform in borderline cases of face recognition, and can an algorithm be developed to effectively fuse human and machine decisions to improve overall accuracy in face recognition tasks?*

Chapter 7 first investigated the challenging cases of FV in [5†], in where machine models fail to predict correctly. By utilizing the C-Score, cases could be identified where the model is not confident in its decision. A user study hypothesized that humans perform better in those edge cases in summary than machine models. To verify this hypothesis, a user study was conducted in [5†], where participants were asked to do an FV task on a selection of image pairs of several benchmark datasets. The study revealed that there exist cases where humans can verify the given facial images better than machine models and, importantly, with higher confidence. To answer the second part of the research question, a simple fusion strategy was developed to combine the predictions of a model and human operators based on their C-Scores. The results showed that the fusion strategy can improve the overall accuracy of FR systems on several datasets.

A possible direction for future work could be a calibration of the C-Score of the human operators because this is a subjective value and could somehow be normalized. Another possibility is the utilization of X-Maps from machine models to provide human operators with additional information to make their decisions. Furthermore, the fusion strategy could be further improved by utilizing more sophisticated algorithms or even incorporating machine learning techniques to decide not only based on the C-Score but also on other factors, such as the images themselves, the quality of the images, or the X-Map results. In general, the study conducted in Chapter 7 is limited due to the small number of participants, and a more extensive study could be conducted to further validate the results.

8.2. General Outlook

After having answered the research questions, a more general outlook on exciting directions for future work in the field of CR FR is given in the following:

Explainability. The field of *Explainable Artificial Intelligence* (XAI) is growing and has seen much interest in recent years. The ability to explain the decisions made by a model is crucial for the acceptance of *Artificial Intelligence* (AI) in real-world applications [222]. In the context of robust FR, explainability is important for developers to understand the model's decisions and identify potential to make them more robust towards image quality or resolution.

Network Architectures. Recent advancements tend to show a shift in research towards utilizing the *Vision Transformer* (ViT) technology in robust FR models and could be a promising direction to follow up. The exploration of hybrid architectures that combine the strengths of CNNs and ViTs also represents a promising direction for future work in FR.

Bias. The issue of bias in FR has been a topic of interest in recent years. Previous approaches to bias mitigation in FR have primarily focused on pre-processing the training data, incorporating penalties during training to mitigate bias, or post-processing predictions to reduce bias. However, these methods have shown limited success in addressing bias in challenging FR tasks. Future work could focus more on the networks themself, such as very recently Dooley *et al.* [223] did, who proposed a neural architecture search for fairness, jointly with a search for hyper-parameters.

Synthetic Face Generation. Advancements in synthetic face generation have also led to the creation of synthetic datasets. Wood *et al.* [224] introduced a dataset of $100,000$ fake images generated by combining a procedurally-generated parametric 3D face model with a comprehensive library of hand-crafted assets. The dataset is designed to be used for training and evaluating computer vision analysis, such as landmark localization and face parsing. The authors show that synthetic data can both match real data in accuracy as well as open up new approaches where manual labeling would be impossible. This is a promising direction and could probably be used to create more realistic and much larger synthetic datasets that incorporate a wide range of image resolutions. Training on such datasets, in turn, could potentially lead to more robust FR models and push the boundaries of the current state-of-the-art.

"What we call the beginning is often the end. And to make an end is to make a beginning. The end is where we start from."

— Thomas Stearns Eliot

Notation

This appendix briefly introduces the notation used throughout this work. The notation is based on the style manual published by the Institute of Electrical and Electronics Engineers (IEEE) [225].

References

Self-citations are highlighted by a †, *e.g.*, [1†]. Any other citations are not highlighted, *e.g.*, [1]. References are ordered by appearance in the text. Multi-references are separated by a comma in alphanumerical order, not in the order of appearance in the text, *e.g.*, the works of Knoche [2] and Rigoll [1] were both completed in 2024, or in other words, the works [1,2] were both completed in 2024. To avoid clutter from frequently recurring references, each reference is usually repeated only once per paragraph unless an attribution is ambiguous. Besides, footnotes are marked by Roman letters.[i]

Acronyms

Each acronym is introduced at least once in the text at the point of its first occurrence in every chapter. The assignment of articles to acronyms is guided by the pronunciation of the first letter of the acronym within the context of the research field. *E.g.*, we refer to 'an' face recognition (FR) system and 'a' convolutional neural network (CNN), but 'a' residual network (ResNet). To prevent ambiguity in plural forms, an 's' is added to the end of acronyms for pluralization, as in 'CNNs' being the plural form of CNN.

Mathematics

The mathematical notation follows the *International Organization for Standardization* (ISO) 80000–2 standard where possible [32]. The most important aspects are summarized below:

- Scalars are written in lower-case letters, *e.g.*, a.
- Constant scalars or scores are written in capital letters, *e.g.*, C.

[i] This is a footnote

- Vectors are written in bold letters, *e.g.*, $\boldsymbol{v}$.
- Matrices and tensors are written in bold capital letters, *e.g.*, $\boldsymbol{X}$.
- Sets are written in calligraphic capital letters, *e.g.*, $\mathcal{B}$.
- Functions are written in normal font, *e.g.*, f or in script font, *e.g.*, $\mathscr{L}$.
- The cardinality of a set is denoted as $|\cdot|$.

In addition to ISO 80000—2, the following rules apply:

- The i-th element of a vector $\boldsymbol{v}$ or a set $\mathcal{V}$ is denoted by v_i or in cases where a second index is needed as $\boldsymbol{v}^{(i)}$.
- If subscript is already occupied, methods or image resolution variants are denoted by a bracket superscript, *e.g.*, $A^{[\mathrm{m}]}$ or $I^{[\mathrm{LR}]}$.
- The identity of a facial image $\boldsymbol{I}$ is denoted by $\mathrm{id}\big(\boldsymbol{I}\big)$.
- The natural logarithm is denoted by log.
- Downscaling an image $\boldsymbol{I}$ by a factor s is denoted by $\downarrow_s(\boldsymbol{I})$.
- Upscaling an image $\boldsymbol{I}$ by a factor s is denoted by $\uparrow_s(\boldsymbol{I})$.
- The distance between two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is denoted by $\mathrm{d}(\boldsymbol{v}_1, \boldsymbol{v}_2)$.
- The absolute value of a scalar a is denoted by $|a|$.
- The floor function, which rounds a real number a down to the nearest integer, is denoted by $\lfloor a \rfloor$.
- The L2-norm of a vector $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|$.
- The max function with zero as lower bound of a scalar a is denoted by $[a]_+$.
- The min function with zero as upper bound of a scalar a is denoted by $[a]_-$.
- The function composition of g_θ and c_θ is denoted by $\mathrm{g}_{\theta^1} \circ \mathrm{c}_\theta$.
- The Δ symbol is used to denote a difference or change, *e.g.*, $\Delta C = C_2 - C_1$.
- The ∇ symbol is used to denote the gradient of a function, *e.g.*, ∇f.
- f_θ denotes an artificial neural network or a feature extraction network, g_θ a layer of a neural network, and c_θ a classifier network.

Supervised Student Works

The following list gives an overview of the supervised works during the course of the dissertation:

- Simsek Selim. "Repulsion Loss — Detecting Pedestrians in a Crowd". In: *Scientific Seminar*. 2018. Technical University of Munich.
- Goeb Stefan. "Wide Compression — Tensor Ring Nets". In: *Scientific Seminar*. 2018. Technical University of Munich.
- Habermayr Lukas. "Enhanced Deep Residual Networks for Single Image Super-Resolution". In: *Scientific Seminar*. 2019. Technical University of Munich.
- Germer Lena. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *Scientific Seminar*. 2019. Technical University of Munich.
- Gilg Johannes. "Attention Mechanisms in an Appearance Matching Network". In: *Research Internship*. 2019. Technical University of Munich.
- Gilg Johannes. "Single Image and Multi Image Super Resolution for Face Identification". In: *Master's Thesis*. 2019. Technical University of Munich.
- Maas Michael. "Implementation of 'Visualization of Activations' in a State-of-the-Art Face Recognition Network". In: *Research Internship*. 2020. Technical University of Munich.
- Ma Bowen. "Comparison of SOTA Super-Resolution Methods in the Domain of Face Recognition". In: *Research Internship*. 2020. Technical University of Munich.
- Maas Michael. "Single-Output-Neuron Cross-Resolution Face Verification". In: *Master's Thesis*. 2021. Technical University of Munich.
- Hamila Yassine. "Generating and Evaluating a Dataset for Realistic Cross-Resolution Face Recognition". In: *Bachelor's Thesis*. 2021. Technical University of Munich.
- Elkadeem Mohamed. "Quintuplet Loss X-Resolution Face Recognition". In: *Master's Thesis*. 2022. Technical University of Munich.
- Hazic Muhammad. "Human Performance on Cross-Resolution Face Verification". In: *Bachelor's Thesis*. 2022. Technical University of Munich.

- Gu Ying. “Image Resolution Impact on Activation Maps”. In: *Research Internship*. 2022. Technical University of Munich.
- Schwendtner Daniel. “Machine Learning-Based CI Test Results Analysis”. In: *Master’s Thesis*. 2023. BMW Group.
- Milosavljevic Luka. “Implementation and Evaluation of a Live Face Recognition System”. In: *Bachelor’s Thesis*. 2023. Technical University of Munich.
- Welling Cornelius Simon. “Transformation and Quantization of Machine Learning Models”. In: *Engineering Internship*. 2023. Technical University of Munich.

List of Acronyms

ADAGRAD Adaptive Gradient Algorithm

ADAM Adaptive Moment Estimation

ADAMW Adaptive Moment Estimation with Weight Decay

AgeDB Age Database

AI Artificial Intelligence

ANN Artificial Neural Network

BRISQUE Blind Referenceless Image Spatial Quality Evaluator

CALFW Cross-Age Labeled Faces in the Wild

CFP-FF Celebrities in Frontal-Profile — Frontal-Frontal

CFP-FP Celebrities in Frontal-Profile — Frontal-Profile

CLT Contrastive Loss Training

CNN Convolutional Neural Network

CPLFW Cross-Pose Labeled Faces in the Wild

CR Cross Resolution

C-Score Confidence Score

DL Deep Learning

EN European Norm

EER Equal Error Rate

ER Equal Resolution

FAR False Acceptance Rate

FIQ Face Image Quality

FP False Positive

FPR False Positive Rate

FR Face Recognition

FRR False Rejection Rate

FN False Negative

FV Face Verification

GAN Generative Adversarial Network

GFMT Glasgow Face Matching Test

HR High Resolution

IEEE Institute of Electrical and Electronics Engineers

IJB Intelligence Advanced Research Projects Activity Janus Benchmark

iResNet Improved Residual Layer Network

ISO International Organization for Standardization

LFW Labeled Faces in the Wild

LR Low Resolution

MB-CLT Multi-Branch Contrastive Loss Training

MLFW Masked Labeled Faces in the Wild

MS1M Microsoft 1 Million

MS1M-V2 Microsoft 1 Million Version 2

MTCNN Multi-Task Cascaded Convolutional Networks

OFMT Oxford Face Matching Test

OLT Octuplet Loss Training

PFE Probabilistic Face Embedding

RAT Resolution Augmentation Training

ResNet Residual Layer Network

RFW Racial Faces in the Wild

RGB Red Green Blue

RMSProp Root Mean Square Propagation

ROC Receiver Operating Characteristic

S-Map Similarity Map

SER-FIQ Stochastic Embedding Robustness Face Image Quality

SGD Stochastic Gradient Descent

SLLFW Similar-Looking Labeled Faces in the Wild

SR Super Resolution

SXQLFW Semi Cross-Quality Labeled Faces in the Wild

TALFW Transferable Adversarial Labeled Faces in the Wild

TAR True Acceptance Rate

TP True Positive

UCCS Unconstrained College Students

ViT Vision Transformer

XAI Explainable Artificial Intelligence

X-Map Explanation Map

XQLFW Cross-Quality Labeled Faces in the Wild

List of Symbols

Greek

α Margin for the contrastive loss, triplet loss, and octuplet loss

β Stride for the systemic image occlusion algorithm

γ Size of a square-shaped patch in horizontal and vertical direction

η Learning rate for training a model

θ Parameters of a model

σ Sigma value of a Gaussian kernel

Latin

$\boldsymbol{A}$ Anchor image

$\boldsymbol{B}$ Blended image

$\boldsymbol{D}$ Difference image between two images

$\boldsymbol{I}$ Image — typically a facial image

$\boldsymbol{M}$ Occlusion mask

$\boldsymbol{N}$ Negative image

$\boldsymbol{O}$ Occlusion map

$\boldsymbol{P}$ Positive image

$\boldsymbol{S}$ Similarity map

$\boldsymbol{X}$ X-Map

$\boldsymbol{f}$ Facial feature vector

$\boldsymbol{p}$ Probability vector

$\boldsymbol{v}$ Vector

$\boldsymbol{y}$ Output vector of a neural network

B Size of a batch of samples

C Confidence score of a model's decision

D Number of distances in a set of distances

F Dimension of a feature vector

K Number of channels in an image

L Number of classes in a classification task

M Size of an image in horizontal direction

N Size of an image in vertical direction

Q Quality score of an image

W Number of occlusion maps and masks

b Bias of the sigmoid function

d Distance between the camera and the face, or between feature vectors

e Shift of the sigmoid function

f Focal length of the camera in the pinhole camera model

h Height of a face

i Index variable

j Index variable

k Steepness of the sigmoid function

l Class label of an image

m Denominator of a method name

n Dimension of a vector

r Raw resolution of an image in both dimensions (horizontal and vertical)

s Scaling factor for downscaling and upscaling operations

t Threshold for the distance between feature vectors or for the confidence score of a model

v Maximum value of the sigmoid function

x Horizontal coordinate of a pixel in an image

y Vertical coordinate of a pixel in an image

$\mathcal{B}$ Batch of samples

$\mathcal{D}$ Set of feature distances

$\mathcal{F}$ Set of facial feature vectors

$\mathcal{M}$ Set of occlusion masks

$\mathcal{O}$ Set of occluded images

$\mathcal{P}$ Set of patches

$\mathcal{R}$ Set of image resolutions

$\mathcal{T}$ Set of triplet samples

$\mathcal{X}$ Set of images, also known as a dataset

$\mathbb{N}$ Natural numbers

$\mathbb{R}$ Rational numbers

$\mathscr{L}$ Loss function

List of Figures

List of Tables

References

[1†] Knoche Martin, Hörmann Stefan, Rigoll Gerhard. "Susceptibility to image resolution in face recognition and training strategies to enhance robustness". In: *Leibniz Transactions on Embedded Systems* 8.1 (2022), pp. 1–20.

[2†] Knoche Martin, Elkadeem Mohamed, Hörmann Stefan, Rigoll Gerhard. "Octuplet loss: Make face recognition robust to image resolution". In: *17th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2023, pp. 1–8.

[3†] Knoche Martin, Hörmann Stefan, Rigoll Gerhard. "Cross-Quality LFW: A database for analyzing cross-resolution image face recognition in unconstrained environments". In: *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2021, pp. 1–5.

[4†] Knoche Martin, Teepe Torben, Hörmann Stefan, Rigoll Gerhard. "Explainable model-agnostic similarity and confidence in face verification". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2023, pp. 711–718.

[5†] Knoche Martin, Rigoll Gerhard. "Tackling face verification edge cases: In-depth analysis and human-machine fusion approach". In: *Proceedings of the 18th International Conference on Machine Vision and Applications (MVA)*. IEEE. 2023, pp. 1–5.

[6†] Hörmann Stefan, Cao Zhenxiang, Knoche Martin, Herzog Fabian, Rigoll Gerhard. "Face aggregation network for video face recognition". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 2973–2977.

[7†] Hörmann Stefan, Xia Zhibing, Knoche Martin, Rigoll Gerhard. "A coarse-to-fine dual attention network for blind face completion". In: *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2021, pp. 01–08.

[8†] Hörmann Stefan, Kong Tianlin, Teepe Torben, Herzog Fabian, Knoche Martin, Rigoll Gerhard. "Face morphing: Fooling a face recognition system is simple!" In: arXiv (2022). arXiv: 2205.13796.

[9†] Hörmann Stefan, Moiz Abdul, Knoche Martin, Rigoll Gerhard. "Attention fusion for audio-visual person verification using multi-scale features". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2020, pp. 281–285.

[10†] Hörmann Stefan, Zhang Zeyuan, Knoche Martin, Teepe Torben, Rigoll Gerhard. "Attention-based partial face recognition". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 2978–2982.

[11†] Hörmann Stefan, Knoche Martin, Babaee Maryam, Köpüklü Okan, Rigoll Gerhard. "Outlier-robust neural aggregation network for video face identification". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 1675–1679.

[12†] Hörmann Stefan, Knoche Martin, Rigoll Gerhard. "A multi-task comparator framework for kinship verification". In: *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2020, pp. 863–867.

[13] Slater Alan, Quinn Paul C. "Face recognition in the newborn infant". In: *Infant and Child Development: An International Journal of Research and Practice* 10.1-2 (2001), pp. 21–24.

[14] Nelson Charles A. "The development and neural bases of face recognition". In: *Infant and Child Development: An International Journal of Research and Practice* 10.1-2 (2001), pp. 3–18.

[15] Carey Susan, Diamond Rhea, Woods Bryan. "Development of face recognition: A maturational component?" In: *Developmental Psychology* 16.4 (1980), p. 257.

[16] Lawrence Steve, Giles C. Lee, Tsoi Ah Chung, Back Andrew D. "Face recognition: A convolutional neural-network approach". In: *IEEE Transactions on Neural Networks* 8.1 (1997), pp. 98–113.

[17] Rowley Henry A., Baluja Shumeet, Kanade Takeo. "Neural network-based face detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1 (1998), pp. 23–38.

[18] Taigman Yaniv, Yang Ming, Ranzato Marc'Aurelio, Wolf Lior. "Deepface: Closing the gap to human-level performance in face verification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1701–1708.

[19] Sun Yi, Wang Xiaogang, Tang Xiaoou. "Deep learning face representation from predicting 10,000 classes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1891–1898.

[20] Lu Chaochao, Tang Xiaoou. "Surpassing human-level face verification performance on LFW with GaussianFace". In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. Vol. 29. 2015.

[21] O'Toole Alice J., Phillips P. Jonathon, Jiang Fang, Ayyad Janet, Penard Nils, Abdi Herve. "Face recognition algorithms surpass humans matching faces over changes in illumination". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.9 (2007), pp. 1642–1646.

[22] Feldstein Steven. *The global expansion of AI surveillance*. Vol. 17. Carnegie Endowment for International Peace, 2019, pp. 7–10.

[23] Martin Kirsten. *Ethics of data and analytics: Concepts and cases*. CRC Press, 2022, pp. 170–177.

[24] Smith Marcus, Miller Seumas. "The ethical application of biometric facial recognition technology". In: *Ai & Society* (2022), pp. 1–9.

[25] Phillips P. Jonathon, Yates Amy N., Hu Ying, Hahn Carina A., Noyes Eilidh, Jackson Kelsey, Cavazos Jacqueline G., Jeckeln Géraldine, Ranjan Rajeev, Sankaranarayanan Swami, Chen Jun-Cheng, Castillo Carlos D., Chellappa Rama, White David, O'Toole Alice J. "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms". In: *Proceedings of the National Academy of Sciences* 115.24 (2018), pp. 6171–6176.

[26] Fisher Caitlinrose. “Code of ethics for facial recognition”. In: *Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium* (2022).

[27] Berle Ian. “The future of face recognition technology and ethico: Legal issues”. In: (2020), pp. 163–184.

[28] Selinger Evan, Leong Brenda. “The ethics of facial recognition technology”. In: *SSRN Electronic Journal* (2021).

[29] Chochia Archil, Nässi Teele. “Ethics and emerging technologies - facial recognition”. In: *IDP Revista de Internet Derecho y Política* (2021).

[30] Ketley Isabella Tomaz. “Case study: Code of ethics for facial recognition technology”. In: *Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium* (2022).

[31] Huang Gary B., Ramesh Manu, Berg Tamara, Learned-Miller Erik. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, 2007.

[32] Standardization International Organization for. *ISO 80000-2: Quantities and units: Part 2: Mathematical signs and symbols to be used in the natural sciences and technology*. ISO, 2009. URL: https://books.google.de/books?id=Q6zJAQAACAAJ.

[33] Lanczos Cornelius. “An iteration method for the solution of the eigenvalue problem of linear differential and integral operators”. In: (1950).

[34] Lepcha Dawa Chyophel, Goyal Bhawna, Dogra Ayush, Goyal Vishal. “Image super-resolution: A comprehensive review, recent trends, challenges and applications”. In: *Information Fusion* 91 (2023), pp. 230–260.

[35] Li Juncheng, Pei Zehua, Zeng Tieyong. “From beginner to master: A survey for deep learning-based single-image super-resolution”. In: arXiv (2021). arXiv: 2109.14335.

[36] Mittal Anish, Moorthy Anush Krishna, Bovik Alan Conrad. “No-reference image quality assessment in the spatial domain”. In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[37] Kamble Vipin, Bhurchandi K. “No-reference image quality assessment algorithms: A survey”. In: *Optik - International Journal for Light and Electron Optics* 126 (May 2015), pp. 1090–1097.

[38] Terhörst Philipp, Kolf Jan Niklas, Damer Naser, Kirchbuchner Florian, Kuijper Arjan. “SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5651–5660.

[39] McCulloch Warren S., Pitts Walter. “A logical calculus of the ideas immanent in nervous activity”. In: *The Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–133.

[40] Sun Yi, Liang Ding, Wang Xiaogang, Tang Xiaoou. “DeepID3: Face recognition with very deep neural networks”. In: arXiv (2015). arXiv: 1502.00873.

[41] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, Houlsby Neil. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *19th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

[42] Chopra Sumit, Hadsell Raia, LeCun Yann. "Learning a similarity metric discriminatively, with application to face verification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. IEEE. 2005, pp. 539–546.

[43] Sun Yi, Chen Yuheng, Wang Xiaogang, Tang Xiaoou. "Deep learning face representation by joint identification-verification". In: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[44] Schroff Florian, Kalenichenko Dmitry, Philbin James. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823.

[45] Shrivastava Abhinav, Gupta Abhinav, Girshick Ross. "Training region-based object detectors with online hard example mining". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 761–769.

[46] Hermans Alexander, Beyer Lucas, Leibe Bastian. "In Defense of the Triplet Loss for Person Re-Identification". In: arXiv (2017). arXiv: 1703.07737.

[47] Deng Jiankang, Guo Jia, Xue Niannan, Zafeiriou Stefanos. "Arcface: Additive angular margin loss for deep face recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4690–4699.

[48] Kelley Henry J. "Gradient theory of optimal flight paths". In: *Ars Journal* 30.10 (1960), pp. 947–954.

[49] Bryson Arthur E. "A gradient method for optimizing multi-stage allocation processes". In: *Proc. Harvard Univ. Symposium on Digital Computers and their Applications*. Vol. 72. 1961, p. 22.

[50] Werbos Paul J. "Applications of advances in nonlinear sensitivity analysis". In: *Proceedings of the 10th IFIP Conference on System Modeling and Optimization*. Springer. 1981, pp. 762–770.

[51] Kingma Diederik P., Ba Jimmy. "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations (ICLR)*. 2015.

[52] Duchi John, Hazan Elad, Singer Yoram. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.7 (2011).

[53] Hinton Geoffrey, Srivastava Nitish, Swersky Kevin. "Neural networks for machine learning lecture 6a: Overview of mini-batch gradient descent". In: *Lecture Notes by Geoffrey Hinton, University of Toronto* 14.8 (2012), p. 2.

[54] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[55] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. “Identity mappings in deep residual networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 630–645.

[56] Wang Hao, Wang Yitong, Zhou Zheng, Ji Xing, Gong Dihong, Zhou Jingchao, Li Zhifeng, Liu Wei. “Cosface: Large margin cosine loss for deep face recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 5265–5274.

[57] Kim Yonghyun, Park Wonpyo, Roh Myung-Cheol, Shin Jongju. “GroupFace: Learning latent groups and constructing group-based representations for face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5621–5630.

[58] Su Weicong, Wang Yali, Li Kunchang, Gao Peng, Qiao Yu. “Hybrid token transformer for deep face recognition”. In: *Pattern Recognition* 139 (2023), p. 109443.

[59] Sun Zhonglin, Tzimiropoulos Georgios. “Part-based face recognition with vision transformers”. In: *Proceedings of the 33rd British Machine Vision Conference (BMVC)*. BMVA Press, 2022, p. 611.

[60] He Lin, He Lile, Peng Lijun. “CFormerFaceNet: Efficient lightweight network merging a CNN and transformer for face recognition”. In: *Applied Sciences* 13.11 (2023), p. 6506.

[61] Islam Khawar, Zaheer Muhammad Z., Mahmood Arif. “Face pyramid vision transformer”. In: *Proceedings of the 33rd British Machine Vision Conference (BMVC)*. BMVA Press, 2022, p. 758.

[62] Guo Yandong, Zhang Lei, Hu Yuxiao, He Xiaodong, Gao Jianfeng. “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 87–102.

[63] Zhang Hongxin, Chi Liying. “End-to-end spatial transform face detection and recognition”. In: *Virtual Reality & Intelligent Hardware* 2.2 (2020), pp. 119–131.

[64] Zhong Yuanyi, Chen Jiansheng, Huang Bo. “Toward end-to-end face recognition through alignment learning”. In: *IEEE Signal Processing Letters* 24.8 (2017), pp. 1213–1217.

[65] Chi Liying, Zhang Hongxin, Chen Mingxiu. “End-to-end face detection and recognition”. In: arXiv (2017). arXiv: 1703.10818.

[66] Huang Yuge, Wang Yuhan, Tai Ying, Liu Xiaoming, Shen Pengcheng, Li Shaoxin, Li Jilin, Huang Feiyue. “Curricularface: adaptive curriculum learning loss for deep face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5901–5910.

[67] Kim Yonghyun, Park Wonpyo, Shin Jongju. “BroadFace: Looking at tens of thousands of people at once for face recognition”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 536–552.

[68] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, Qiao Yu. “Joint face detection and alignment using multitask cascaded convolutional networks”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503.

[69] Hörmann Stefan. “Robust face recognition under adverse conditions”. PhD thesis. Technical University of Munich, 2023.

[70] Wen Yandong, Zhang Kaipeng, Li Zhifeng, Qiao Yu. "A discriminative feature learning approach for deep face recognition". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 499–515.

[71] Bansal Ankan, Castillo Carlos, Ranjan Rajeev, Chellappa Rama. "The do's and don'ts for cnn-based face verification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017, pp. 2545–2554.

[72] Zhou Erjin, Cao Zhimin, Yin Qi. "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" In: arXiv (2015). arXiv: 1501.04690.

[73] Zhang Yaobin, Deng Weihong. "Class-balanced training for deep face recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 824–825.

[74] Zhu Zheng, Huang Guan, Deng Jiankang, Ye Yun, Huang Junjie, Chen Xinze, Zhu Jiagang, Yang Tian, Lu Jiwen, Du Dalong, Zhou Jie. "Webface260m: A benchmark unveiling the power of million-scale deep face recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 10492–10502.

[75] Yi Dong, Lei Zhen, Liao Shengcai, Li Stan Z. "Learning face representation from scratch". In: arXiv (2014). arXiv: 1411.7923.

[76] Bansal Ankan, Nanduri Anirudh, Castillo Carlos D., Ranjan Rajeev, Chellappa Rama. "Umdfaces: An annotated face dataset for training deep networks". In: *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2017, pp. 464–473.

[77] Parkhi Omkar, Vedaldi Andrea, Zisserman Andrew. "Deep face recognition". In: *Proceedings of the 26th British Machine Vision Conference (BMVC)*. British Machine Vision Association. 2015.

[78] Cao Qiong, Shen Li, Xie Weidi, Parkhi Omkar M., Zisserman Andrew. "Vggface2: A dataset for recognising faces across pose and age". In: *18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2018, pp. 67–74.

[79] Deng Jiankang, Zhou Yuxiang, Zafeiriou Stefanos. "Marginal loss for deep face recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017, pp. 60–68.

[80] *Trillion pairs*. (accessed Feb. 12, 2024). 2019. URL: http://trillionpairs.deepglint.com/.

[81] Nech Aaron, Kemelmacher-Shlizerman Ira. "Level playing field for million scale face recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7044–7053.

[82] An Xiang, Zhu Xuhan, Gao Yuan, Xiao Yang, Zhao Yongle, Feng Ziyong, Wu Lan, Qin Bin, Zhang Ming, Zhang Debing, Fu Ying. "Partial fc: Training 10 million identities on a single machine". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 1445–1449.

[83] Wang Fei, Chen Liren, Li Cheng, Huang Shiyao, Chen Yanjie, Qian Chen, Loy Chen Change. "The devil of face recognition is in the noise". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 765–780.

[84] Taigman Yaniv, Yang Ming, Ranzato Marc'Aurelio, Wolf Lior. "Web-scale training for face identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 2746–2754.

[85] Zhang Yaobin, Deng Weihong, Wang Mei, Hu Jiani, Li Xian, Zhao Dongyue, Wen Dongchao. "Global-local gcn: Large-scale label noise cleansing for face recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7731–7740.

[86] Hupont Isabelle, Fernández Carles. "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition". In: *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2019, pp. 1–7.

[87] Wang Mei, Deng Weihong, Hu Jiani, Tao Xunqiang, Huang Yaohai. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 692–702.

[88] Jin Chi, Jin Ruochun, Chen Kai, Dou Yong. "A community detection approach to cleaning extremely large face database". In: *Computational Intelligence and Neuroscience* 2018 (2018).

[89] Meng Qiang, Zhao Shichao, Huang Zhida, Zhou Feng. "Magface: A universal representation for face recognition and quality assessment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14225–14234.

[90] Sun Yifan, Cheng Changmao, Zhang Yuhan, Zhang Chi, Zheng Liang, Wang Zhongdao, Wei Yichen. "Circle loss: A unified perspective of pair similarity optimization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6398–6407.

[91] Zhong Yaoyao, Deng Weihong. "Face transformer for recognition". In: arXiv (2021). arXiv: 2103.14803.

[92] Zheng Tianyue, Deng Weihong. *Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments*. Tech. rep. 18-01. Beijing University of Posts and Telecommunications, 2018.

[93] Sengupta Soumyadip, Chen Jun-Cheng, Castillo Carlos, Patel Vishal M., Chellappa Rama, Jacobs David W. "Frontal to profile face verification in the wild". In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*. IEEE. 2016, pp. 1–9.

[94] Zheng Tianyue, Deng Weihong, Hu Jiani. "Cross-Age LFW: A Database for studying cross-age face recognition in unconstrained environments". In: arXiv (2017). arXiv: 1708.08197.

[95] Moschoglou Stylianos, Papaioannou Athanasios, Sagonas Christos, Deng Jiankang, Kotsia Irene, Zafeiriou Stefanos. "AgeDB: The first manually collected, in-the-wild age database". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017, pp. 1997–2005.

[96] Cheng Zhiyi, Zhu Xiatian, Gong Shaogang. "Surveillance face recognition challenge". In: arXiv (2018). arXiv: 1804.09691.

[97] Deng Weihong, Hu Jiani, Zhang Nanhai, Chen Binghui, Guo Jun. "Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership". In: *Pattern Recognition* 66 (2017), pp. 63–73.

[98] Zhong Yaoyao, Deng Weihong. "Towards transferable adversarial attack against deep face recognition". In: *IEEE Transactions on Information Forensics and Security* (2020).

[99] Wang Zhongyuan, Huang Baojin, Wang Guangcheng, Yi Peng, Jiang Kui. "Masked face recognition dataset and application". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 5.2 (2023), pp. 298–304.

[100] Kemelmacher-Shlizerman Ira, Seitz Steven M, Miller Daniel, Brossard Evan. "The megaface benchmark: 1 million faces for recognition at scale". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4873–4882.

[101] Klare Brendan F., Klein Ben, Taborsky Emma, Blanton Austin, Cheney Jordan, Allen Kristen, Grother Patrick, Mah Alan, Jain Anil K. "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1931–1939.

[102] Whitelam Cameron, Taborsky Emma, Blanton Austin, Maze Brianna, Adams Jocelyn, Miller Tim, Kalka Nathan, Jain Anil K., Duncan James A., Allen Kristen, Cheney Jordan, Grother Patrick. "Iarpa janus benchmark-b face dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017, pp. 90–98.

[103] Maze Brianna, Adams Jocelyn, Duncan James A., Kalka Nathan, Miller Tim, Otto Charles, Jain Anil K., Niggel W. Tyler, Anderson Janet, Cheney Jordan, Grother Patrick. "Iarpa janus benchmark-c: Face dataset and protocol". In: *IEEE International Conference on Biometrics (ICB)*. IEEE. 2018, pp. 158–165.

[104] Wolf Lior, Hassner Tal, Maoz Itay. "Face recognition in unconstrained videos with matched background similarity". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2011, pp. 529–534.

[105] Liu Weiyang, Wen Yandong, Yu Zhiding, Li Ming, Raj Bhiksha, Song Le. "Sphereface: Deep hypersphere embedding for face recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 212–220.

[106] Deng Jiankang, Guo Jia, Yang Jing, Lattas Alexandros, Zafeiriou Stefanos. "Variational prototype learning for deep face recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11906–11915.

[107] Chrysos Grigorios G, Moschoglou Stylianos, Bouritsas Giorgos, Panagakis Yannis, Deng Jiankang, Zafeiriou Stefanos. "P-Nets: Deep polynomial neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7325–7335.

[108] Qin Lixiong, Wang Mei, Deng Chao, Wang Ke, Chen Xi, Hu Jiani, Deng Weihong. "SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023), p. 1.

[109] Dan Jun, Liu Yang, Xie Haoyu, Deng Jiankang, Xie Haoran, Xie Xuansong, Sun Baigui. "Trans-Face: Calibrating transformer training for face recognition from a data-centric perspective". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 20642–20653.

[110] Kumar Neeraj, Berg Alexander C., Belhumeur Peter N., Nayar Shree K. "Attribute and simile classifiers for face verification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2009, pp. 365–372.

[111] Kim Geunsu, Park Gyudo, Kang Soohyeok, Woo Simon S. "S-ViT: Sparse vision transformer for accurate face recognition". In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. 2023, pp. 1130–1138.

[112] Boom Bastiaan J., Beumer GM., Spreeuwers Luuk J., Veldhuis Raymond NJ. "The effect of image resolution on the performance of a face recognition system". In: *9th International Conference on Control, Automation, Robotics and Vision*. IEEE. 2006, pp. 1–6.

[113] Li Pei, Flynn Patrick J., Prieto Loreto, Mery Domingo. "Face recognition in low quality images: A survey". In: *ACM Computing Surveys* 1.1 (2019).

[114] Li Pei, Prieto Loreto, Mery Domingo, Flynn Patrick J. "On low-resolution face recognition in the wild: Comparisons and new techniques". In: *IEEE Transactions on Information Forensics and Security* 14.8 (2019), pp. 2000–2012.

[115] Marciniak Tomasz, Chmielewska Agata, Weychan Radoslaw, Parzych Marianna, Dabrowski Adam. "Influence of low resolution of images on reliability of face detection and recognition". In: *Multimedia Tools and Applications* 74 (2015), pp. 4329–4349.

[116] Abadi Mart'in, Agarwal Ashish, Barham Paul, Brevdo Eugene, Chen Zhifeng, Citro Craig, Corrado Greg S., Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Goodfellow Ian, Harp Andrew, Irving Geoffrey, Isard Michael, Jia Yangqing, Jozefowicz Rafal, Kaiser Lukasz, Kudlur Manjunath, Levenberg Josh, Man'e Dandelion, Monga Rajat, Moore Sherry, Murray Derek, Olah Chris, Schuster Mike, Shlens Jonathon, Steiner Benoit, Sutskever Ilya, Talwar Kunal, Tucker Paul, Vanhoucke Vincent, Vasudevan Vijay, Vi'egas Fernanda, Vinyals Oriol, Warden Pete, Wattenberg Martin, Wicke Martin, Yu Yuan, Zheng Xiaoqiang. *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[117] Norm Standard European. *DIN EN 62676-4: 2016 Video surveillance systems for use in security applications - Part 4: Application guidelines*. EN, 2015. URL: https://www.vde-verlag.de/normen/0800325/din-en-62676-4-vde-0830-71-4-2016-07.html.

[118] Rosenberg Louis B. "The effect of interocular distance upon operator performance using stereoscopic displays to perform virtual depth tasks". In: *Proceedings of IEEE Virtual Reality Annual International Symposium*. IEEE. 1993, pp. 27–32.

[119] Hayashi Shinji, Hasegawa Osamu. "Face detection in low-resolution images". In: *International Symposium on Visual Computing*. Springer. 2005, pp. 199–206.

[120] Singh Maneet, Nagpal Shruti, Singh Richa, Vatsa Mayank, Majumdar Angshul. "MagnifyMe: Aiding cross resolution face recognition via identity aware synthesis". In: arXiv (2018). arXiv: 1802.08057.

[121] Wang Zhifei, Miao Zhenjiang, Wu QM. Jonathan, Wan Yanli, Tang Zhen. "Low-resolution face recognition: A review". In: *The Visual Computer* 30.4 (2014), pp. 359–386.

[122] Jiang Junjun, Wang Chenyang, Liu Xianming, Ma Jiayi. "Deep learning-based face super-resolution: A survey". In: *ACM Computing Surveys (CSUR)* 55.1 (2021), pp. 1–36.

[123] Kalarot Ratheesh, Li Tao, Porikli Fatih. "Component attention guided face super-resolution network: Cagface". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 370–380.

[124] Wang Chenyang, Jiang Junjun, Liu Xianming. "Heatmap-aware pyramid face hallucination". In: *IEEE International Conference on Multimedia and Expo (ICME)*. 2021, pp. 1–6.

[125] Li Jichun, Bare Bahetiyaer, Zhou Shili, Yan Bo, Li Ke. "Organ-branched CNN for robust face super-resolution". In: *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.

[126] Wang Huan, Hu Qian, Wu Chengdong, Chi Jianning, Yu Xiaosheng, Wu Hao. "Dclnet: Dual closed-loop networks for face super-resolution". In: *Knowledge-Based Systems* 222 (2021).

[127] Liu Shuang, Xiong Chengyi, Gao Zhirong. "Face super-resolution network with incremental enhancement of facial parsing information". In: *25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 7537–7543.

[128] Lu Yongyi, Tai Yu-Wing, Tang Chi-Keung. "Attribute-guided face generation using conditional cyclegan". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 282–297.

[129] Yu Xin, Fernando Basura, Hartley Richard, Porikli Fatih. "Super-resolving very low-resolution face images with supplementary attributes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 908–917.

[130] Yu Xin, Fernando Basura, Hartley Richard, Porikli Fatih. "Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.11 (2019), pp. 2926–2943.

[131] Xin Jingwei, Wang Nannan, Jiang Xinrui, Li Jie, Gao Xinbo, Li Zhifeng. "Facial attribute capsules for noise face super resolution". In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 12476–12483.

[132] Zhang Kaipeng, Zhang Zhanpeng, Cheng Chia-Wen, Hsu Winston H., Qiao Yu, Liu Wei, Zhang Tong. "Super-identity convolutional neural network for face hallucination". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 183–198.

[133] Kazemi Hadi, Taherkhani Fariborz, Nasser M. Nasrabadi. "Identity-aware deep face hallucination via adversarial face verification". In: *IEEE International Conference on Biometrics Theory Applications and Systems*. 2019.

[134] Dogan Berk, Gu Shuhang, Timofte Radu. “Exemplar guided face image super-resolution without facial landmarks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019, pp. 1814–1823.

[135] Bayramli Bayram, Ali Usman, Qi Te, Lu Hongtao. “FH-GAN: Face hallucination and recognition using generative adversarial network”. In: *International Conference on Neural Information Processing*. Springer. 2019, pp. 3–15.

[136] Huang Huaibo, He Ran, Sun Zhenan, Tan Tieniu. “Wavelet domain generative adversarial network for multi-scale face hallucination”. In: *International Journal of Computer Vision* 127.6 (2019), pp. 763–784.

[137] Lai Shun-Cheung, He Chen-Hang, Lam Kin-Man. “Low-resolution face recognition based on identity-preserved face hallucination”. In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 1173–1177.

[138] Abello Antonio Augusto, Hirata Roberto. “Optimizing super resolution for face recognition”. In: *32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE. 2019, pp. 194–201.

[139] Grm Klemen, Scheirer Walter J., Štruc Vitomir. “Face hallucination using cascaded super-resolution and identity priors”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 2150–2165.

[140] Cheng Fangfang, Lu Tao, Wang Yu, Zhang Yanduo. “Face super-resolution through dual-identity constraint”. In: *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.

[141] Kim Jonghyun, Li Gen, Yun Inyong, Jung Cheolkon, Kim Joongkyu. “Edge and identity preserving network for face super-resolution”. In: *Neurocomputing* 446 (2021), pp. 11–22.

[142] Cansizoglu Esra A., Jones Michael, Zhang Ziming, Sullivan Alan. “Verification of very low-resolution faces using an identity-preserving deep face super-resolution network”. In: arXiv (2019). arXiv: 1903.10974.

[143] Li Mengyan, Zhang Zhaoyu, Yu Jun, Chen Chang Wen. “Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement”. In: *Transactions on Multimedia* 23 (2020), pp. 468–483.

[144] Cheng Xiaojuan, Lu Jiwen, Yuan Bo, Zhou Jie. “Identity-preserving face hallucination via deep reinforcement learning”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.12 (2019), pp. 4796–4809.

[145] Zangeneh Erfan, Rahmati Mohammad, Mohsenzadeh Yalda. “Low resolution face recognition using a two-branch deep convolutional neural network architecture”. In: *Expert Systems with Applications* 139 (2020).

[146] Khazaie Vahid R., Bayat Nicky, Mohsenzadeh Yalda. “IPU-Net: Multi scale identity-preserved U-Net for low resolution face recognition”. In: arXiv (2020). arXiv: 2010.12249.

[147] Ronneberger Olaf, Fischer Philipp, Brox Thomas. "U-net: Convolutional networks for biomedical image segmentation". In: *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241.

[148] Szegedy Christian, Liu Wei, Jia Yangqing, Sermanet Pierre, Reed Scott, Anguelov Dragomir, Erhan Dumitru, Vanhoucke Vincent, Rabinovich Andrew. "Going deeper with convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[149] Hsu Chih-Chung, Lin Chia-Wen, Su Weng-Tai, Cheung Gene. "Sigan: Siamese generative adversarial network for identity-preserving face hallucination". In: *IEEE Transactions on Image Processing* 28.12 (2019), pp. 6225–6236.

[150] Kazemi Hadi, Taherkhani Fariborz, Nasrabadi Nasser M. "Identity-aware deep face hallucination via adversarial face verification". In: *10th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE. 2019, pp. 1–10.

[151] Ghosh Soumyadeep, Vatsa Mayank, Singh Richa. "SUPREAR-NET: Supervised resolution enhancement and recognition network". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2022).

[152] Lu Ze, Jiang Xudong, Kot Alex. "Deep coupled resnet for low-resolution face recognition". In: *IEEE Signal Processing Letters* 25.4 (2018), pp. 526–530.

[153] Zeng Dan, Chen Hu, Zhao Qijun. "Towards resolution invariant face recognition in uncontrolled scenarios". In: *IEEE International Conference on Biometrics (ICB)*. IEEE. 2016, pp. 1–8.

[154] Massoli Fabio Valerio, Amato Giuseppe, Falchi Fabrizio. "Cross-resolution learning for face recognition". In: *Image and Vision Computing* (2020).

[155] Talreja Veeru, Taherkhani Fariborz, Valenti Matthew C., Nasrabadi Nasser M. "Attribute-guided coupled GAN for cross-resolution face recognition". In: *10th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–10.

[156] Ge Shiming, Zhao Shengwei, Li Chenyu, Li Jia. "Low-resolution face recognition in the wild via selective knowledge distillation". In: *IEEE Transactions on Image Processing* 28.4 (2018), pp. 2051–2062.

[157] Sun Jingna, Shen Yehu, Yang Wenming, Liao Qingmin. "Classifier shared deep network with multi-hierarchy loss for low resolution face recognition". In: *Signal Processing: Image Communication* 82 (2020), p. 115766.

[158] Mudunuri Sivaram Prasad, Sanyal Soubhik, Biswas Soma. "GenLR-Net: Deep framework for very low resolution face and object recognition with generalization to unseen categories". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE. 2018, pp. 602–60209.

[159] Lai Shun-Cheung, Lam Kin-Man. "Deep Siamese network for low-resolution face recognition". In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2021, pp. 1444–1449.

[160] Zha Juan, Chao Hongyang. “TCN: Transferable coupled network for cross-resolution face recognition”. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3302–3306.

[161] Mishra Nayaneesh Kumar, Dutta Mainak, Singh Satish Kumar. “Multiscale parallel deep CNN (mpdCNN) architecture for the real low-resolution face recognition for surveillance”. In: *Image and Vision Computing* 115 (2021), p. 104290.

[162] Li Peiying, Tu Shikui, Xu Lei. “Deep rival penalized competitive learning for low-resolution face recognition”. In: *Neural Networks* (2022), pp. 183–193.

[163] Terhörst Philipp, Ihlefeld Malte, Huber Marco, Damer Naser, Kirchbuchner Florian, Raja Kiran, Kuijper Arjan. *QMagFace: Simple and accurate quality-aware face recognition*. 2021. arXiv: 2111.13475.

[164] Zhao Xuan. *Homogeneous low-resolution face recognition method based correlation features*. 2021. arXiv: 2111.13175.

[165] Tang Su, Zhou Shan, Kang Wenxiong, Wu Qiuxia, Deng Feiqi. “Finger vein verification using a siamese CNN”. In: *IET Biometrics* 8.5 (2019), pp. 306–315.

[166] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander C., Fei-Fei Li. “Imagenet large scale visual recognition challenge”. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[167] Gómez-Silva María José, Armingol Jose M., Escalera Arturo de la. “Triplet permutation method for deep learning of single-shot person re-identification”. In: *9th International Conference on Imaging for Crime Detection and Prevention (ICDP)*. IET. 2019, pp. 56–61.

[168] Kaya Mahmut, Bilge Hasan Şakir. “Deep metric learning: A survey”. In: *Symmetry* 11.9 (2019), p. 1066.

[169] Duta Ionut Cosmin, Liu Li, Zhu Fan, Shao Ling. “Improved residual networks for image and video recognition”. In: *25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9415–9422.

[170] Sandler Mark, Howard Andrew, Zhu Menglong, Zhmoginov Andrey, Chen Liang-Chieh. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520.

[171] Loshchilov Ilya, Hutter Frank. “Decoupled weight decay regularization”. In: *7th International Conference on Learning Representations (ICLR) (Poster)*. 2019.

[172] Boutros Fadi, Damer Naser, Kirchbuchner Florian, Kuijper Arjan. “Self-restrained triplet loss for accurate masked face recognition”. In: *Pattern Recognition* 124 (2022), p. 108473.

[173] Feng Yushu, Wang Huan, Hu Haoji Roland, Yu Lu, Wang Wei, Wang Shiyan. “Triplet distillation for deep face recognition”. In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 808–812.

[174] Lu Ze, Jiang Xudong, Kot Alex. “Deep coupled ResNet for low-resolution face recognition”. In: *IEEE Signal Processing Letters* 25.4 (2018), pp. 526–530.

[175] Zeng Dan, Chen Hu, Zhao Qijun. “Towards resolution invariant face recognition in uncontrolled scenarios”. In: *IEEE International Conference on Biometrics (ICB)*. 2016, pp. 1–8.

[176] Ji Xiaozhong, Cao Yun, Tai Ying, Wang Chengjie, Li Jilin, Huang Feiyue. “Real-world super-resolution via kernel estimation and noise injection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 466–467.

[177] Zhang Kai, Zuo Wangmeng, Zhang Lei. “Deep plug-and-play super-resolution for arbitrary blur kernels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1671–1681.

[178] Zhou Ruofan, Susstrunk Sabine. “Kernel modeling super-resolution on real low-resolution images”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2433–2443.

[179] Bell-Kligler Sefi, Shocher Assaf, Irani Michal. “Blind super-resolution kernel estimation using an internal-GAN”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 284–293.

[180] Grgic Mislav, Delac Kresimir, Grgic Sonja. “SCface - surveillance cameras face database”. In: *Multimedia Tools and Applications* 51.3 (2011), pp. 863–879.

[181] Huang Zhiwu, Shan Shiguang, Zhang Haihong, Lao Shihong, Kuerban Alifu, Chen Xilin. “Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset”. In: *11th Asian Conference on Computer Vision (ACCV)*. Springer. 2012, pp. 589–600.

[182] Wong Yongkang, Chen Shaokang, Mau Sandra, Sanderson Conrad, Lovell Brian C. “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2011, pp. 81–88.

[183] Cheng Zhiyi, Zhu Xiatian, Gong Shaogang. “Low-resolution face recognition”. In: *14th Asian Conference on Computer Vision (ACCV)*. Springer. 2018, pp. 605–621.

[184] Sapkota Archana, Boult T. “Large scale unconstrained open set face database”. In: *6th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2013), pp. 1–8.

[185] Khryashchev Vladimir, Nenakhov Ilya, Lebedev Anton, Priorov Andrey. “Evaluation of face image quality metrics in person identification problem”. In: *19th IEEE Conference of Open Innovations Association FRUCT*. Vol. 420. Nov. 2016, pp. 80–87.

[186] Yang Fei, Shao Xiaohu, Zhang Lijun, Deng Pingling, Zhou Xiangdong, Shi Yu. “DFQA: Deep face image quality assessment”. In: *International Conference on Image and Graphics*. Springer. 2019, pp. 655–667.

[187] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, Polosukhin Illia. “Attention is all you need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008.

[188] Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos. "" 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*. 2016, pp. 1135–1144.

[189] Mery Domingo, Morris Bernardita. "On black-box explanation for face verification". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 3418–3427.

[190] Mery Domingo. "True black-box explanation in facial analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1596–1605.

[191] Lin Yu-Sheng, Liu Zhe-Yu, Chen Yu-An, Wang Yu-Siang, Chang Ya-Liang, Hsu Winston H. "xCos: An explainable cosine metric for face verification task". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.3s (2021), pp. 1–16.

[192] Cao Chunshui, Liu Xianming, Yang Yi, Yu Yinan, Wang Jiang, Wang Zilei, Huang Yongzhen, Wang Liang, Huang Chang, Xu Wei, Ramanan Deva, Huang Thomas S. "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2956–2964.

[193] Selvaraju Ramprasaath R., Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, Batra Dhruv. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[194] Chattopadhay Aditya, Sarkar Anirban, Howlader Prantik, Balasubramanian Vineeth N. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 839–847.

[195] Draelos Rachel Lea, Carin Lawrence. "HiResCAM: Explainable multi-organ multi-abnormality prediction in 3D medical images". In: arXiv (2020). arXiv: 2011.08891.

[196] Ramaswamy Harish G., Desai Saurabh. "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 983–991.

[197] Wang Haofan, Wang Zifan, Du Mengnan, Yang Fan, Zhang Zijian, Ding Sirui, Mardziel Piotr, Hu Xia. "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2020, pp. 24–25.

[198] Fu Ruigang, Hu Qingyong, Dong Xiaohu, Guo Yulan, Gao Yinghui, Li Biao. "Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs". In: *Proceedings of the 31st British Machine Vision Conference (BMVC)*. BMVA Press, 2020.

[199] Li Kunpeng, Wu Ziyan, Peng Kuan-Chuan, Ernst Jan, Fu Yun. "Tell me where to look: Guided attention inference network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 9215–9223.

[200] Dabkowski Piotr, Gal Yarin. “Real time image saliency for black box classifiers”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[201] Khakzar Ashkan, Baselizadeh Soroosh, Khanduja Saurabh, Kim Seong Tae, Navab Nassir. “Improving feature attribution through input-specific network pruning”. In: arXiv (2019). arXiv: 1911.11081.

[202] Shi Yichun, Jain Anil K. “Probabilistic face embeddings”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6902–6911.

[203] Chang Jie, Lan Zhonghao, Cheng Changmao, Wei Yichen. “Data uncertainty learning in face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5710–5719.

[204] Schlett Torsten, Rathgeb Christian, Henniger Olaf, Galbally Javier, Fierrez Julian, Busch Christoph. “Face image quality assessment: A literature survey”. In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–49.

[205] Huber Marco, Terhörst Philipp, Kirchbuchner Florian, Damer Naser, Kuijper Arjan. “Stating comparison score uncertainty and verification decision confidence towards transparent face recognition”. In: *Proceedings of the 33rd British Machine Vision Conference (BMVC)*. BMVA Press, 2022, p. 506.

[206] Neto Pedro C., Sequeira Ana Filipa, Cardoso Jaime S., Terhörst Philipp. “PIC-Score: Probabilistic interpretable comparison score for optimal matching confidence in single- and multi-biometric face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2023, pp. 1021–1029.

[207] Voglis C., Lagaris IE. “A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization”. In: *WSEAS International Conference on Applied Mathematics*. Vol. 7. 2004.

[208] Grinberg Miguel. *Flask web development: developing web applications with python*. "O’Reilly Media, Inc.", 2018.

[209] Howard John J., Rabbitt Laura R., Sirotin Yevgeniy B. “Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making”. In: *Plos One* 15.8 (2020), e0237855.

[210] Fysh Matthew C., Bindemann Markus. “Human-computer interaction in face matching”. In: *Cognitive Science* 42.5 (2018), pp. 1714–1732.

[211] Sandford Adam, Ritchie Kay L. “Unfamiliar face matching, within-person variability, and multiple-image arrays”. In: *Visual Cognition* 29.3 (2021), pp. 143–157.

[212] Burton A. Mike, White David, McNeill Allan. “The Glasgow face matching test”. In: *Behavior Research Methods* 42.1 (2010), pp. 286–291.

[213] White David, Guilbert Daniel, Varela Victor PL., Jenkins Rob, Burton A. Mike. “GFMT2: A psychometric measure of face matching ability”. In: *Behavior Research Methods* 54.1 (2022), pp. 252–260.

[214] Stantic Mirta, Brewer Rebecca, Duchaine Bradley, Banissy Michael J., Bate Sarah, Susilo Tirta, Catmur Caroline, Bird Geoffrey. "The Oxford face matching test: A non-biased test of the full range of individual differences in face perception". In: *Behavior Research Methods* 54.1 (2022), pp. 158–173.

[215] Duchaine Brad, Nakayama Ken. "The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants". In: *Neuropsychologia* 44.4 (2006), pp. 576–585.

[216] Abudarham Naphtali, Grosbard Idan, Yovel Galit. "Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition". In: *Cognitive Science* 45.9 (2021), e13031.

[217] Zheng Xueyi, Wang Ruixuan, Zhang Xinke, Sun Yan, Zhang Haohuan, Zhao Zihan, Zheng Yuanhang, Luo Jing, Zhang Jiangyu, Wu Hongmei, Huang Dan, Zhu Wenbiao, Chen Jianning, Cao Qinghua, Zeng Hong, Luo Rongzhen, Li Peng, Lan Lilong, Yun Jingping, Xie Dan, Zheng Wei-Shi, Luo Junhang, Cai Muyan. "A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology". In: *Nature Communications* 13.1 (2022), p. 2790.

[218] Carragher Daniel J., Hancock Peter JB. "Simulated automated facial recognition systems as decision-aids in forensic face matching tasks". In: *Journal of Experimental Psychology: General* (2022), p. 1286.

[219] Wang Mei, Deng Weihong, Hu Jiani, Tao Xunqiang, Huang Yaohai. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 692–702.

[220] Phillips P. Jonathon, O'toole Alice J. "Comparison of human and computer performance across face recognition experiments". In: *Image and Vision Computing* 32.1 (2014), pp. 74–85.

[221] Johnston Robert A., Edmonds Andrew J. "Familiar and unfamiliar face recognition: A review". In: *Memory* 17.5 (2009), pp. 577–596.

[222] Theis Sabine, Jentzsch Sophie, Deligiannaki Fotini, Berro Charles, Raulf Arne Peter, Bruder Carmen. "Requirements for explainability and acceptance of artificial intelligence in collaborative work". In: *International Conference on Human-Computer Interaction*. Springer. 2023, pp. 355–380.

[223] Dooley Samuel, Sukthanker Rhea, Dickerson John, White Colin, Hutter Frank, Goldblum Micah. "Rethinking bias mitigation: Fairer architectures make for fairer face recognition". In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2024).

[224] Wood Erroll, Baltrušaitis Tadas, Hewitt Charlie, Dziadzio Sebastian, Cashman Thomas J., Shotton Jamie. "Fake it till you make it: Face analysis in the wild using synthetic data alone". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 3681–3691.

[225] *IEEE editorial style manual for authors*. (Accessed on Jan. 4, 2024). 2023. URL: https://journals.ieeeauthorcenter.ieee.org/wp-content/uploads/sites/7/IEEE-Editorial-Style-Manual-for-Authors.pdf.