

Dissertation

Learning Transferable Descriptive Disentangled Representations

Mohammad Hasan Sarhan





Technische Universität München
TUM School of Computation, Information and Technology

Learning Transferable Descriptive Disentangled Representations

Mohammad Hasan Sarhan

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. habil. Alois Christian Knoll

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab
2. Prof. Dr. Bernhard Kainz

Die Dissertation wurde am 22.05.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 28.11.2024 angenommen.

Mohammad Hasan Sarhan

Learning Transferable Descriptive Disentangled Representations

Dissertation

Technische Universität München

TUM School of Computation, Information and Technology

Lehrstuhl für Informatikanwendungen in der Medizin

Boltzmannstraße 3

85748 Garching bei München

Abstract

Successful machine learning algorithms highly rely on good data representations. These representations can be designed with domain knowledge to suit a specific application. The preprocessing design step is a crucial part of machine learning models. Although this step deploys human ingenuity into the process of feature extraction, it is labor-intensive and does not provide an automatic mean for extracting informative discriminative features from data. Having machine learning algorithms that require less feature engineering is a highly desirable property, especially for applying machine learning to novel or rapidly changing applications. With modern deep neural networks, representations that capture factors of variation in the data could be learned automatically and used for a plethora of applications. However, such representations may entangle or hide some factors of variations which hinders the expressiveness and abstraction of the representation.

There are multiple characteristics desired in learned representations that could lead to better performance on mainstream tasks. Enforcing representations of similar data points to be close in the embedding space leads to more *descriptive* representation. Beyond descriptiveness, it is valuable to learn *disentangled* representations by disentangling the factors of variation in the data. Disentangled representations lead to more explainable, transferable, abstract representations. Moreover, learning representations on one task and *transferring* them to solve a similar task is beneficial for utilizing extracted information that could be useful for another task.

This thesis proposes representation learning algorithms suitable for various applications. These representations, depending on the application, are transferable, invariant to some nuisances such as domain shifts, descriptive, and disentangle the factors of generation of data. To facilitate this, two main contributions are proposed. First, enhancing diabetic retinopathy detection by using transferred representations of a lesion segmentation task. Triplet embedding loss is utilized to learn more descriptive representations suitable for the segmentation task on one hand and segmentation representations are transferred to predict better diabetic retinopathy risk on the other hand. Second, creating novel disentangled, fair, invariant representations in a non-adversarial manner that accommodates a plethora of applications and data types. These representations disentangle information into orthogonal application-specific and sensitive components which allows fairer representations that are invariant to sensitive information. Deep feature extractors are proposed and various experiments on multiple datasets are conducted to evaluate the representations characteristics and performance on mainstream tasks. The results are competing with or exceeding state-of-the-art methods results. Strengths and weaknesses of each method are discussed. Novel approaches presented in this thesis demonstrate the importance of learning representations with the correct characteristics to promote better, fairer, explainable results.

Zusammenfassung

Erfolgreiche maschinelle Lernalgorithmen sind stark von guten Datenrepräsentationen abhängig. Diese Repräsentationen können auf Basis von Domänenwissen entworfen werden, um bei einer spezifischen Anwendung verwendet zu werden. Die Art der Vorverarbeitung von Daten ist ein entscheidender Teil von maschinellen Lernmodellen. Obwohl dieser Schritt menschliche Finesse in den Prozess der Merkmalsextraktion einbringt, ist er arbeitsintensiv und bietet keine automatische Möglichkeit zur Extraktion informativer, diskriminierender Merkmale aus den Daten. Maschinellernalgorithmen, die weniger Merkmalsentwicklung erfordern, sind eine äußerst wünschenswerte Eigenschaft, insbesondere für die Anwendung von maschinellem Lernen auf neuartige oder sich schnell ändernde Anwendungen. Mit modernen, tiefen neuronalen Netzen könnten automatisch Repräsentationen erlernt werden, die Faktoren der Variation in den Daten erfassen und für eine Vielzahl von Anwendungen verwendet werden können. Allerdings könnten solche Repräsentationen einige Faktoren der Variation verflechten oder verbergen, was die Ausdruckskraft und Abstraktion der Repräsentation beeinträchtigt.

Es gibt mehrere gewünschte Eigenschaften in erlernten Repräsentationen, die zu einer besseren Leistung bei Mainstream-Aufgaben führen könnten. Das Erzwingen von Repräsentationen ähnlicher Datenpunkte, nahe im Einbettungsraum zu liegen, führt zu einer *deskriptiveren* Repräsentation. Über die Deskriptivität hinaus ist es wertvoll, eine *entwirrte* Repräsentation zu lernen, indem die Faktoren der Variation in den Daten entwirrt werden. Entwirrte Repräsentationen führen zu erklärbareren, übertragbareren, abstrakteren Repräsentationen. Darüber hinaus ist es vorteilhaft, Repräsentationen mit einer Aufgabe zu lernen und sie auf eine ähnliche Aufgabe zu *übertragen*, um extrahierte Informationen zu nutzen, die für eine andere Aufgabe nützlich sein könnten.

Diese Arbeit schlägt Repräsentationslernalgorithmen vor, die für verschiedene Anwendungen geeignet sind. Diese Repräsentationen sind, abhängig von der Anwendung, übertragbar, invariant gegenüber bestimmten Störungen wie Domänenverschiebungen, deskriptiv und entwirren die Faktoren der Datenerzeugung. Um dies zu erleichtern, werden zwei Hauptbeiträge vorgeschlagen. Erstens, Verbesserung der Detektion von diabetischer Retinopathie durch Verwendung übertragener Repräsentationen einer Läsionssegmentierungsaufgabe. Der Triplet-Einbettungsverlust wird genutzt, um deskriptivere Repräsentationen zu erlernen, die für die Segmentierungsaufgabe einerseits geeignet sind, und andererseits werden Segmentierungsrepräsentationen übertragen, um das Risiko der diabetischen Retinopathie besser vorherzusagen. Zweitens, die Schaffung von neuartigen entwirrten, fairen, invariante Repräsentationen auf nicht-adversarielle Weise, die auf eine Vielzahl von Anwendungen und Datentypen anwendbar sind. Diese Repräsentationen entwirren Informationen in orthogonale, anwendungsspezifische und sensible Komponenten,

was fairere Repräsentationen ermöglicht, die gegenüber sensiblen Informationen invariant sind. Es werden tiefe Merkmalsextraktoren vorgeschlagen und verschiedene Experimente mit mehreren Datensätzen werden durchgeführt, um die Eigenschaften der Repräsentationen und deren Leistung bei Mainstream-Aufgaben zu bewerten. Die Ergebnisse konkurrieren mit oder übertreffen die Ergebnisse von State-of-the-Art-Methoden. Stärken und Schwächen jeder Methode werden diskutiert. Neuartige Ansätze, die in dieser Arbeit präsentiert werden, zeigen die Bedeutung des Erlernens von Repräsentationen mit den richtigen Eigenschaften, um bessere, fairere und erklärbarere Ergebnisse zu fördern.

Acknowledgments

I would like to express my sincere gratitude to Prof. Nassir Navab for his outstanding supervision, flexibility, and always valuable inputs. I am grateful for the possibilities he created and all the knowledge and love he poured to CAMP chair that built a family-like innovative environment. My deepest gratitude goes to my mentors Dr. Abouzar Eslami and Dr. Shadi Albarqouni. Dr. Eslami has created a unique environment at Carl Zeiss Meditec where innovation and research met application. The freedom and encouragement he gave are unmatched. Dr. Albarqouni has guided me scientifically and mentally. His belief in me and the work we did together is what kept me going and I am forever grateful for that.

I had the honor to work with intelligent colleagues from the CAMP chair at the technical university of Munich and at Carl Zeiss Meditec. I thank my colleagues in Carl Zeiss Meditec in Munich, USA, and India for all inspiring years of working together.

A special loving thanks goes to my soulmate Ghalia, her laugh and love is what's worth living for. Last but not least, I thank my family for all the support and endless giving they provided and I thank my father whose light in my heart shines through me.

Contents

1	Introduction	1
1.1	Outline	1
1.2	Background and Motivation for Learning Representation	1
1.3	Concepts of Representation Learning	2
1.3.1	Learning Deep Representations	2
1.3.2	Disentangled Representations	3
1.3.3	Fair Representations	4
1.3.4	Transfer Learning	6
1.3.5	Metric Learning	7
2	Learning Disentangled Fair Representations	9
2.1	Motivation and Challenges	9
2.2	Related Work	11
2.2.1	Disentangled Representations Literature	11
2.2.2	Fair Representations Literature	14
2.3	Contributions	18
2.3.1	Fairness by Learning Orthogonal Disentangled Representations (ECCV 2020)	20
2.3.2	On the Fairness of Privacy-Preserving Representations in Medical Applications (MICCAI DCL Workshop 2020)	24
2.3.3	Learning Interpretable Disentangled Representations using Adversarial VAEs (MICCAI DART Workshop 2019)	25
3	Learning Representations for Diabetic Retinopathy	29
3.1	Motivations and Challenges	29
3.2	Related work	32
3.3	Contributions	33
3.3.1	Multi-scale microaneurysms segmentation using embedding triplet loss (MICCAI 2019)	34
3.3.2	Microaneurysms Segmentation and Diabetic Retinopathy Detection by Learning Discriminative Representations (IET Image Processing 2021)	38
3.3.3	Deep Learning for Automatic Diabetic Retinopathy Grading of Ultra-widefield Fundus Images (ARVO 2019)	39
3.3.4	Diabetic Retinopathy Detection in Widefield Images by Using Transfer Learning from Handheld Narrow Field Images (ARVO 2020)	40
3.3.5	Machine Learning Techniques for Ophthalmic Data Processing: A Review (IEEE JBHI 2020)	41
4	Conclusion	43

4.1 Discussion	43
4.2 Future Work	44
A List of Authored and Co-authored Publications	49
B Abstracts of Publications not Discussed in this Thesis	51
C Fairness by Learning Orthogonal Disentangled Representations	53
D On the Fairness of Privacy-Preserving Representations in Medical Applications	77
E Learning Interpretable Disentangled Representations using Adversarial VAEs	91
F Multi-scale Microaneurysms Segmentation Using Embedding Triplet Loss	103
G Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations	117
H Machine Learning Techniques for Ophthalmic Data Processing: A Review	131
Bibliography	147
List of Figures	155

Introduction

1.1 Outline

1.2 Background and Motivation for Learning Representation

Representations are the cornerstones for building good machine learning models that perform well on downstream tasks. Machine learning models use representations as inputs to their models in order to perform a meaningful task. The better a representation is, the easier it is for an algorithm to extract useful information from it. All machine learning algorithms require some sort of a representation. The simplest form of a representation is the raw format information of the input data. However, when input data are complex, models suffer to extract information from it. Another form of representations requires a manual feature extraction step before applying the machine learning routine. This feature extraction step is used to create data representation suitable for the machine learning model. When representations are not suitable for the machine learning model, the learning process might be hindered as the model is required to internally apply some preprocessing to the data to extract meaningful information. While having manual feature extraction with domain knowledge is beneficial for including human intelligence in the development process, it is also limiting because it presents a bottleneck in the development process for adapting the system to new inputs as it is a labor-intensive step that requires human resources. Learning representation automatically greatly aids the development process of a machine learning model. Learning the feature extraction process alongside the objective function of a downstream task ensures that learned representations are suitable for the task at hand. Modern deep neural networks learn representations that capture factors of variation and allow the feature extraction to be done automatically. However, vanilla deep learning methods are prone to learning sub-optimal representations. Such representations may be highly sensitive to some variations in the data. This sensitivity may lead unexpected or undesired behavior. Moreover, such representations may entangle multiple factors of variation of the data. This entanglement limits the abstraction power of the representation which in turn may affect the performance on the downstream task.

In this thesis, methods of learning good representations for general purpose and medical tasks are proposed. The methods focus on characteristics of learned representations and examine how such characteristics affect performance on down stream tasks. The main contributions of this work are as follows:

- Imposing fairness constraints on representations by learning to disentangle sensitive information and downstream task information with orthogonality posed as a proxy for independence. This disentanglement promotes better trade-off in the performance on mitigating sensitive information leakage into the representation and performance on downstream tasks.
- Proposing adversarial disentanglement variational autoencoders with total correlation constraints for learning disentangled representations that are evaluated on a medical dataset and perform well on multiple considered metrics.
- Enhancing microaneurysms detection which are early signs for diabetic retinopathy. The detection is done by using a multi-scale approach with metric learning loss that imposes similarity constraints on representations of microaneurysms patches. Microaneurysms segmentation results are proven experimentally in the thesis to improve diabetic retinopathy grading and detection.
- Proposing a transfer learning approach to improve diabetic retinopathy detection and grading by utilizing knowledge from abundant narrow field fundus images in the learning on scarce wide field fundus images.
- Compiling state-of-the-art works on machine learning methods on ophthalmic data analysis into a review article and listing publicly available datasets and current limitations and future outlooks of the domain in the article.

1.3 Concepts of Representation Learning

In this section, various concepts related to representation learning are introduced. First, learning deep representations with automatic feature extraction is introduced. Disentangling generative factors in the representations and the benefits of such disentanglement are presented afterwards. Then, I present what is fairness in learning representations and why it matters to have fair representations in some applications. The concept of transferring knowledge from one application to the other is then present in transfer learning section. Lastly, learning representation by imposing similarity or dissimilarity measures is introduced in the metric learning section.

1.3.1 Learning Deep Representations

"To fight an exponential, it seems reasonable to arm oneself with other exponentials." - Bengio *et. al* [8]. Learning representations with deep architectures allows hierarchical extraction of meaningful features. Deeper architectures allow progressive extraction of features where each layer has potentially a higher level of abstraction compared to previous layers. This is an important concept in building more *abstract* representations as more abstract concepts are built by means of less abstract ones. The higher the abstraction level of a concept it, the more invariant it is to local changes of the input. This abstraction allows easier separation of features and explanatory factors in the data and it could be achieved by learning complicated functions

that can represent this level of abstraction [8, 9]. To learn such complicated functions, it might be necessary to resolve to deep architectures [8].

Another crucial property that enables better abstraction and is enabled by deep architectures is distributed representations. To understand distributed representations, one can start by understanding the limitations of undistributed or local representations. In simple terms, a representation is local when each neuron represents a configuration of the input. For example, an input of a skin lesion that is small and circular is represented by a single neuron different from a neuron to represent large and circular lesion. Hence to represent n different variations, n neurons are required in such symbolic representations [27]. These representations are easy to learn or code by hand but are inefficient when data has componential structure [81]. In a distributed representation, concepts are represented by a pattern of firing neurons rather than single neurons. This is represented by a many-to-many relationship between neurons and concepts such that each concept is represented by multiple neurons and a neuron is engaged in representing multiple concepts. This allows for more efficient representations especially when concepts are described by various factors of generation. In the simple case of a representation of n binary features which can represent 2^n configuration each could correspond to a different concept. Distributed representations with deep architectures make it possible to re-use extracted features by building a hierarchy of features. The re-use of different parts of deep architecture (paths) grows exponentially with the depth of the network. Learning theory suggests that representing a family of functions with fewer parameters permits learning with fewer examples. This results in better computational efficiency given less nodes are visited and better statistical efficiency given less parameters are required to be learned by re-use of parameters over multiple inputs.

1.3.2 Disentangled Representations

Although deep learning architectures have unmatched success in the past few years [96, 34, 95], such algorithms lack multiple qualities of biological intelligence such as robustness and generality. Hence, the latent representations learned by such algorithms via supervised objectives tailored for the downstream task seem to overfit to the learned task and are not suitable for transfer to other tasks. This is mitigated in the literature in various manners such as data augmentation [43], auxiliary tasks [38], and multi-task learning. Another approach to overcome these problems is to learn generic task-agnostic representations that capture vital features for good performance on a plethora of tasks. It is argued that disentangled representation could be helpful for achieving such representations [7, 77].

One definition of a disentangled representation is a representation where each component of the latent units describes changes of one generative factor of the data and is invariant to other generative factors. Plenty of generative factors of data are independent and they change independently from each other in the input distribution. Moreover, only few of them change at a time when considering real world differences between inputs. The complexity of data in that sense comes from the interaction of multiple factors of variation which makes downstream tasks of learning more complex. For example, in a face detection application on problem might arise is the direction of the light source which changes the appearance of the face and might emphasize different borders. These changes makes it difficult for an entangled

representation to detect the same face under different lighting conditions as face appearance and light source are correlated for such representation. When this information is disentangled in the representation, it becomes easier for a subsequent AI algorithm to understand better the data and create more informed decisions despite the complex interactions between the factors of variation.

Some works in the literature aim to learn representations that disentangle as many factors of variations as possible in an unsupervised manner [35, 13, 41, 86]. In such approaches the objective is to emphasize factorization in the representation components as constraint to achieve better disentanglement without imposing assumptions related to which factors of variations should be disentangled. Other works use some form of supervision to achieve better disentanglement with few amounts of labels [58] or with weak supervision [57]. Another form of disentanglement prominent in the literature is disentangling information in the data into two (or more) representations each is responsible for a generative process of the data. In such cases factorization of the components is not done and each representation of each generative process could be entangled. However, information related to the two generative processes are separated by the representations. Common applications are to disentangle style from content in natural images [48, 37] or to separate anatomical factors from modality-related factors in medical imaging [10, 76].

It is worth noting that invariant and disentangled representations differ conceptually. In invariant representations, the representation is insensitive to changes of data that are uninformative to the downstream task while disentangled representations are equivariant to changes of data that are relevant to the downstream task. Hence, the concepts differ in how much information each preserves. In various tasks, it is difficult to decide which information should be treated as relevant and which is not. Moreover, learned representations might be used in multiple tasks which makes it more difficult to decide what information is relevant and what is not. In such cases it seems reasonable to disentangle as many factors of variation as possible while losing as little information as possible.

1.3.3 Fair Representations

After the success of machine learning and deep model, the adoption of such models in critical applications such as granting loans and medical decision support systems has increased. When deep feature extractors learn application specific features that optimize the objective of a downstream task, they may encapsulate attributes that represent historical biases or reveal some information that was not intended to be revealed. Hence, such representations may lead to applications that are not fair towards some protected groups. This potential discriminatory behavior has raised concerns about fairness of artificial intelligence and statistical inference algorithms [63, 69]. Lately, the European Union has released a report for "Ethics guidelines for trustworthy AI"¹ and in that report it is stressed that unfairness and biases must be avoided.

Many types of biases can be found in data and these biases should be avoided to produce more fair algorithms. Such biases include but are not limited to historical biases, representation biases, measurement biases, evaluation biases, and Simpson's paradox biases. Historical

¹Ethics guidelines for trustworthy AI, <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

bias occur because of social biases that exist and might creep into the data even if data sampling process is perfect while representation bias occur when the sampling process from the population is not well defined. Measurement bias happen because of the measurement method of the data or a specific feature while evaluation bias happens when evaluating the model. Lastly, Simpson's paradox affects the analysis of heterogeneous data that consist of multiple subgroups. The effect prevails when a trend appears in subgroups of the data but when aggregated over the whole dataset the trend might disappear or might be even reversed.

The objective of learning fair representations is to learn a representation that describes well the necessary variation in the data to perform well a target task (downstream application) while being invariant to changes in the data regarding a sensitive attribute. The target attribute resembles the original task on which the model is required to perform well, this could be predicting the salary of an employee based various features that define an employee. The sensitive attribute on the other hand is the predefined direction of invariance that the model is required to be invariant to. This could be the race of the employee in the salary prediction example. The problem is defined such that it is required to predict the salary of an employee (target attribute) regardless of their race (sensitive attribute). The sensitive and target attributes can be independent in some scenarios and the problem is called a domain adaptation problem in such case. An example of such a problem is to predict a certain pathology in medical imaging regardless of the device manufacturer. The presence of the disease is independent from the manufacturer of the device (assuming all manufacturers devices are able to capture the pathology); hence, it is a domain adaptation problem. On the other hand, when the sensitive and target attributes are not independent, then the problem is a fairness problem. One example is the salary prediction problem where race and salary might be correlated because of historical biases. Fig. 1.1 shows the dependency graph of the domain adaptation and fairness cases and the generation process of an observation given the graphs.

A common definition of fairness that is widely used in algorithmic fairness literature [80, 106, 89] is to require the sensitive attribute to be statistically independent from a target attribute. For example, in the salary prediction usecase, it is required to predict the salary of the employee regardless of a sensitive information that we have the belief it should not affect the prediction of salary such as race or gender. Hence, $p(\mathbf{y} = \text{high salary} | \mathbf{x}, s = \text{white})$ should be equal to $p(\mathbf{y} = \text{high salary} | \mathbf{x}, s = \text{hispanic})$. In most evaluations of fair representations in the literature, the representation is used as input to predictive model to predict the sensitive attribute and depending how well the model predicts the sensitive attribute the representation is judged for how much sensitive information is leaked into the the representation. A good fair representation would score as low as a random prediction on the sensitive attribute prediction task. Other definitions exist in the literature where fairness is defined by equal opportunities [33]. In this definition the probability of a data point (e.g. employee) to have a positive outcome (e.g. high salary) is equal for all protected classes (e.g. employee race). This means the true positive rates for the protected classes should be equal. Other definition is called fairness through awareness [22], here it is assumed that similar individuals according to a predefined distance metric should have similar predictions. This is different from fairness through unawareness [29] where the algorithm is considered fair as long as sensitive attributes are not used explicitly in the decision making process. One problem with such a definition is

that some input feature might reveal information about the protected group in a non trivial manner which leads to sensitive information leakage. Lastly demographic parity fairness requires predictions of the target task for all sensitive groups to be the same.

Fairness stems from a belief that is driven by societal factors. In a dataset, some features might be discriminative for the prediction but these features belong to a protected group and we wish for an application built upon the data to be invariant to these features. In that sense, fairness could be viewed as introducing invariance to a model in directions we believe a model should be invariant towards. In applications of learning invariant representations, this belief could be related to a noise profile that the prediction should not be affected by, hence the model predicts regardless of the presence of the noise. This adds robustness to such models towards nuisance. In fairness applications, this noise is what the model is built to be invariant towards and is driven by values and beliefs that are derived by societal factors. As mentioned earlier the target attribute could be correlated with sensitive attribute and in this scenario the sensitive attribute of the protected group would a discriminatory feature of learning algorithm. Therefore, introducing invariance in the direction of the discriminatory sensitive attribute may introduce a drop in the performance on the target task especially if historical biases are strongly represented in the dataset. This suggests a possible trade-off fair models introduce between performance on the target task and the mitigation of sensitive information leakage. In other words, models that are able to hide sensitive information from a representation while keeping performance on the target task intact are considered better models compared to models that hide sensitive information while sacrificing more on the target performance. If sensitive and target information are not independent, then this trade-off is inevitable and better models would give better trade-off of the two performances. In cases of domain adaptation applications where "sensitive" and target information are likely to be independent, such as noise introduced by a measurement device as a sensitive attribute, an improvement in the prediction of target attribute prediction might be introduced when the noise information presence is mitigated in the learned representation. When designing a fairness application, trade-offs between the performance on the target task and mitigation of sensitive information leakage should be kept in mind and reduced as much as possible.

1.3.4 Transfer Learning

Transfer learning allow a learning algorithm to leverage common information among multiple tasks to transfer knowledge between the tasks and eliminate the need for training neural networks from scratch, but rather use the knowledge in a trained model as an initial point for training a new model on different dataset or a different task. Transfer learning could benefit the training convergence in three different manners [99]; it allows the training to have a higher performance start, it allows faster training convergence, and it allows higher performance of the algorithm after convergence. In most machine learning algorithms, the assumption is that training and testing data come from similar distributions. However, this assumption is not always held, especially when collecting training data of certain applications is expensive or time consuming. In such scenarios transfer learning could be of help where a learner is trained on available affordable data (denoted as source domain data) and the knowledge extracted using this learner is utilized for training a new learner on a limited

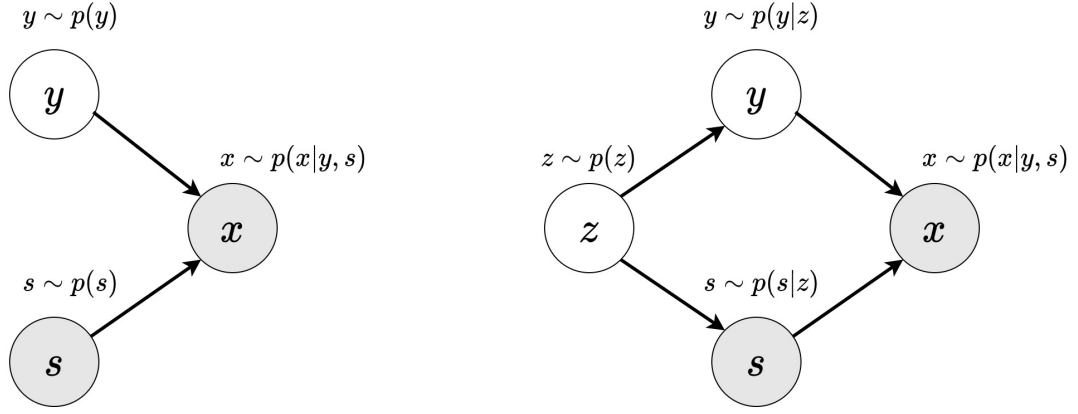


Fig. 1.1. Dependency graph between the target attribute (y), the sensitive attribute (s), and observation (x). **Left:** y and s are marginally independent and generating an observation x is done as follows $s \sim p(s), y \sim p(y), x \sim p(x|y, s)$. This dependency scenario is referred to as domain adaptation applications. **Right:** y and s are not marginally independent and generating an observation x is done as follows $z \sim p(z), s \sim p(s|z), y \sim p(y|z), x \sim p(x|y, s)$. This dependency scenario is referred to as fairness applications.

amount of the more difficult to collect data (denoted as target domain data) that may share some commonalities with the original training data.

According to [98], transfer learning algorithms can be divided into four categories; (1) instance-based transfer learning where some weight-adjusted instances of the source domain are used to supplement the target domain dataset, (2) mapping-based transfer learning where the source and target domains are both mapped into a shared domain, (3) adversarial-based transfer learning where an adversary is used to learn a representation that is domain-agnostic, and lastly (4) network-based transfer learning where a network or part of a network pre-trained on the source domain is transferred to be part of network used for the target domain. In this manuscript the mention of transfer learning corresponds to network-based transfer learning unless mentioned otherwise. A depiction of network-based transfer learning paradigm is shown in Fig.1.2. Network-based transfer learning is widely used in the literature especially by using networks pre-trained on the large ImageNet dataset [20] and then transfer it to a variety of applications [70, 102, 111, 103, 12]. The success of network-based transfer learning approaches come from the hierarchical abstraction the deep networks do. More complex concepts are built based on simpler ones. Hence, in deeper networks, the layers closer to the input are assumed to be extracting lower level information such as edge or blob detectors while deeper layers closer to the output extract higher-level information such as concepts. Transferring knowledge from lower layers of pre-trained network is practiced in the literature because of the fact that these layers are assumed to carry lower level feature extractors that transfer well between tasks.

1.3.5 Metric Learning

Contrastive learning is training technique that learns to distinguish similar and dissimilar samples based on a similarity metric. The goal of metric learning is to learn an embedding function that maps data to a representation where this has some constraints based on the

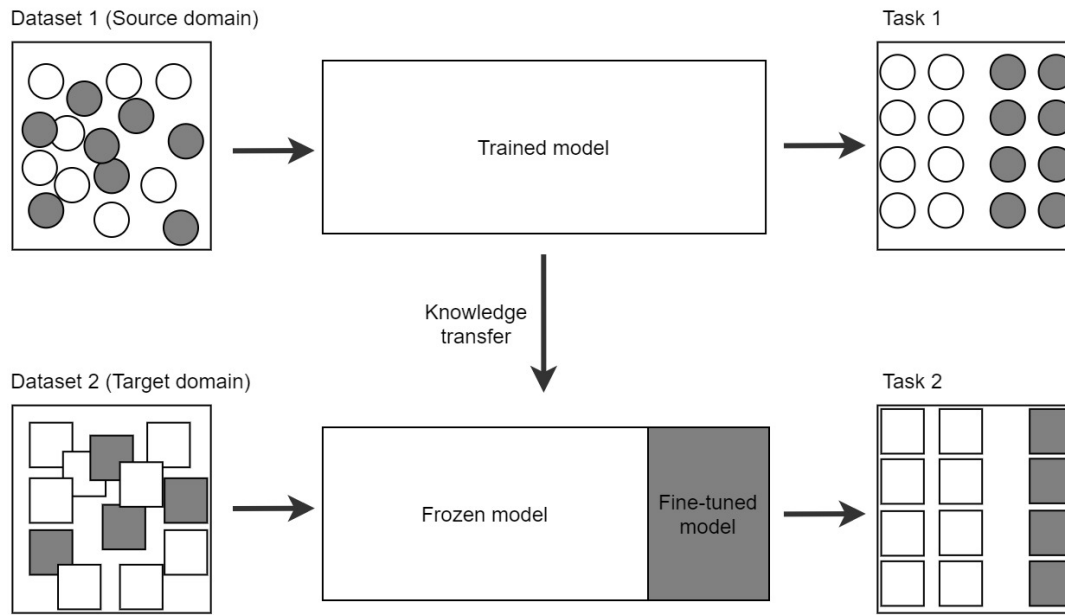


Fig. 1.2. A schematic view of general transfer learning application. The top part is where knowledge is extracted from one dataset for one task. The bottom part utilizes the already trained network on the first dataset and apply changes to parts (or the whole) network based on a new dataset and/or task

similarities between samples. The learning process forces the function to map similar samples from the data space into metrically close samples in the representation space and map dissimilar samples in the data space into metrically far points in the representation space. In deep learning models, one of the early uses of metric learning is the use of pair-wise contrastive loss [15]. In this loss pairs of similar or dissimilar samples are fed to the model and mapped into a learned representation space, the aim of the loss is then to minimize the embedding distance of samples that are similar to each other and maximize the distance of embeddings of dissimilar samples. Assuming that each point is associated with a label, inputs with similar labels are considered as similar data inputs and inputs with different labels are considered as dissimilar data inputs. Later, triplet embedding loss was proposed [92]. Triplet loss allows learning of a relative similarity metric by using a triplet of inputs rather than two similar or dissimilar inputs. The triplet contains an anchor, a positive, and a negative sample. The anchor is sample selected from the dataset around which the other two samples revolve. The positive sample is a sample similar to the anchor while the negative is dissimilar to the anchor. The loss brings the positive sample closer to the anchor and pushes the negative one further from the anchor. Such learning techniques are widely used in person re-identification problems where an embedding is used to match a query image to an image in a database. Such metric losses learn embeddings that are descriptive of the data and are usable for matching samples of the data to extract similar or dissimilar samples.

Since the deeper layers of a neural network contain higher level information that have a higher level of abstraction compared to earlier layers, they are mostly used the embedding layer for such metric learning approaches. It is desired that semantically similar samples are similar in the embedding space and a representation rich in semantic information is one with a high level of abstraction. Metric losses are not only applied to person re-identification problems but also to medical application [108, 6].

Learning Disentangled Fair Representations

2.1 Motivation and Challenges

It is suggested that learning disentangled representations allows for more compact, abstract, transferable, and interpretable representations [7]. The definition of a disentangled representation is still up to debate but a common starting point is to define a representation where one latent unit is sensitive towards one factor of generation of the data while being invariant to other factors of generation as a disentangled representation [7, 35, 13, 41]. Compactness of such representations come from the fact that they have highly distributed structure rather than representing variations in a local representation. Distribution allows for better reuse of neurons in deep architecture which suggests better learning with fewer examples [7]. Abstraction is insured by hierarchically extracting features through the deep architecture of neurons that allow building more abstract concepts on top of less abstract ones. Disentangled representations are more generic in the sense that the same representations could be used for a plethora of applications which allows for better transferability. In terms of complexity, interpretability of a model could be achieved in two main ways, either (1) intrinsically by model design such as in the case of decision trees or attention models [107] or (2) in a post-hoc manner by building a highly complex model and perform reverse engineering steps to extract meaningful explanations of after the model is trained [93, 54]. Approaching interpretability intrinsically could be done in a supervised manner such as decision trees and in an unsupervised manner such as in variational autoencoders (VAEs) [42] or similar models [13, 35, 41]. Learning disentangled representations improves the intrinsic interpretability of a deep model trained in an unsupervised or semi-supervised manner [7, 56]. Since these disentangled representations encapsulate information about different independent generative factors in separate latent components of the representation, they create a better framework for interpreting results on downstream tasks that use the representations [7, 86]. Therefore, having a representation that disentangles factors of variation in the data is becoming a crucial step to building better general purpose representations that could be suitable for various downstream tasks. It is also argued that learning a factorized representation of the data generative factors in an unsupervised manner is a pre-requirement for human-like artificial intelligence algorithms [35, 44] success. In such scenarios, knowledge transfer is highly required to reuse learned representations in multiple applications.

It is argued that disentangled representations provide a fairer framework compared to non-disentangled general purpose representations [55]. It has been shown by Locatello *et al.* [55] that increased disentanglement in the learned representations has a consistent correlation with fairness on downstream applications on a wide range of datasets and applications. This concept allowed Creager *et al.* [17] to build flexibly fair representations

that achieve group and subgroup fairness. Since machine learning applications are rapidly being adopted for critical decision-making procedures such as granting loans and autonomous driving, it became crucial to take fairness into consideration when building such critical applications. The fairness concept is highly related to invariance in the representation where a representation is required to be insensitive towards local changes in the input or changes along the direction of an uninformative variance to the task.

One of the main challenges to learning disentangled representations is the choice of an objective function that suits the problem. For example, in the downstream classification of prediction models, the objective is conceptually clear as to make as many correct classifications with the fewest amount of misclassifications. On the other hand, the objective for learning a representation (a disentangled or fair one for example) is not obvious, especially in unsupervised scenarios. Approaches that aim at learning disentangled representations either go in the direction of disentangling as many factors of variation as possible into the components of the representation [35, 41, 13] or split the representation into two (or more) representations where each describes one generative factor of the data (*e.g.* a domain-invariant content space a domain-specific attribute space [48, 37]). Another challenge for disentangled representation is the evaluation of such representations. How to say for a specific representation that it disentangles enough factors of variation and that they are disentangled well within the learned representation. Directions taken in the literature that evaluate these representations rely on some sort of mutual information calculation between components of the representation and some ground truth factors of variation that are believed to be generating the dataset. Such annotation is difficult to obtain for some applications such as medical ones, hence, limited approaches have tackled the problem of evaluating disentanglement on medical imaging datasets.

As for learning fair representations, the main challenge is to separate sensitive information from relevant (target) information with as little sensitive information leakage as possible while not compromising the performance of the representation on downstream target tasks. Learning representations that are invariant to some nuisance has been studied in the literature. Most of the recent works on fairness purge sensitive information from representations by employing an adversarial learning paradigm to create an adversary game between a discriminator that learns a target task and an adversary that aims to extract sensitive information from the representation. While learning representations in an adversarial manner could lead to good performances, it is more difficult to reach an equilibrium in such minimax game setup [80] and it is unnecessary and may lead to counterproductive results [64]. Hence, learning such fair invariant representations in an efficient non-adversarial manner is beneficial to the community. Moreover, learning an invariant representation is not the only challenge in the case of learning fair representations. Since in many cases of fair applications the target and sensitive attributes are not independent, there is a trade-off between mitigating sensitive information leakage and performance on target applications. Moreover, such fair representations utility is not well discussed in the literature in the case of medical applications. The question for medical scenarios remains, when should socially derived fairness be applied in medical applications? When using general-purpose applications, it is imperative that a drop in performance would occur on the account of a fairer model. This drop may be tolerated in such applications as the value of fairness may outweigh the drop in the performance of the model. Nevertheless, this question becomes more difficult to answer in medical applications where attributes that are

believed to be related to fairness in general purpose applications may be important predictors in case of medical applications.

2.2 Related Work

Achieving representations that are disentangled and/or fair have been recently studied widely in the literature. Approaches that disentangle components of representations rely on modified versions of variational autoencoders that enforce some extra constraints on the representations while approaches that disentangle attributes of the data into a representation per attribute rely on adversarial losses for the generation and consistency losses similar to cycle GANs [113]. Regarding fairness, most recent applications enforce invariance by removing sensitive information from a representation with the help of an adversary and an adversarial loss to reach a minimax game equilibrium. Methods for learning disentangled representation and fair representations are covered separately.

2.2.1 Disentangled Representations Literature

Since learning disentangled representations is mostly approached by a modified version of variational autoencoders [42], the author introduces variational autoencoder models first, then dives deeper into other versions that are developed for disentanglement.

Let $\mathbf{x} \in \mathbb{R}^N$ be a dataset of samples that comes from a distribution parametrized by some latent generative factors $\mathbf{z} \in \mathbb{R}^M$ as the ground truth data generation process. Variational Autoencoder (VAEs) [42] learn to maximize the log-likelihood of the observed data $\mathbf{x}$ expectation over the distribution of the ground truth generative factors $\mathbf{z}$:

$$\max_{\phi, \theta} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] \quad (2.1)$$

where q_{ϕ} represents the encoder as a neural network parameterized by ϕ as learned parameters and p_{θ} represents the decoder function as a neural network parameterized by θ as learned parameters. This can be written as:

$$\log p_{\theta}(\mathbf{x}|\mathbf{z}) = D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) + \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) \quad (2.2)$$

where $D_{KL}(\cdot || \cdot)$ is the Kullback–Leibler divergence between the true $p(\mathbf{z})$ and the approximate $q(\mathbf{z}|\mathbf{x})$ posterior. D_{KL} is non-negative which makes maximizing $\mathcal{L}$ equivalent to maximizing the lower bound to the actual objective in Eq. 2.1. This equates to:

$$\log p_{\theta}(\mathbf{x}|\mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \quad (2.3)$$

The objective in Eq. 2.3 is intractable and requires some assumptions to be made to be usable in practice. The prior $p(\mathbf{z})$ and posterior $q(\mathbf{z}|\mathbf{x})$ distributions are assumed to be of a certain distribution that is a design choice for the VAE. One common choice generally taken in the literature and mostly taken in cases where disentanglement is enforced is to parametrize the distributions as Gaussian with a diagonal covariance matrix. Typically the

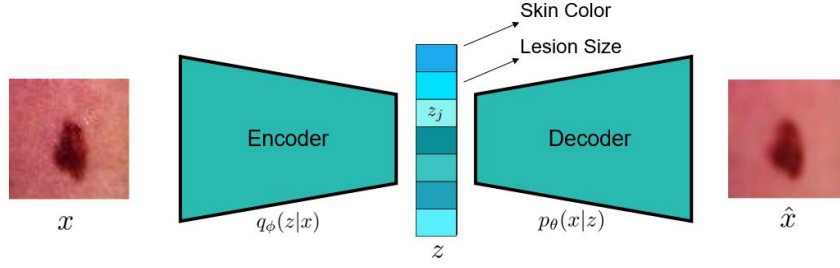


Fig. 2.1. A depiction of general VAE-like architecture. The constraints applied to the representation z lead to different disentanglement approaches [35, 13, 41].

prior is set to the isotropic unit Gaussian $\mathcal{N}(0, 1)$. These assumptions are crucial for the practicality of VAEs training and gradients backpropagation as they allow the usage of the reparametrization trick. With this technique, it is possible to estimate the gradients of $\mathcal{L}$ with respect to ϕ as each component z_i of the representation z is sampled from the respective Gaussian $z_i \sim q_\phi(z_i|x) = \mathcal{N}(\mu_i, \sigma_i)$ as differentiable scale and translate transformation of a noise variable sampled from a unit Gaussian $\epsilon \sim \mathcal{N}(0, 1)$ such as $z_i = \mu_i + \sigma_i \epsilon$. Fig. 2.1 shows a depiction of the general architecture of a VAE where various constraints on the latent code z may lead to different disentanglement behaviors.

Since the prior $p(z)$ and the posterior $q(z|x)$ are assumed to be Gaussians with a diagonal covariance matrix, this motivated Higgins *et al.* [35] to utilize the implicit independence imposed by this prior assumption to enforce a factorization of the latent code components that are associated with better disentanglement. A simple way to enforce this factorization on z is to emphasize the divergence term in Eq. 2.3 by giving it a different weighting β that insures more focus while training on imposing the independence of components in z which acts a proxy for disentanglement. The objective for β -VAE is:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction objective } \mathcal{L}_{rec}} - \beta D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \quad (2.4)$$

Different values of β give different representations with different degrees of emphasis on disentanglement. If $\beta > 1$, then the training takes into account disentangling the independent data generative factors. It is worth noting that not all generative factors of the data are conditionally independent and the size of the representation z must be larger than the number of independent generative factors to learn a descriptive representation that contains not only independent factors of generation but also dependent ones.

It is clear from Eq. 2.4 that using larger β would encourage disentanglement while sacrificing the reconstruction power of the representation. This trade-off implies that information described in the representation while disentangled by independence, is less descriptive of the data compared with a vanilla VAE. This inspired Chen *et al.* [13] to decompose the KL divergence term to better understand the sources of disentanglement in the objective. After decomposing the KL divergence term similar to [36] they got the following objective

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \underbrace{I_q(\mathbf{z}, \mathbf{x}) + \beta D_{KL}(q_\phi(\mathbf{z}) || \prod_j q_\phi(z_j)) + \sum_j D_{KL}(q_\phi(z_j) || p(z_j))}_{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \text{ decomposition } (\mathcal{L}_{prior})} \quad (2.5)$$

The first term is the KL divergence decomposition is referred to as index-code mutual information. It is the mutual information between the data and latent code variable. Some have proposed that higher index-code mutual information would lead to better disentanglement [14] while others proposed removing the term during the optimization [61, 112]. On the contrary, recent works suggest that penalized mutual information motivates more compact and disentangled representations when applied to the information bottleneck [2]. The second term in the KL divergence decomposition is the total correlation (TC) which is a generalization of mutual information to more than two random variables. It acts as a measure of dependence between variables and a higher penalty on this term imposes more independence and disentanglement in the representation components. It is argued that penalizing this term is what makes β -VAE disentangle the data factors of variation. The third term is referred to as dimension-wise KL which keeps individual components of the representation close to their corresponding priors. It is argued by [13] that the TC term is the most important term in the decomposition to achieve disentanglement and use the weighting factor β only on the TC term to achieve not only better disentanglement, but also higher fidelity reconstructions of the data. The model is termed β -TCVAE because of the focus on the total correlation term during the training. To estimate $\log q(\mathbf{z})$ for training the authors proposed a minibatch weighted sampling approach inspired by importance sampling which does not require setting extra hyperparameters for the training. Kim *et al.* [41] proposed another approach for estimating total correlation by using a discriminator and cross-entropy loss on top of vanilla VAE loss. The discriminator learns to differentiate between samples from $q(\mathbf{z}|\mathbf{x})$ and a randomly permuted version for $q(\mathbf{z}|\mathbf{x})$ across a batch as standard trick that is used as a proxy for independence in the literature [5].

Since there is no agreed-upon definition of disentangled representation, there is not one agreed-upon metric that calculates disentanglement utility. One commonly used qualitative metric is latent traversals, which is visualizing the changes in the reconstructions while changing gradually one component of the latent space and then repeating this for all components of the representation. It is good to have a metric to visually assess if the model does any disentanglement at all. However, because of its qualitative nature, it is not suitable to be used for comparing multiple algorithms or multiple versions of one algorithm. Fig. 2.2 shows an example of latent traversal reconstructions of β -TCVAE model trained on a skin lesion dataset. It could be seen that some components control the size and circularity of the lesion which are examples of generative factors of skin lesion data. Other components to be entangled as these generative factors might be not independent in the dataset or the model failed to enforce independence in such scenarios (*e.g.* second to the last component which seem to have information about homogeneity and skin color). On the quantitative end, there have been multiple suggestions for measuring disentanglement that require some sort of supervision (labeling) in the process. The simplest form is the β -VAE [35] metric. It measures the error rate of a linear classifier that is trained on data generated and labeled especially to

measure the proposed disentanglement. The labeled data for the linear classifier is generated by fixing one factor of variation v_k while changing the others and generating L pairs of representations. Absolute difference between these representations is taken and the average of these differences is used as an input sample to the classifier such as $\frac{1}{L} \sum_{l=1}^{l=L} |z_l^{(1)} - z_l^{(2)}|$ where random vectors $z_l^{(1)}, z_l^{(2)}$ are sampled as $x \sim p(x|v_k), z \sim q(z|x)$ for any fixed value of v_k . The label for such sample is the fixed factor k where representation at its index is supposed to be zero if the associated component solely represents factor v_k . The weakness of such an approach comes from the linear classifier where the choice of hyperparameters might be a factor in the performance and it might capture a linear combination of components that could represent a factor. To include axis alignment (having one latent variable that is related to each factor) and remove hyperparameters sensitivity from the metric, Kim *et al.* [41] have proposed using the index of the component with the lowest variance as the input to the classification model such as $d^* = \operatorname{argmin}_j \operatorname{Var}_{q(z_j|v_k)}[z_j]$ and k is the label. The deterministic majority voting error for these input-output pairs represents the metric. Chen *et al.* [13] suggested that classifier-based metrics are merely measuring the reduction in entropy of z when a factor of variation is observed. They have proposed a new metric named Mutual Information Gap (MIG). In this metric mutual information between the latent variable z_j and a ground truth generative factor, v_k is estimated. In order to ensure axis alignment, the mutual information scores of the two latent variables with the highest and second highest mutual information are subtracted. This way, the metric ensures that only one latent variable has a high mutual information score with the observed generative factor. Lastly, the score is normalized using the entropy score of the generative factor v_k . There exist other metrics for measuring disentanglement that share some similarities with discussed metrics. Ridgeway & Mozer [78] have suggested the modularity score which calculates a deviation of an ideal case where a component is dependent on at most one factor of generation. Eastwood & Williams [23] have proposed a disentanglement metric where the entropy after normalizing each component based on the importance in the prediction of a factor of generation is calculated as a metric for disentanglement.

2.2.2 Fair Representations Literature

Approaching fairness has conceptually changed over time, early in the literature, data massaging techniques were used to ensure that samples of each group have an equal number of positive examples. A classifier is then trained on the new labels with the assumption that equal opportunities in the training set would translate to equal opportunities in the testing set. [39, 74]. Data massaging techniques suffer from potential under-utilization of the available data samples and add complexity to the logistics of the data collection process that when possible is arguably beneficial to the final product. The other category under which most of the newer works fall in is termed regularization-based strategy. In this category, a regularization term is added to the training objective that minimizes bias or awareness of the protected/unprotected group status to derive a group-agnostic model that performs a target task equally well on both groups. Zemel *et al.* [110] proposed a semi-supervised fair clustering method that clusters data points while constraining the distribution of protected groups to be similar in each cluster. Later, the Variational Fair Autoencoder (VFAE) [59] was proposed to learn a representation that describes well the data yet is invariant to some arbitrary attribute. These invariant fair representations could be then used for another downstream task where fairness is required.

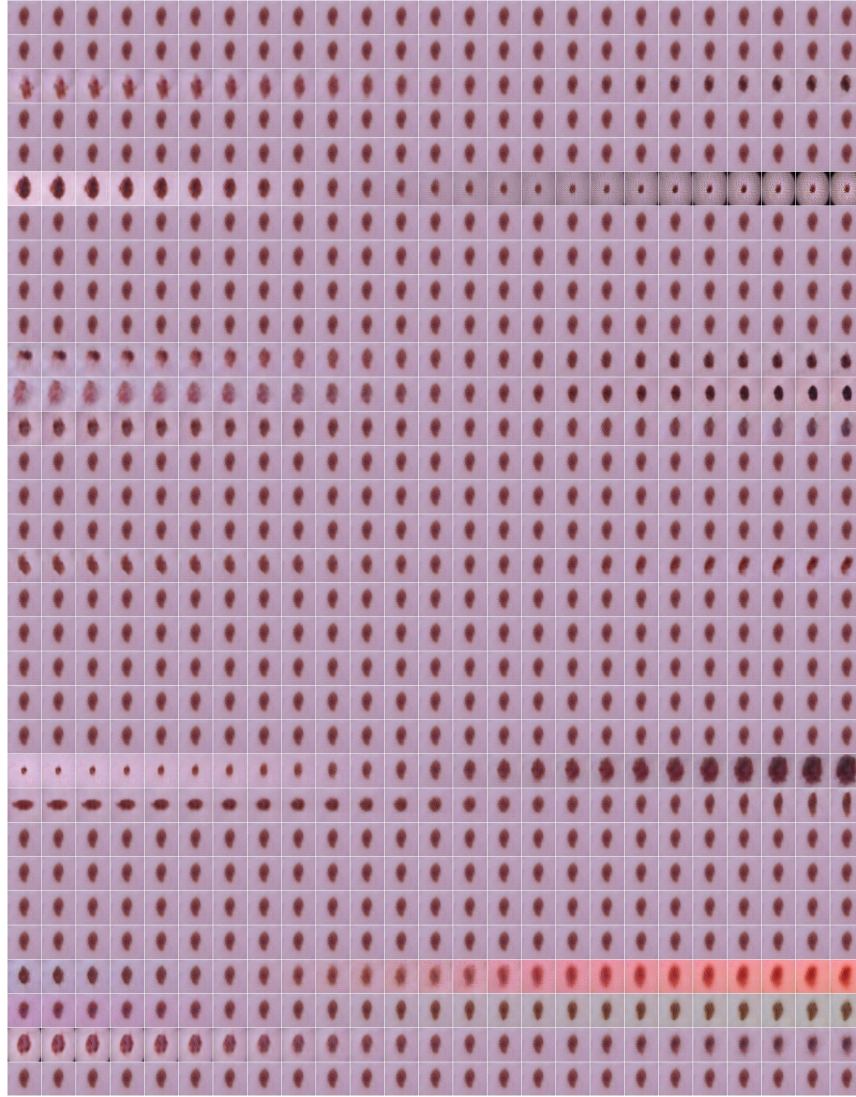


Fig. 2.2. Latent traversal reconstructions of a proposed adversarial disentanglement framework trained on skin lesion dataset[100]. Each row represents one component of the representation (32 components) and columns represent gradual changes in the component value in the range $[-3, 3]$

More modern approaches to learning representations rely on concepts from Generative Adversarial Networks (GANs) [28] to generate more robust invariant representations [106, 80, 60]. The invariance problem is set as a minimax game between an encoder that maps inputs to a latent representation space with the help of a target task, and an adversary that learns to extract sensitive information from the encoded representations. The encoder and adversary are trained alternatively with the encoder objective to minimize the likelihood of the adversary while the adversary is forced to discriminate sensitive information in the representation. More formally, the task of a fair representation application is to learn from input data x a representation z that is descriptive of a target attribute y while being invariant to a sensitive attribute s . Assuming that $p(y|x)$ and $p(s|x)$ are the ground truth labels of the target and sensitive attributes, the adversarial game is composed of three components: (1) an encoder E that maps inputs to representations as a deterministic function $z = E(x; \theta_E)$, (2) a target predictor T that learns a distribution $q_{\phi_T}(y|z)$ to model the conditional distribution $p(y|x)$, the predictor network predicts target attribute y given the representation as input, and

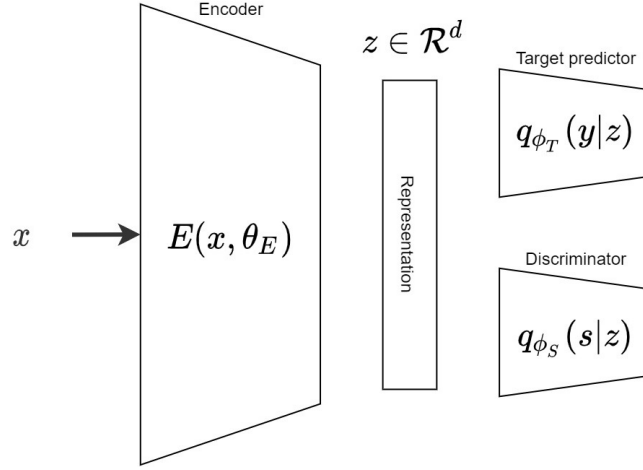


Fig. 2.3. A general framework of adversarial fairness where a representation z is trained to be descriptive of a target task to predict an attribute y while being invariant to a sensitive attribute s that contains information about a protected group.

(3) a discriminator that learns a distribution $q_{\phi_S}(s|z)$ to model the conditional distribution $p(s|x)$, the discriminator extracts sensitive information s from the learned representation z . A depiction of this general adversarial fairness framework is shown in Fig.2.3. Xie *et al.* [106] has posed the problem as a three-player zero-sum adversarial game:

$$\begin{aligned}
 & \min_{\theta_E, \phi_T} \max_{\phi_S} J_1(\theta_E, \phi_T) - \alpha J_2(\theta_E, \phi_S) \\
 & J_1(\theta_E, \phi_T) = KL(p(y|x) || q_{\phi_T}(y|z)); \quad z = E(x; \theta_E) \\
 & J_2(\theta_E, \phi_S) = KL(p(s|x) || q_{\phi_S}(s|z)); \quad z = E(x; \theta_E)
 \end{aligned} \tag{2.6}$$

where θ_E, ϕ_T, ϕ_S are the parameters of the encoder, the target predictor, and the discriminator respectively. α is a trade-off parameter. While problems posed as minimax zero-sum games perform well in the literature, they sometimes suffer from convergence stability problems and require some regularization to create a more stable training paradigm. Roy *et al.* [80] have proposed a solution to the problem by changing the rules of the game. Rather than posing the problem as a zero-sum minimax game, they posed it as a non-zero sum minimization game where two objectives are competing when optimizing for metrics that are related but not exactly the same.

$$\begin{aligned}
& \min_{\phi_S} V_1(\theta_E, \phi_S) \\
& \min_{\theta_E, \phi_T} V_2(\theta_E, \phi_T) + \alpha V_3(\theta_E, \phi_S) \\
& V_1(\theta_E, \phi_S) = KL(p(\mathbf{s}|\mathbf{x}) \parallel q_{\phi_S}(\mathbf{s}|\mathbf{z})); \quad \mathbf{z} = E(\mathbf{x}; \theta_E) \\
& V_2(\theta_E, \phi_T) = KL(p(\mathbf{y}|\mathbf{x}) \parallel q_{\phi_T}(\mathbf{y}|\mathbf{z})); \quad \mathbf{z} = E(\mathbf{x}; \theta_E) \\
& V_3(\theta_E, \phi_S) = KL(q_{\phi_S}(\mathbf{s}|\mathbf{z}) \parallel \mathcal{U}); \quad \mathbf{z} = E(\mathbf{x}; \theta_E)
\end{aligned} \tag{2.7}$$

where α is the trade-off hyperparameter for the two competing objectives, $\mathcal{U}$ is the uniform distribution. The competing objectives V_1 and V_3 are working against each other but they are optimizing different metrics as V_1 optimizes cross-entropy while V_3 is entropy maximization loss (minimized KL divergence from the uniform distribution). Hence, the problem is considered a non-zero-sum game in these settings. The entropy loss forces the representation to be agnostic to information about sensitive attributes which is a desired property of fair representations. In an attempt to utilize disentangled representation in fairness applications Creager *et al.* [17] have proposed a framework for learning fair representations by disentanglement. The model allows for flexibly manipulating sensitive information at test time and combining multiple sensitive attributes at test time to achieve subgroup fairness. The disentanglement is achieved by splitting the latent representation into a factorized sensitive latent space and a non-sensitive latent space. The factorization of the sensitive latent space allows for easier control over the components of the representation at test time. The Independence between the sensitive and non-sensitive representations, and the factorization of the sensitive latent space are achieved by adversarial training where an adversary is trained alternatively with other objectives to differentiate between an aggregated posterior coming from the model denoted as real samples and simulated independent distribution coming from the products of components of the sensitive distribution and the sensitive representation distribution denoted as fake samples. Lastly, there exist other approaches in the literature that utilize other forms of mutual information estimations to avoid the adversarial training model as they are less stable during training, more difficult to converge, and are not always necessary [64].

The evaluation of fairness could be done quantitatively and qualitatively. Quantitative evaluation of fair representations answers mainly two questions: *How much sensitive information is leaked to the representation?* and *how much target information is retained in the representation?*. To answer these questions, the learned representations are used after the training in two evaluation tasks. The first is to classify target attributes based on the learned fair representations. The higher the classification accuracy on this task the better the representation is in retaining information about target attributes which is a desired property of a fair representation. The second evaluation task is to classify sensitive attributes based on the learned fair representations. The representation in an ideal case should be uninformed about sensitive attributes. Hence, the accuracy performance of the classifier should be similar to random guessing of the over-represented class in the dataset. On the other hand, qualitative evaluation includes visualization of the 2D projections of the latent codes in fairness and non-fairness scenarios. It should be visible from the projections that in the case of fairness,

target classes are easily separable while sensitive classes are more mixed and difficult to separate. Fig.2.4 shows an example of qualitative analysis on the Health heritage dataset ¹ with the target task being the hospital readmission status of a patient and the sensitive attribute is the age of the patient. Fairness is achieved using the author's proposed framework [90]. It could be seen from the figure that when fairness is used it is more difficult to conclude meaningful information about sensitive attributes which is not the case when fairness constraints are not applied.

2.3 Contributions

Learning representations difficulty stems from the fact that there are still debates about which characteristics are most important in a representation and what type of objectives satisfy such characteristics. One commonly agreed-upon characteristic of a good representation is the disentanglement of the factors of variation in the data. This achieves more explainable, abstract, distributed representations that also perform well on downstream tasks [101]. Most of the works in the literature focus on toy datasets to build better disentanglement and new metrics of disentanglement. While this is valuable for advancing the theory on the topic, it sometimes fails to tailor the model and objectives to real-world datasets that may have different requirements to achieve a successful disentanglement. Moreover, works that quantitatively assess disentanglement properties on medical imaging data have not been investigated because of smaller datasets associated with medical applications and the lack of ground truth labels of the generative factors that are necessary for the quantitative evaluation of disentanglement. Most disentanglement-related solutions applied to medical datasets focus on image translation applications to separate content and style components of an image to be used for domain adaptation [10, 109] and multi-modal registration[76] applications. In the author's opinion, a good disentangled representation in the medical domain should quantitatively and qualitatively perform well in discovering some factors of generation in the data that are responsible for pathologies or abnormalities regarding the specific application. Ideally, one factor of variation or a combination of multiple factors would be responsible for a disease or a case of interest for a downstream application.

The domain of learning fair representations has not yet been extensively explored, especially with deep neural network techniques. However, it is gaining more attention lately as critical applications rely now on automated decision-making features powered by machine learning models. When such models are not fair towards a protected group that there is a belief of its sensitivity, then the trust of such models would be affected and their adoption might be hindered by such unfairness. It has been recently shown that a disentangled representation has the potential to be more fair compared to entangled representations[55]. This idea of using disentanglement of representations in fairness applications is not yet well explored in the community and has the potential to achieve better fairness while having the perks of a disentangled representation. Moreover, most of the current fairness publications in the literature focus on achieving fairness by adversarial training. While this may lead to good performance, it has been shown that it sometimes lacks the stability and ease of convergence [64, 80]. A method that achieves fairness without the need for an adversarial

¹Heritage Health dataset, www.heritagehealthprize.com

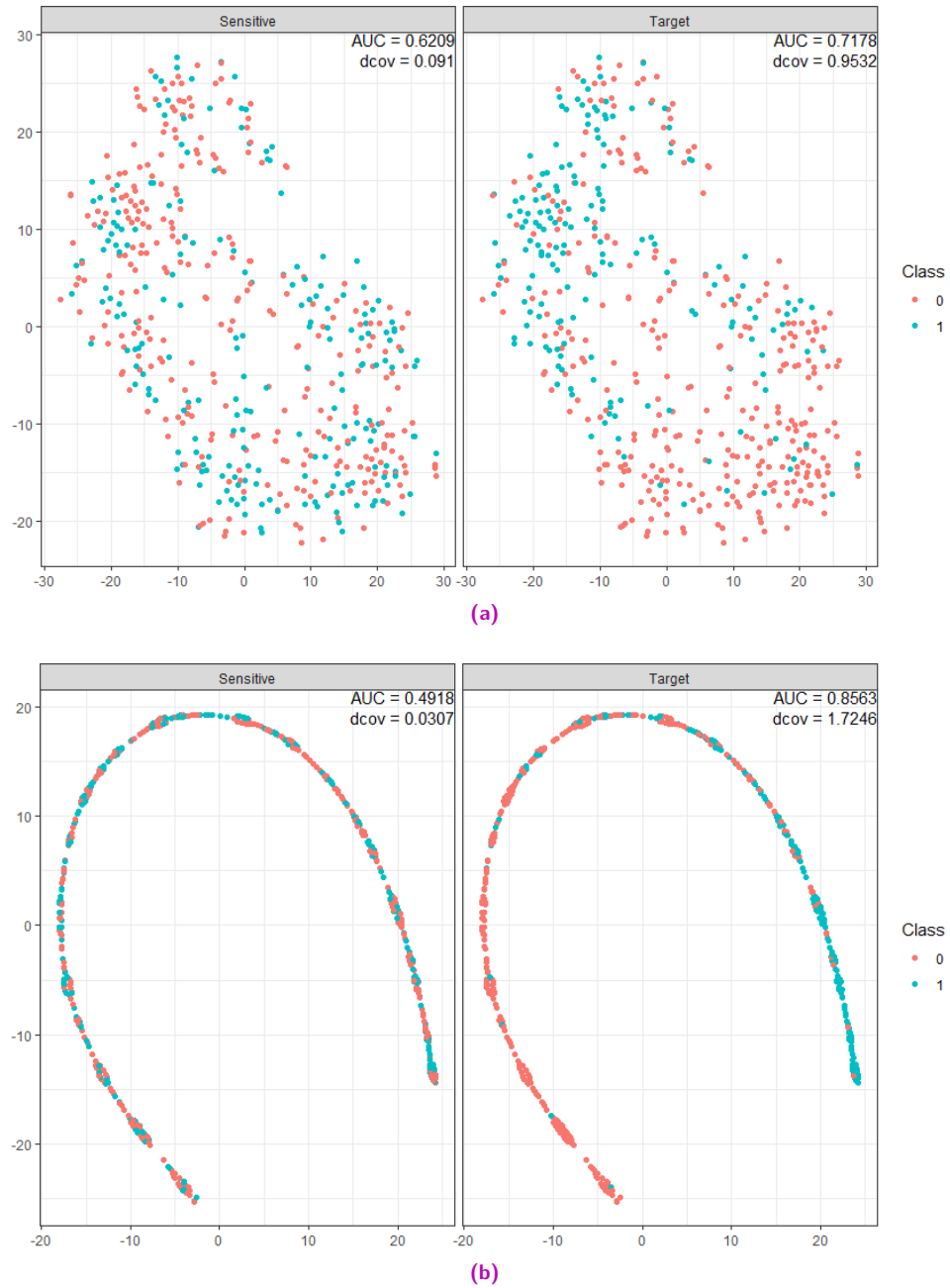


Fig. 2.4. t-SNE projections of the health heritage dataset representations without (a) and with (b) fairness constraints. The left figures highlight the sensitive attribute labels. The right figures highlight the target attribute labels. AUCs of classifiers trained on the data with sensitive and target labels are shown for comparison. Distance covariance is also calculated for both cases. (a) It is possible to extract some information about sensitive attributes using a simple classifier. (b) Data points that are close in the projected representation space have close target labeling but not necessarily a close sensitive labeling which is a desired property in fair representations.

training paradigm while staying competitive or better than state-of-the-art adversarial methods in terms of sensitive and target performance trade-off is valuable in terms of applicability and performance. An area of debate that is yet to be settled is the concept of fairness in the medical domain. It is not clear by definition what fairness entails for a medical application. In many definitions, fair representations are invariant to some sensitive attributes that there is a

societal belief of their sensitivity. However, the commonly agreed-upon protected groups in society might contain some features that are necessary for the assessment of some diseases. It is necessary to discuss and evaluate fairness for such applications to understand better what affects some pathologies and what could be considered as a protected feature of a population.

In the coming sections, the contributions of the author towards enhanced fairness and disentanglement of representations in general purpose and medical applications are listed and described. Each work aims to propose a new methodology and/or application for learning better representations and evaluating these works qualitatively and quantitatively. In all contributions, a new methodology for achieving better representation learning is proposed and then evaluated either on medical scenarios when applicable or on general-purpose datasets.

2.3.1 Fairness by Learning Orthogonal Disentangled Representations (ECCV 2020)

Recent developments on fairness focus on purging sensitive information from a learned representation to create an invariant representation towards arbitrary nuisance. These efforts are mostly focused on creating an adversary to mitigate sensitive information leakage into the representation. Moreover, many of these approaches do not exploit disentanglement concepts towards achieving a better fairness with more control over the representations components. The author of this work proposes a fairness framework that exploits the abstractness and distribution power of disentangled representation to build a better fairness trade-off with more control over the representation. The framework is non-adversarial and exploits concepts of orthogonality and entropy maximization to achieve disentanglement of sensitive and non-sensitive information.

One motivation behind the proposed approach is advances in disentangled representations where it became more evident that because of their descriptive power, such representations have the potential to perform better in fairness applications. The reasoning for this is that disentangled representations are supposed to capture information about multiple generative factors and separate them into different components of the representation. This separation limits dependence and mutual information between the latent dimensions which in turn incentivizes a predictor to rely only on parts of the code that are relevant to a target task of the predictor and not on the dimensions corresponding to factors of generation of sensitive information. Notations from Section 2.2.2 are used to describe the concept more formally. Assuming a disentangled representation is trained to separate sensitive and target information as two generative factors of the data. Hence $\mathbf{y}$ and $\mathbf{s}$ are separated into two independent components $\mathbf{z}_T$ and $\mathbf{z}_S$ respectively. In such disentanglement setup, $\mathbf{z}_T$ depends only on $\mathbf{y}$ and $\mathbf{z}_S$ depends only on $\mathbf{s}$. According to [55], in such scenario, an optimal classifier would learn to ignore the part of the input that is not informative about the target task and independent from it. An optimal classifier suffices the relationship $p(\hat{\mathbf{y}}|\mathbf{z}) = p(\mathbf{y}|\mathbf{z})$ where $\mathbf{z}$ is the learned representation. In the aforementioned disentanglement of sensitive and target information scenario, this relation will hold $p(\hat{\mathbf{y}}|\mathbf{z}) = p(\mathbf{y}|\mathbf{z}) = p(\mathbf{y}|\mathbf{z}_T, \mathbf{z}_S) = p(\mathbf{y}|\mathbf{z}_T)$ since $\mathbf{y}$ is not dependent on $\mathbf{z}_S$. Such optima classifier on representation $\mathbf{z}$ has the potential

of being fairer than an optimal classifier on the data x . However, it might have a worse prediction power on the target task. It has been shown by experimental validation that a more disentangled representation is correlated with more fairness on downstream tasks by using various learning techniques, disentanglement evaluation techniques, and datasets. This strong evidence suggests that taking disentanglement into consideration when building a fairness framework is beneficial to the overall achieved fairness.

The definition of fairness used in this manuscript contributions is one that is commonly used in the literature of fairness [106, 60, 80] which requires the sensitive information to be independent from the target information. Hence the prediction of $p(y|x)$ is independent from the sensitive information *i.e.*

$$p(y|x) = p(y|x, s) \quad (2.8)$$

As an example, when the task is to predict the salary of an employee regardless of their race (race is the sensitive attribute), a prediction under a fair model shall have $p(y = \text{high salary}|x, s = \text{race}_1)$ equal to $p(y = \text{good salary}|x, s = \text{race}_2)$. The problem is formulated by assuming that the sensitive and target information are two generative factors of the data that are promoted to be independent. This independence is achieved by maximizing the log likelihood of the conditional distribution $\log p(y, s|x)$ while holding the fairness assumption in Eq. 2.8

$$\begin{aligned} p(y, s|x) &= \frac{p(y, s, x)}{p(x)} \\ &= \frac{p(y|s, x)p(s, x)}{p(x)} \\ &= \frac{p(y|s, x)p(x|s)p(s)}{p(x)} \\ &= \frac{p(y|x)p(x|s)p(s)}{p(x)} \\ &= p(y|x)p(s|x) \end{aligned} \quad (2.9)$$

To promote disentanglement of sensitive and target information as two generative factors, the input x is mapped to a target and sensitive representations z_T, z_S respectively. These two intermediate representations are introduced so $p(y|x)$ and $p(s|x)$ are not evaluated directly but through z_T and z_S such that

$$p(y, s|x) = p(y|z_T)p(z_T|x)p(s|z_S)p(z_S|x) \quad (2.10)$$

The overall objective is to maximize the log-likelihood of Eq. 2.10 two constraints to satisfy disentanglement and fairness assumptions: (1) $p(z_S|x)$ and $p(z_T|x)$ are independent, and (2) z_T is invariant to sensitive information s .

$$J = -\log p(y, s|x) \text{ s.t. } \text{MI}(z_T, z_S) = 0 \text{ and } \text{KL}(p(s|z_T), \mathcal{U}) = 0 \quad (2.11)$$

$$J = -\log p(y, s|x) \text{ s.t. } \underbrace{\text{MI}(z_T, z_S) = 0}_{\mathcal{N}([1,0], I), \mathcal{N}([0,1], I)} \text{ and } \underbrace{\text{KL}(p(s|z_T), \mathcal{U}) = 0}_{\mathcal{L}_E}$$

$$p(y, s|x) = \underbrace{p(y|z_T)}_{\mathcal{L}_T} \underbrace{p(z_T|x)}_{\mathcal{L}_{z_T}} \underbrace{p(s|z_S)}_{\mathcal{L}_S} \underbrace{p(z_S|x)}_{\mathcal{L}_{z_S}}$$

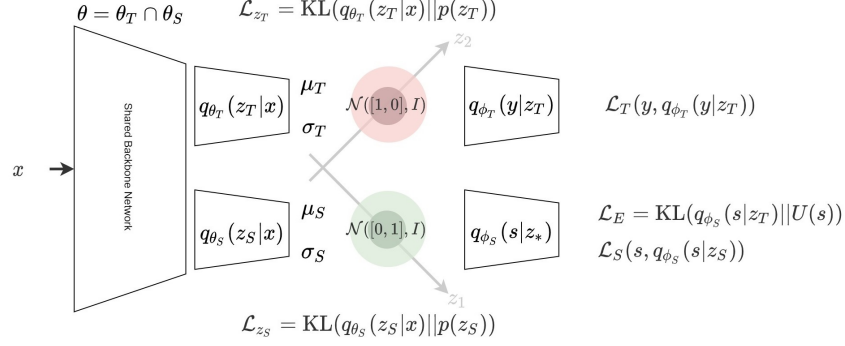


Fig. 2.5. A depiction of the proposed fairness framework. An input data x is fed to a shared encoder and projected into two orthogonal disentangled subspaces representing target z_T and residual sensitive z_S information. The representations are then fed to target and sensitive discriminator and the composite loss J is optimized. Each loss term and constraint is realized by one or more losses which are shown under each term at the top part of the figure.

where $\mathcal{U}(s)$ is the uniform distribution and KL is kullback-leibler divergence. The overall framework is shown in Fig. 2.5. Each term in Eq. 2.10 and each constraint in Eq. 2.11 correspond to a term in the overall loss function of the framework. First an input sample x is fed to the backbone shared network, then projected into two latent representations namely the target and residual (sensitive) representations by utilizing two encoders that take the output of the backbone as their inputs and produce the aforementioned two sub-spaces; $q_{\theta_T}(z_T|x)$, and $q_{\theta_S}(z_S|x)$, where θ represents the parameters of the shared backbone network, i.e. $\theta = \theta_S \cap \theta_T$. This is different from framework presented in Section 2.2.2 where only one representation is produced and then fed to a classification network. In the proposed work the two representations are fed to two different classification networks for target attributes classification, $q_{\phi_T}(y|z_T)$, and the other is for sensitive attributes classification $q_{\phi_S}(s|z_S)$ to train two supervised losses related to the terms $p(s|z_S)$ and $p(z_S|x)$ in Eq. 2.10:

$$\mathcal{L}_T(\theta_T, \phi_T) = -\sum p(\mathbf{y}|\mathbf{x}) \log[q_{\phi_T}(\mathbf{y}|\mathbf{z}_T)] \quad (2.12)$$

$$\mathcal{L}_S(\theta_S^*, \phi_S) = -\sum p(\mathbf{s}|\mathbf{x}) \log[q_{\phi_S}(\mathbf{s}|\mathbf{z}_S)] \quad (2.13)$$

where $\theta_S^* = \theta_S \setminus \theta$ represents the parameters of the sensitive representation encoder without the shared backbone network to mitigate sensitive information leakage in the feature extractors of the backbone network.

The generated target and sensitive subspaces distributions are enforced to be close to be close to two orthogonal factorized normal distributions that promote orthogonality of the two representations and disentanglement of the components of each representation as a proxy for independence of the two representations. The component-wise disentanglement property is

promoted by forcing a factorized representation on the components of $\mathbf{z}_T$ and $\mathbf{z}_S$. Similar to the prior enforcing objective of β -VAE [35] presented in Section 2.2.1, a factorized normal distribution is enforced on the representations. The losses for creating these two normally distributed representations are:

$$\mathcal{L}_{\mathbf{z}_T}(\theta_T) = \text{KL}(q_{\theta_T}(\mathbf{z}_T|\mathbf{x}) || p(\mathbf{z}_T)) \quad (2.14)$$

$$\mathcal{L}_{\mathbf{z}_S}(\theta_S) = \text{KL}(q_{\theta_S}(\mathbf{z}_S|\mathbf{x}) || p(\mathbf{z}_S)) \quad (2.15)$$

where $p(\mathbf{z}_T)$ and $p(\mathbf{z}_S)$ are normally distributed priors with orthogonal means $\boldsymbol{\mu}_T \perp \boldsymbol{\mu}_S$. The orthogonality is achieved by enforcing the two representations to be projected into orthogonal subspaces. In a simple 2-dimensional case the priors are set as follows $p(\mathbf{z}_S) = \prod_{i=1}^{N_S} p(z_S^i) = \mathcal{N}([0, 1]^T, I)$, and $p(\mathbf{z}_T) = \prod_{i=1}^{N_T} p(z_T^i) = \mathcal{N}([1, 0]^T, I)$. An illustration of this can be shown in Fig. 2.5 where the two priors lie on orthogonal subspaces. Note that the components of the priors are independent and the priors have a diagonal covariance matrix which enforces a factorized representation.

Lastly, to mitigate sensitive information leakage into the target representation, the target representation is forced to be agnostic to sensitive information using an entropy maximization approach similar to [80]. The entropy of the sensitive discriminator q_{ϕ_S} given the target representation $\mathbf{z}_T$ such as

$$\mathcal{L}_E(\phi_S, \theta_T) = \text{KL}(q_{\phi_S}(\mathbf{s}|\mathbf{z}_T) || \mathcal{U}(\mathbf{s})). \quad (2.16)$$

where $\mathcal{U}$ is the uniform distribution and the term in Eq. 2.16 is minimized during training as a smaller KL divergence from the uniform distribution corresponds to higher entropy. To put the previous equations in perspective the objective function for the framework is as follows:

$$\begin{aligned} \arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} & \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) \\ & + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} (\mathcal{L}_{\mathbf{z}_T}(\theta_T) + \mathcal{L}_{\mathbf{z}_S}(\theta_S)) \end{aligned} \quad (2.17)$$

where λ_E , and λ_{OD} are weighing hyperparameters for the *Entropy* loss and the *Orthogonal-Disentangled* loss, respectively.

One limitation of the proposed framework is that the training process requires the definition of a target task priory and have the target annotations readily available for creating the model. This could limit the applicability of the framework in usecases where the sensitive attributes are already defined but the representations are required to be multi-purpose where various downstream tasks could utilize them. A solution for this limitation could be using a reconstruction fidelity objective as an auxiliary target task which allows the extraction of meaningful features without the definition of target labels similar to [105, 59]. A potentially beneficial development of the framework could be replacing the handcrafted definition of the orthogonal subspaces with an extra loss term to learn orthogonality intrinsically using low rank constraints on the leaned representations.

The proposed disentanglement fairness framework provides a solid training paradigm to create fair models of various applications and datasets without the need for alternate training of two competing objectives as in adversarial models. The effect of disentanglement orthogonality loss term is validated by ablative studies and the overall framework achieves state-of-the-art or competitive performance on all 5 datasets that were used in the evaluation process. The dataset were tabular and visual dataset of binary and multi-class classification applications. The proposed model enhances the mitigation of the sensitive information in the target representation while affecting the target performance minimally compared to other state-of-the-art models.

2.3.2 On the Fairness of Privacy-Preserving Representations in Medical Applications (MICCAI DCL Workshop 2020)

To further expand on the idea of disentangled fairness, the author proposes to utilize the parts of the fairness framework discussed in Section 2.3.1 to evaluate the utility of fair representations on medical applications. Recent development of fairness and the increased awareness of its importance and potentials in medical applications [72, 11] have naturally led to the question: *Should medical applications also include fairness in their designs? And what types of sensitive attributes are suitable for medical applications?*. The difficulty of answering these questions stems from the fact the sensitive attributes related to protected groups of general purpose applications have to potential to be discriminative important factors for the detection of some diseases and pathologies. Some diseases appear differently with different ratios in different populations which makes apriori knowledge about the population a beneficial feature for a prediction model.

To study this two medically related datasets are used, (1) one is more related to management of healthcare systems which is the *health heritage* dataset ². In this dataset the health records and insurance claims of over 60k patients are used where the target task is to predict revisitation of the patients in order to use it for insurance claims and the sensitive attribute is the sex of the patient. The other dataset (2) is the *ABIDE* dataset[21] for Autism Spectrum Disorder (ASD) detection based on fMRI scans and demographic data (e.g. age, sex, site). The target task for this dataset was the prediction if the patient is diagnosed with ASD and two sensitive tasks for two sensitive attributes are proposed. The first is the binarized age of the patient and the second one is the sex of the patient.

The performance on the health dataset was similar to other state-of-the-art approaches in the mitigation of sensitive information leakage into the representation. However, the trade-off on the target task performance was better than other state-of-the-art approaches. This application is related to the medical domain but it is not related to diagnosis of a specific disease which puts it in a category in between a medical and a general purpose application. The message from this experiment is that it is possible to accurately predict revisitation of a patient to hospital without information about their sex. On the other hand, the ABIDE dataset use case is closer to medical application scenario where information about a disease is to be extracted based on patient information. First experiment on the ABIDE dataset is by using age as a sensitive attribute, the task is to predict if the patient has ASD regardless of their age. It

²Heritage Health dataset, www.heritagehealthprize.com

is noted in other literature[73] that age might be a redundant demographic feature that does not positively affect the diagnosis of ASD. When used as an extra feature with graph convolutional neural networks, the results did not get better and were sometimes worse [73]. This is a suitable use case for examining fairness since there is a belief that this feature is not responsible for the prediction of the disease. The results of the proposed fairness framework proposed in this manuscript go inline with the findings from [73]. When attempting to remove age information from the learned target representation, the age information leakage into the representation is minimal while the performance on the target task is better than baseline performance and on par with state-of-the-art results. This could suggest that age information is not crucial for the prediction of ASD and might act as noise to the prediction. However, the results for the sex as a sensitive attribute show a different trend. When hiding sex information from the target representation, the mitigation of leakage performance is good but the performance on the target task drops under baseline model performance. This could be explained by the fact that sex information might be crucial for the prediction of the disease and hiding such information would sacrifice predictive performance. This conclusion is implicitly supported by results from Parisot *et al.* [73] where the emphasis on sex information in their building their graph convolutional network led to better performance on the target task. It is worth noting that not being able to separate sex and ASD information without a sacrifice in the target prediction performance may suggest dependence on hand but also it may suggest that within the boundaries of the proposed model it was not possible to efficiently disentangle these two features given the independence relaxation constraints of the model.

To summarize, using fairness in medical applications should be done with special care. It is important to distinguish two use cases when applying fairness constraints on medical data. First use case is when the sensitive attribute does not contribute to the diagnostic decision such as the case of age in the ABIDE data set. In this use case removing the sensitive information from the representation may lead to improved performance on the target task. The second use case is when the sensitive attribute is contributing to the prediction of the disease, in such scenarios, removing information about the sensitive attribute may lead to decreased performance on the target task. Hence, the problem formulation for the medical use cases of fairness is closer to domain adaptation problem as the sensitive attribute is independent from the target and acts as noise source. One limitation of this work is the limited amount of datasets it is evaluated on and limited size of the ABIDE dataset where the real world diagnostic scenario was derived from. It is important to validate these conclusions on more and larger datasets where potential sensitive attributes are available and may contribute to predictions. A good example is the usage of fundus imaging dataset for the detection of diabetic retinopathy as a target task and the population, sex or age as sensitive attributes.

2.3.3 Learning Interpretable Disentangled Representations using Adversarial VAEs (MICCAI DART Workshop 2019)

Works in the literature regarding disentanglement of representations has been on the rise lately with works focusing on different realizations of the latent code component-wise independence that is believed to be a precursor to a good disentangled representation. Such approaches

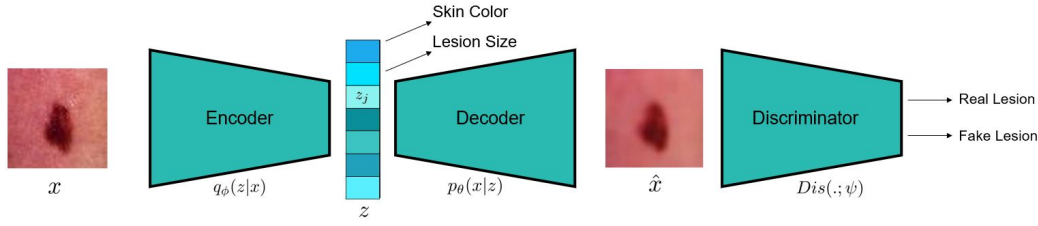


Fig. 2.6. The proposed adversarial disentanglement framework. An input image x is fed to an encoder $q_\phi(z|x)$ to generate an encoded representation z of the input that is constrained with a Gaussian prior and total correlation constraint. A decoder utilizes representation z to reconstruct the input image and produce a fake generated image $\hat{x}$. Then either the fake image $\hat{x}$ or both the real x and fake $\hat{x}$ images are fed to the discriminator network alternatively depending if it is a generator or discriminator optimization step.

are mostly evaluated on toy datasets that contain carefully designed factors of variation with complete annotation for the generation process in order to have a systematic evaluation process of the proposed works. However, to the knowledge of the author, there exists no works that discuss the disentanglement on medical datasets and suggest a disentanglement framework that is more suitable for medical data. Moreover, most of the works focus on achieving better disentanglement scores without giving enough attention to the sacrifice in the reconstruction fidelity scores. The author of this manuscript suggests that a good disentangled representation on medical datasets must adhere to the qualities of a good disentangled representation on other applications while sacrificing as little as possible of the reconstruction power of the representation. This stems from the fact that medical imaging datasets contain some fine details that may be crucial for detection and analysis of some diseases and if the reconstruction fidelity is compromised, it implies that some of these details may have been lost in the encoding decoding process which would negatively affect discriminative models that may use the learned representations on down stream tasks.

The author hypothesises that enhancing the reconstruction fidelity would also enhance the disentanglement of the representation components by virtue of detecting more fine details of the data which might be affecting some generative factors of the data or be generative factors themselves as well. Hence, what is proposed is proposed is a disentanglement framework that has basis similar to autoencoder-based works discussed in Section 2.2.1 with the addition of a discriminator network that encourages reconstruction to be more realistic compared with real reconstructions by using an adversarial learning scheme where a generator (decoder in this case) tries to fool the discriminator by feeding it fake reconstructed images to differentiate it from real images sampled from the dataset distribution. Fig.2.6 shows a depiction of the proposed approach and input image x is fed to an encoder $q_\phi(z|x)$ to generate an encoded representation z of the input that is constrained with a Gaussian prior and total correlation constraint [13]. A decoder then uses representation z to reconstruct the input image data and produce a fake generated image $\hat{x}$. Then either the fake image $\hat{x}$ or both the real x and fake $\hat{x}$ images are fed to the discriminator network alternatively depending if it is a generator or discriminator optimization step. The task of the discriminator is learn a rich similarity metric that focuses on the semantic similarities between images rather than merely comparing pixel-wise difference. This supposed not to only enhance the reconstructions of the images but also allow for extracting more descriptive representations of the data [46]. In such a setup, the pixel level loss (e.g. Mean squared error) would act as a content loss focusing

on pixel level information while the adversarial loss would act as a style loss focusing on the semantic information of the images [25]. Starting from Eq. 2.5 of β -TCVAE the proposition is to add a discriminator $Dis(;\psi)$ parameterized by trainable parameters ψ and train it as counterfit image classifier in an alternative manner similar to other generative adversarial networks approaches [28]. The overall objective of the proposed adversarial disentanglement framework is:

$$\begin{aligned}\arg \min_{\phi, \theta} [\mathcal{L}_{gen}] &= \arg \min_{\phi, \theta} [\mathcal{L}_{rec} + \mathcal{L}_{prior} - \log(Dis(\hat{x}))] \\ \arg \min_{\psi} [\mathcal{L}_{disc}] &= \arg \min_{\psi} [-\log(Dis(\mathbf{x})) - \log(1 - Dis(\hat{x}))]\end{aligned}\tag{2.18}$$

where $\mathcal{L}_{rec}$ and $\mathcal{L}_{prior}$ are the reconstruction and prior loss from Eq. 2.5 respectively. $\mathbf{x}$ is the real image sampled from $p(\mathbf{x})$ and $\hat{x}$ is a generated fake image sampled from $p_{\theta}(\mathbf{x}|\mathbf{z})$.

One challenge for the evaluation of disentanglement in medical applications is the lack of labeling for ground truth generative factors which is essential for evaluating the quality of disentanglement. The author proposes handcrafting few generative factors labels by using segmentation masks provided for a part of ISIC 2018 dataset [100]. Based on these segmentation masks few features related the generation process of skin lesion images could be extracted. Size, eccentricity, and orientation of a lesion are extracted and the disentanglement is evaluated over these generative factors. Such extraction method on ISIC 2018 skin lesion dataset for disentanglement evaluation could pave the way for new methods to utilize existing datasets to develop disentanglement on medical data. The proposed adversarial disentanglement VAE performs well on the studied medical dataset on multiple metrics. The MIG disentanglement score on the handcrafted features is improved compared with other state-of-the-art approaches and performance of using the representation on downstream task of skin lesion type classification performs better than other unsupervised disentangled representation learning approaches. On the reconstruction fidelity, the proposed model performs better than other disentanglement models but fails to beat vanilla VAE because it has the least emphasis on factorization which creates a better focus on the reconstruction fidelity. One limitation of the proposed work is that it is validated on a skin lesion dataset which is more complex than toy examples used in other approaches but not as complex as brain eye scans where more generative factors are interacting and may lead to a more difficult problem of disentanglement. Hence, it would be interesting to evaluate disentanglement on such more complex scenarios but the limitation comes from the available annotations of generative factors for such datasets. Another interesting direction to investigate is the usage of few labels on the generative factors during the training process and during the model selection process. It has been recently shown that using a completely unsupervised approach for disentangling factors of variation in the data may lead to unexpected results and it is not controllable nor guaranteed to extract meaningful factors without any guidance for the disentanglement [56, 58]. It is important to evaluate newly developed disentanglement framework by utilizing few labels in the training and model selection process.

Learning Representations for Diabetic Retinopathy

3.1 Motivations and Challenges

Diabetic Retinopathy (DR) is one of the leading causes of vision impairment and blindness in middle-aged groups [49]. Approximately one third of the population that is diagnosed with diabetes mellitus worldwide have signs of DR. Hence, the interest in the detection and treatment of DR is gaining interest widely. Screening programs for DR are being adopted in order to allow the early detection of the disease especially in large populations where the problem is more prominent. The International Clinical Diabetic Retinopathy Disease Severity (ICDRDS) [4] scale is a widely used grading system for DR severity levels grading. According to ICDRDS the grading of DR can fall in one of five categories that ascend in severity namely: none, mild, moderate, severe, and Proliferative Diabetic Retinopathy (PDR). Table 3.1 shows the details of the ICDRDS scale of each level.

In the early stages of DR, the capillaries weaken which causes microaneurysms to appear in retinal imaging. In fundus imaging microaneurysms appear as small red dots on the retina. When the weak capillaries rupture, hemorrhages happen. They appear as darker and larger red areas compared to microaneurysms. The capillaries become permeable because of DR which causes hard exudates to appear. In fundus imaging they are bright yellow areas with hard edges. If the capillary support fails, then soft exudates appear which look like pale yellow areas with soft edges in fundus imaging. If the case is more severe, new small vessels start to grow because of nutrition deficiency signals from tissues. These fragile new vessels can cause a sudden loss of vision when they grow into the vitreous humour. The creation of these new vessels is called neovascularisation. Fig. 3.1 shows an example fundus image with the main anatomies and lesions related to diabetic retinopathy highlighted.

Segmentation and detection contribution: There are various image modalities that are used for the diagnosis of the eye such as optical coherence tomography, confocal imaging, and fundus imaging. Colored fundus imaging is widely used in the screening programs of DR because of its ease of access and use and wide availability of acquisition devices. Since screening programs aim is to support as many of a population as possible, a large amount of fundus images is acquired and require to be processed by experts for the screening results. This becomes a tedious and financially demanding task very quickly when the amount of collected data is increasing rapidly. Hence, developing an automatic system for detecting DR and its lesions is highly beneficial for screening programs. As presented in Table 3.1, the early detection of DR signs highly relies on the detection of microaneurysms which appear in the early development stages of the disease. These lesions appear as small red dots in fundus images and it is challenging to detect them reliably as they constitute less than 1% of a fundus

Tab. 3.1. International Clinical Diabetic Retinopathy Disease Severity (ICDRDS) scale.

DR severity level	Findings and observable anomalies
0 - No apparent retinopathy	No abnormalities
1 - Mild DR	Microaneurysms only
2 - Moderate DR	More than just microaneurysms but less than Severe
3 - Severe DR	No signs of proliferative retinopathy and one of the following: - Extensive (>20) intraretinal hemorrhages in each of 4 quadrants - Definite venous beading in two or more quadrants - Prominent Intraretinal Microvascular Abnormalities (IRMA) in at least one quadrant
4 - Proliferative DR	One or more of the following: - Neovascularization - Vitreous/preretinal hemorrhage

image [24]. Another challenging part in the detection of microaneurysms is the low contrast between the lesions and the fundus background which creates more difficulties for automatic detection algorithms. Moreover, other structures and lesions in the fundus image might be confused with microaneurysms especially for automatic system. Structures as the crossing of two blood vessels or small hemorrhages are commonly mistaken for microaneurysms. The author believes that a good algorithm for the early detection of DR should start from a good detection of microaneurysms structures because they are crucial components of the early signs of DR progression. Since microaneurysms might be confused with other types of lesions and anatomical structures, having a descriptive representation that describes microaneurysms differently enough from other structures that a classifier can distinguish the two well would be a good solution to the lesion ambiguity problem. Moreover, as microaneurysms are important for the early detection of DR, transferring knowledge from lesion detection task to DR classification task would greatly benefit the DR classification task especially to distinguish early stages of the disease.

Transfer learning contribution: Traditional narrow field funds imaging lacks the view of peripherals of the retina where lesions could appear earlier in some cases. Fundus Wide Field Imagin (WFI) provides a powerful mean to observe and monitor the retinal peripherals. In narrow-field imaging, the field of view is normally 45° which shows the central part of the retina where the optic disc and the macula are located. In WFI, the field of view is more 45° and in Ultra-Wide Field Imagin (UWFI) the field of view is up to 200° . Narrow field imaging covers around 15% of the retina while UWFI can cover more than 80% of the retina [104] It is reported that 10% of the cases DR was reclassified as more severe when UWFI acquisitions were used [94]. It is then valuable to have an automatic system for detecting DR in WFI and UWFI data. However, one challenge for using such data for the prediction is the scarcity of the wider field data. The systems for acquiring such images are on the rise but are still less adopted than narrow field imaging systems that are more affordable. Moreover, the annotation of wider field images is more time consuming than narrow field imaging because there are more data in the image to be processed by the annotator. Since there exists large

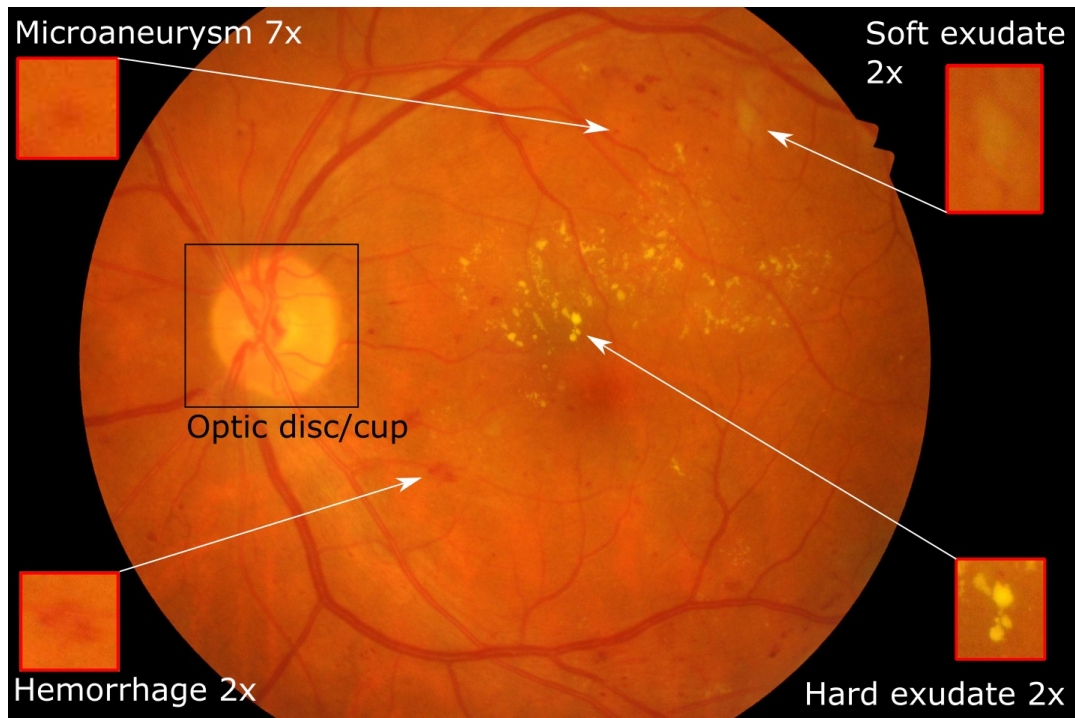


Fig. 3.1. An example fundus image from IDRiD [75] dataset. The lesions most important to diabetic retinopathy are highlighted and the optic disc/cup area is show.

amounts of narrow field data it is useful to utilize this information for any new system that is built for WFI or UWFI based systems. One proposed way to utilize this knowledge is to use transfer learning techniques to transfer knowledge of larger annotated datasets of narrow field images into prediction tasks of wider field fundus imaging. This transfer can solver problems related to data scarcity and annotation difficulty because it requires less data to extract useful information as some of the feature extraction knowledge is transferred from the narrow field imaging. Fig. 3.2 shows an example of a widefield fundus camera and an example ultra-widefield fundus image captured using the camera.

Review contribution: After the wide adoption of machine and deep learning algorithms in various real world and medical applications, they have made their way into ophthalmic applications such as DR, Age-related Macular Degeneration (AMD), and glaucoma detection, and retinal lesions, layers, and vessels segmentation [31, 19, 52, 71, 79, 50, 16]. Moreover, many publicly available datasets exist that are related to the aforementioned tasks. While these works and datasets are plenty and useful for new researches who are contributing to pushing the state of the art further, they are not organized and categorized in one overview document where it is possible to gain more knowledge about the medical motivations for the problems, current solutions, available datasets related to the problems, results on these datasets, and the current limitations and potential future perspectives on these topics. Such knowledge is beneficial to gain an overview of the latest advances on machine learning applications in ophthalmology. The author proposes a review manuscript of recent trends in diagnostic machine learning in ophthalmology for multiple diseases and applications and discusses the technicalities behind some highlighted works. Clinical motivation behind the problems are

¹ZEISS official CLARUS™ 500 page, <https://www.zeiss.com/meditec/int/product-portfolio/retinal-cameras/clarus-500.html>



Fig. 3.2. (a) Example of a widefield fundus CLARUS™ camera¹; (b) and an example ultra-widefield fundus image captured using the camera.

discussed as well with the future perspective in the author's informed opinion presented and discussed.

3.2 Related work

Multiple approaches in the literature use machine learning techniques for the detection of DR or segmentation of DR-related lesions. The most common task of DR detection is either detecting if there are any signs of the disease compared with no apparent retinopathy (level-0 vs. the rest) or the detection of referable DR which is the presence of moderate or more DR signs (level-0 and level-1 vs. the rest). Some approaches use traditional computer vision techniques to extract information about lesions and anatomies (such as blood vessels, optic disk, and macula) of the eye using banks of handcrafted filters. This lesion information could be then passed to k-nearest neighbour classifier as in [82]. In [1], the extraction is done using artificial deep neural networks and then a random forest model is used to fuse all information into a binary classification decision. Later on, the separation between the feature extraction and classification decision has decreased and more end-to-end approaches are used where the input is an image and the output is a classification decision. The feature extraction and classification steps are fused in end-to-end models by using deep neural networks. Such an approach is proposed by [31] where an ensemble of 10 Inception-V3 [97] models are trained using over 100k collected fundus images that are labeled by multiple annotators. The ensemble is trained to detect referable DR among other binary classification tasks. In an attempt to create a deep model that is more informed about salient lesions and anatomies, it has been suggested by [103] to use attention mechanisms in order for the model to focus the training on regions of the image where most discriminative features are. Such model is used to grade DR into one of five ICDRDS scale grades.

Regarding microaneurysms, there are works in the literature that focus on the detection of these lesions in fundus images. Some works use a two step approach where first candidate regions of microaneurysms are suggested and then features that utilize color and shape are extracted around the suggested regions to differentiate positive and negative candidates. Some works use a local rotation feature extractor to detect microaneurysms from the

candidates set[47]. Others use a local convergence index feature extractor that ensures the image gradients of a patch around a candidate converges to point that is considered a microaneurysm [18]. Other approaches use deep neural networks to extract features that represent microaneurysms [32, 45, 71]. Such methods use patches of images as inputs and optimize for the task of classification of the center pixel of a patch. This center pixel classifier is then convolved in a sliding window fashion over a fundus image in order to assign a microaneurysm probability to each pixel of the fundus. Some approaches mix handcrafted and deeply extracted features and use them as input to a classification model [71]. While others use only deep neural network classifiers to classify image patches [32, 45]. To the knowledge of the author, there are no proposed works in the literature that take advantage of metric learning techniques to create embeddings that are suitable for the detection of microaneurysms. This is a crucial point as distinguishing positive and negative candidates reliably requires representations that are discriminative enough to differentiate microaneurysms from other similar lesions such as hemorrhages and similar anatomical structures such as crossing of vessels.

Lastly, the automatic detection of retinal diseases in WFI and UWFI images is newly gaining interest from the community. Some approaches used deep learning for detecting Age-related Macular Degeneration from wider field fundus images and achieved good results in wet-AMD detection [62]. Others have used these imaging techniques to assist the early detection of retinal detachment [68, 53]. For DR detection using widefield fundus, [65] have proposed the usage of deep learning with UWFI to detect proliferative diabetic retinopathy and [51] used deep learning to segment diabetic retinopathy related lesions in UWFI fundus. Lastly, Oh *et al.* [67] have conducted a larger study on wider field fundus imaging and used over 13k UWFI images to detect DR using deep learning techniques and concluded that deep models perform better when using fundus WFI for the detection of DR.

3.3 Contributions

One of the major challenges in automatic detection of DR is the difficulty of detecting lesions responsible for early signs of the disease or the difficulty of collecting enough data (e.g. The case of ultra-widefield fundus imaging). The early signs of DR are defined by the aforementioned small red dots called microaneurysms. These lesions appear at early stages of the disease and remain prominent in later stages. Microaneurysms could rupture and turn into hemorrhages which are correlated with more severe cases of DR. An ideal system for detecting DR should have the capacity to accurately detect microaneurysms presence and locations in order to utilize this information in the DR detection process. However, systems for detecting microaneurysms struggle because the difficulty of the task. Microaneurysms are small and comprise less than 1% of a fundus image. They sometimes do not have enough contrast from the background and are not clearly visible or confused with other lesions such as hemorrhages. The author of the thesis hypothesizes that creating descriptive representations to encapsulate information about microaneurysms is one way to increase the accuracy of detecting these lesions, especially in use cases where annotations about other lesions are not available or not all microaneurysms in the image are annotated. Another difficulty for detecting microaneurysms is the fact that they come in different sizes and shapes. This adds another layer of difficulty to the problem as single scale approaches may suffer from this

variability in microaneurysms sizes. Hence, learning the detection of these lesions on multiple scales of the image is valuable for enhancing the detection.

Another difficulty for detecting DR is the lack of the available annotated data related to DR in some imaging modalities. Ultra-widefield imaging has proven to enhance the detection of DR [94]. However, the available annotated data is not sufficient to build an automatic system for DR detection solely based on this data. On the other hand, plenty of annotated data is available for narrow field fundus imaging and various already successful applications use narrow field data for the detection of DR [1]. To recollect large datasets of wider field fundus imaging and have it annotated by experts is repetitive, time consuming, and expensive. Moreover, discarding all the knowledge accumulated from narrow field imaging would be a waste of resources. Narrow field fundus images share various commonalities with wide field imaging. While wide field imaging covers a larger portion of the retina, lesions in both modalities appear similarly and knowledge about grading or detecting DR gained from narrow field imaging can be transferred to wide field imaging. This allows for the possibility of achieving good performance on wide field fundus imaging without the need for collecting large amounts of data for the model development.

In the following sections contributions by the author regarding the detection microaneurysms, utilizing this knowledge for DR detection and using transfer learning to detect and grade DR in wider field fundus imaging are presented and discussed. A framework for detecting microaneurysms on multiple scales and refining these multi-scale detection results using metric learning techniques that ensure building more descriptive representations of microaneurysms which allows better detection of the lesions when faced with a confusing lesion that might mistaken with microaneurysms. In a later contribution, the microaneurysms segmentation results are utilized in enhancing the detection of DR. The embeddings of segmentation networks and embedding of a grading network are combined to represent and image with more detailed information about microaneurysms segmentation. This new representation is then used for the DR and referable DR detection tasks. Two transfer learning approaches are proposed and discussed, one is to use tabletop narrow field fundus imaging data in the pre-training process and fine-tune using ultra-widefield fundus imaging data. The other approach transfer learning from narrow field fundus imaging data acquired with a handheld camera into ultra-widefield data acquired by a tabletop fundus camera. Lastly, the author presents a review article that summarizes works of machine learning in ophthalmology in the past 4 years and categorizes works into diseases and application and presents publicly available datasets for reader who are interested in the ophthalmic data processing and analysis domain.

3.3.1 Multi-scale microaneurysms segmentation using embedding triplet loss (MICCAI 2019)

In fundus imaging, microaneurysms appear as small red dots scattered in the image. These red dots might have different sizes, shapes, and color shades. Hence, they might be mistaken with other lesions such as hemorrhages and with other anatomical structures or the background. In order to detect microaneurysms on multiple scales, the author proposes to use multi-scale approach for the segmentation of microaneurysms rather than using only one scale that may

overfit to a certain appearance of the data. It has been seen that employing multiple scales for the detection of small structures leads to better results [3]. The author uses two scales for the segmentation realized by two fully convolutional segmentation network that accept patches of a certain scale of the image for each network namely; full scale patches and half scale patches that undergo a downsampling step. The segmentation networks by which the multi-scale processing is achieved are called Hypothesis Generation Networks (HGNs). Since two scales are used in the proposed approach, then two HGN modules are part considered which are the full-scale HGN (HGN $1\times$) and the half-scale HGN (HGN $0.5\times$). HGNs are trained with patches of microaneurysms or healthy samples. Microaneurysm patches are extracted from images with apparent DR while negative healthy patches are extracted from healthy images. This relaxes the requirement that all pixels of an image have to be annotated in order to be used for the segmentation training. Since HGNs are fully convolutional networks, they can be used on a whole fundus image at test time without any need for re-training. Fig. 3.3 top part shows a depiction of the training of the HGNs on the full and half scales.

The use of multiple segmentation networks for multiple scales can solve the detection problem of microaneurysms of different sizes. However, it does not solve the problem of confusing structures that might be detected as microaneurysms. In order to learn a robust classifier that discriminate between various shapes of microaneurysms on one side and other non-microaneurysm structure on the other side, the usage of embedding triplet loss is proposed. Triplet loss is used for learning descriptive representations of data to be used in classification or retrieval tasks such as face recognition [92]. It provides a powerful tool for applying constraints on the embeddings of the training rather than the predictions only. The loss accepts a triplet of images (microaneurysms patches) that represent an anchor, a positive sample and a negative sample. The anchor and positive samples are from the same class while the negative sample is from another class. Triplet loss aims to pull the positive sample and the anchor sample closer to each other in the embedding space while pushing the negative sample and the anchor apart from each other in the embedding space using a predefined metric of similarity. This is realized by using a classification network that distinguishes microaneurysms from non-microaneurysms patches. This classifier is named Patch-wiser Refinement Network (PRN) as it refines the segmentation results of HGN by using a classifier powered by triplet loss. The training of PRN is separate from training HGN modules. Assuming that the anchor patch is a , the positive patch is p , and the negative patch is n , then the triplet embedding loss takes the form:

$$\mathcal{L}_{triplet} = \sum_i^N [d(f(a_i), f(p_i)) - d(f(a_i), f(n_i)) + \alpha]_+ \quad (3.1)$$

Where $f(\cdot)$ is an embedding function that maps an input patch to the last layer of PRN module, α is a margin to ensure a distance between positive and negative pairs, and $d(\cdot, \cdot) \in \mathbb{R}^1$ is a distance measure in the representation space. The angular cosine distance is chosen as distance measure as it shows better performance when training deep neural networks on high dimensional representations [66]. Fig. 3.3 bottom part shows a depiction of the training of the PRN and how triplet loss acts in the embedding space.

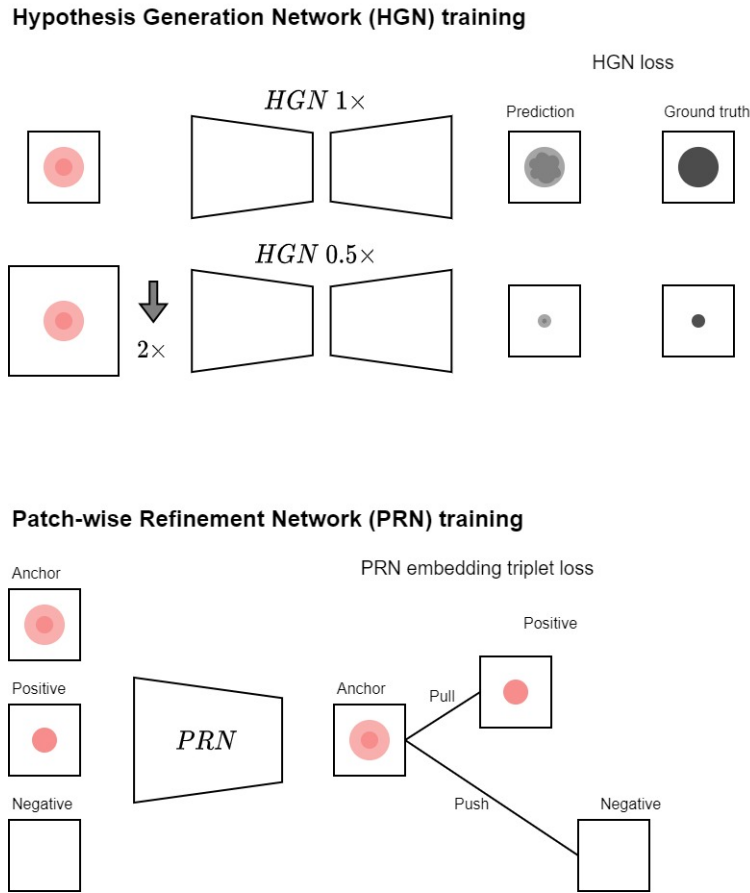


Fig. 3.3. The training of the proposed microaneurysms segmentation framework. The top part describes the *hypothesis generation network (HGN)* modules which are fully convolutional networks that are used as region proposal networks. Each HGN is trained separately on patches of binary segmented images. The bottom part shows the *patch-wise refinement network (PRN)* which is a classifier trained with embedding triplet loss to refine the outputs of the segmentation proposals suggested by HGNs. PRN is trained separately from HGNs.

One challenge for training such a framework is the selection of triplets for the training of PRN module. If all possible triplets are to be considered, the amount of data for training will be prohibitively large. Moreover, microaneurysm patches do not appear in images as healthy patches which creates an imbalance in the dataset that needs to be addressed. To resolve these problems, selective sampling is utilized during the training process [30]. This technique controls what samples are chosen in the next epoch for training. The microaneurysm samples are the under represented class and are chosen uniformly randomly each epoch. However, healthy patches are over represented in the dataset and they are selected for the next epoch with a probability proportional to the error each sample produces after an epoch is complete. Normally, the larger amount of healthy patches and the similar appearance of many of them makes then network insensitive to changes in healthy samples. In selective sampling protocol, not all combinations of triplets are considered and the training is focused on more informative healthy negative samples that produce larger errors.

After HGNs and PRN modules are trained, the inference is done on whole fundus images rather than patches. A depiction of the inference process is shown in Fig. 3.4. Two versions of a whole fundus image (full and half scale) are fed to two HGN modules to produce two

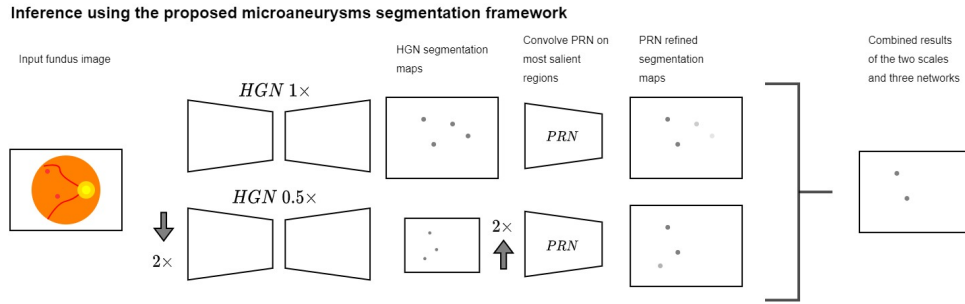


Fig. 3.4. The inference process using the proposed microaneurysms segmentation framework. A whole fundus image with full and half scales is fed to the respective HGN modules to produce to probability maps for microaneurysms segmentation that are used as region proposals for the PRN classification network. The outputs of both modules are combined with arithmetic or geometric averaging to produce a final segmentation map.

probability maps that are used as region proposal for PRN. Patches around high probability areas from HGN are fed to PRN network to refine the selected regions further using the classification network trained with triplet loss. Since PRN is only applied on areas with high probability of being microaneurysms, there is no need to convolve PRN with the whole image which save a significant amount of time during the inference process. After applying PRN over the two HGNs outputs, two new maps are created. The final segmentation result is the combination of the four segmentation maps derived from the two HGNs and the PRN modules. Multiple combination techniques are studied and geometric averaging is found to be beneficial for suppressing false positives.

The results of applying the proposed approach show enhancement on the segmentation of microaneurysms without using annotations of other lesions in the training process. This is useful for various applications such as enhance early detection of DR, improved interpretability of DR detection, big data analysis of populations, and progression monitoring of microaneurysms turnover [26]. One limitation of the proposed work is that it is not an end-to-end approach and may create sub-optimal solutions as it optimizes each problem separately and combines results of each network with the other at inference time. Each network is trained with its own objective function. A more optimal solution trains both modules in one training process where the results of one module would be optimized to have a better fit to the overall objective of the problem rather than optimizing a subproblem at a time. Lastly, the segmentation framework is trained and evaluated on one dataset which may create a limitation to its applicability. This is accounted for in the next section 3.3.2 where segmentation embeddings are utilized for the classification of DR on a different dataset. Another limitation is the lack of global view of the image when performing the segmentation. Since the training is done on patches of data, the network does not see full images of fundus to understand context of the lesions presence. This is partially accounted for with half scale HGN that allow for larger receptive field but it does not give a global view of the fundus image.

3.3.2 Microaneurysms Segmentation and Diabetic Retinopathy Detection by Learning Discriminative Representations (IET Image Processing 2021)

To reflect on a limitation of the contribution in Section 3.3.1, the author quantitatively study the effect of adding microaneurysms segmentation to a DR grading model and evaluate the results based on two tasks namely DR detection and referable DR detection. For this evaluation, a state-of-the-art classification model is used to learn ICDRDS scale 5-grade grading on a large dataset of labeled fundus images and compare the results with and without the addition of segmentation information to the model. The addition is done by using the encoder output of each HGN model (*cf.* Section 3.3.1) as a representative information of microaneurysms statistics in the image. Since this representation encapsulates information used by the decoder of HGN to segment microaneurysms in the image, it is assumed that it holds meaningful information about microaneurysms that could be used for other applications such as DR grading. The grading of DR can benefit from microaneurysms information especially for the earlier scales because they highly rely on microaneurysms information. It is expected that having representative information about microaneurysms would improve the performance of a grading model in differentiating first three levels of the score which correspond to the DR detection and referable DR detection problems.

The grading of DR using segmentation information is done by concatenating representations from three sources and using the new concatenated representation as input to a random forest model that performs the grading task. The first two representations come from the two bottleneck representations in HGNS. These represent the segmentation information in the final classification model. The third representation comes from a classification model trained for the classification task. The last layer before the classification layer is picked as the embedding from the classification model. The last layer is picked as a representation layer because of its high abstraction level richness of information. The classification model and the two HGNS are trained separately before their representations are extracted and used as inputs to train a random forest model. A depiction of the random forest model inputs is shown in Fig. 3.5. The evaluation of the model is done using a different dataset from the training that comes from a different center. This external validation tests how the model would perform in real world use cases where the data does not come from exactly the same source as the training data. Using segmentation information improved the classification results over the vanilla classification model in both DR detection and referable DR detection tasks. This shows the utility of transferring knowledge about microaneurysms segmentation information in enhancing the DR classification and referable DR tasks performance.

Lastly, one limitation of this contribution is that the involved models are trained separately which may lead to sub-optimal solutions when solving each subproblem at a time. The segmentation representations are proven to be beneficial to the classification final objective but training all the model together in an end-to-end fashion would have the benefit of extracting representations that are specifically suitable for the final objective.

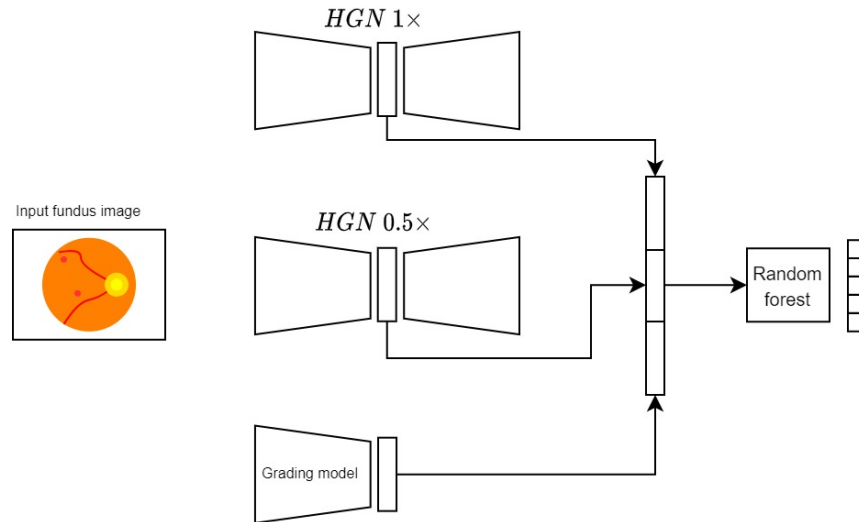


Fig. 3.5. The grading of diabetic retinopathy using representation from hypothesis generation networks and the representation of a classification network used for diabetic retinopathy grading. The training of the hypothesis generation networks and the classification network is done separately then the representations are used to optimize the random forest model for the grading task.

3.3.3 Deep Learning for Automatic Diabetic Retinopathy Grading of Ultra-widefield Fundus Images (ARVO 2019)

The major advantage of wide field fundus scans is that it can show the peripheries where some lesions might go unnoticed when using narrow field fundus imaging. However, the inspection process of wide field and ultra-widefield is tedious and time consuming, especially when the imaging modality becomes more widely adopted and used for screening processes. This leads to the need of an automatic system for the grading and detection of DR using wider field fundus imaging. The challenge for such a task is the scarcity of annotated wider field fundus images that could be used for developing such a model.

To overcome the data scarcity problem, a transfer learning approach is used to develop a model for grading ultra-widefield fundus images using narrow field images as a pre-training step. The rationale comes from the fact the lesions have similar appearances in both modalities which suggest that the transfer of knowledge from a narrow field model to a wide field model should be useful and is a good initialization. The data used for the pre-training of the model comes from a publicly available dataset of over 35k narrow field fundus images coming from tabletop devices of multiple centers. The ultra-widefield data is small dataset that is collected from multiple clinics using CLARUS™ widefield imaging device. To highlight the importance of transfer learning for such an application, three experiments are conducted. The first experiment does not utilize narrow field imaging data and only uses ultra-widefield data for training a deep model on DR grading task. In the second experiment, a deep model is trained until convergence on the narrow field dataset. After the training is complete the model performance is evaluated on the ultra-widefield data without using the ultra-widefield data for the optimization process. In the the third experiment, the ultra-widefield data is utilized in

the training process of the deep model. Similar to the second experiment, the model is trained until convergence on the narrow field datasets. Then the ultra-widefield data is used for fine tuning the trained model. Hence, the narrow field model is used as an initialization for the wider field model.

Since the ultra-widefield dataset is small (119 subjects), the images undergo a series of augmentation operations during the training that include random cropping, translation rotation and flipping. Moreover, both narrow field and ultra-widefield images undergo a preprocessing step for contrast enhancement. The results show a superior performance when including the narrow field imaging data in the training process (Experiments 2 and 3). Extracting information about lesions and DR from the narrow field dataset improved the results on the wider field dataset. In further training of the model on the ultra-widefield data after being trained on narrow field data (Experiment 3) achieves the best results on the ultra-widefield dataset. One limitation of this work is that it utilizes a small amount of ultra-widefield data for the training and evaluation because of the scarcity of such annotated data. Future studies with larger datasets are necessary in order to achieve more significant results.

3.3.4 Diabetic Retinopathy Detection in Widefield Images by Using Transfer Learning from Handheld Narrow Field Images (ARVO 2020)

To further understand the utility of transfer learning on narrow field to widefield knowledge transfer, the author proposes using narrow field fundus images from a handheld fundus imaging device in the pre-training process. This studies not only the transfer capabilities from narrow field to widefield, but also transfer capabilities from handheld fundus imaging devices to tabletop fundus imaging devices. Handheld fundus images devices are a good solution to screening programs because of the ease of use and affordability. The capturing process is fast and patient friendly. However, the wrong alignment of the device might produce image artifacts in the bottom or top part of the image as camera is not aligned well and might be tilted. The quality of the produced images is good for screening programs to find signs of DR but is not on par with other tabletop devices where better camera sensors are utilized. This produces a challenge for automatic algorithms to learn and transfer knowledge from handheld devices to tabletop widefield imaging devices. Fig. 3.6 shows an example image of the VISUSCOUT® handheld device and an example fundus image using the it.

Similar to the setup in Section 3.3.3, three experiments are conducted to evaluate the utility of transfer learning. The First experiment includes training a deep neural network from scratch on ultra-widefield images collected using CLARUS™ widefield imaging device. The dataset contains images from 361 subjects. Data augmentation is used to have a more robust model and to partially account for the lack of variability in the data. In the second experiment a deep neural network is trained using a larger narrow field dataset containing 35k images collected with VISUSCOUT® handheld fundus camera. The evaluation is done on ultra-widefield fundus

²ZEISS official VISUSCOUT® page, <https://www.zeiss.com/meditec/int/product-portfolio/retinal-cameras/visuscout-100-handheld-fundus-camera.html>

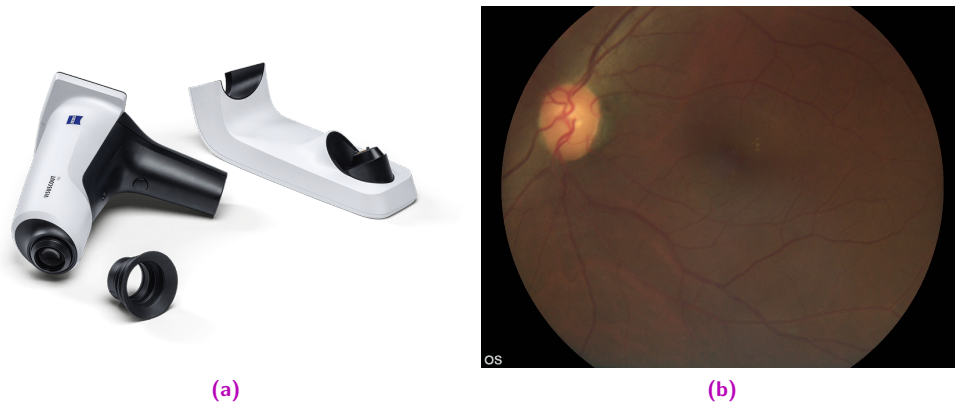


Fig. 3.6. (a) Example of a handheld fundus VISUSCOUT® camera²; (b) and an example fundus image captured using the camera.

images. In this experiment, the ultra-widefield imaging dataset does not contribute to the training process. Hence, it measures how well the narrow field data captures information that are useful to the widefield data without any knowledge about widefield imaging. The third experiment includes the usage of the ultra-widefield data in the fine-tuning process. After training the deep model on narrow field handheld data, the model is further optimized using part of the ultra-widefield data and evaluated on the leftover part of the data. The training task is to detect referable DR from the input fundus images.

The model trained from scratch on a small dataset of ultra-widefield images has the worst performance of the three models. The amount of data is not enough to build meaningful feature extractors that detect patterns in the data useful for referable DR detection. When using handheld narrow field fundus imaging device data for training and evaluating on the ultra-widefield data, it is clear the performance significantly improves on multiple measured metrics. When using the ultra-widefield data for the fine-tuning of the model, the performance on the ultra-widefield data further improves in the referable DR detection task. While this work uses larger ultra-widefield dataset compared with the work presented in Section 3.3.3, the size of the dataset is still considered small compared what is available in other imaging modalities. The usage of larger amounts of ultra-widefield data for the fine-tuning process is expected to produce better results.

3.3.5 Machine Learning Techniques for Ophthalmic Data Processing: A Review (IEEE JBHI 2020)

As machine learning techniques are becoming dominant in various fields where automation is appreciated, the medical domain is not an exception. Various works have utilized machine learning techniques to automate decision making in medical applications. Ophthalmic applications are one of the applications where recently automatic decision support powered by machine and deep learning is being researched and applied. With the fast rise of machine learning methods and their adoption to various ophthalmic applications it became daunting to grasp what are the latest advances and what are the future perspectives for the current approaches. Moreover, plenty of publicly available datasets are being used for developing

and evaluating new methods in the domain. It is highly appreciated to have an overview of the available datasets and how often they are used in the impactful publications of the domain. A review that contains latest advances in ophthalmic data analysis highlighting common diseases and applications, and discussing the limitations is of high value to readers of technical backgrounds who are interested in ophthalmic machine learning domain.

The author has published a review of the recent machine learning related publications that discuss ophthalmic data analysis applications. The included works in the review include the most impactful works in the past 4 years (of the publication date) that appeared in high impact journals or conferences. The included works are further expanded by works that cited impactful works and had a new methodology that has not been published before to the author's knowledge. At the end, more than 60 works are summarized in the review and the more impactful works are discussed in more details. The results of all the included works are compared when possible to have an overview of the current state of the art. Furthermore, 25 publicly available datasets are listed and categorized based on their application area.

The review is sectioned based on common disease in ophthalmology which are most studied in the current literature namely diabetic retinopathy, age-related macular degeneration and glaucoma. For each of the diseases the publications related to it are listed with a brief summary and some publications are discussed in more details. The applications related to each of the diseases are presented and publications are categorized depending on these sub-applications. Applications of diseases include detection, grading, lesion segmentation, and weak localization. Future directions and limitations of each of the applications are discussed at the end of each section. Moreover, applications that are beneficial to general ophthalmic image analysis and monitoring are also presented. The applications are retinal vessels segmentation based on fundus imaging and retinal layers segmentation based on optical coherence tomography imaging. Lastly, future directions of using machine learning in ophthalmology and promising new techniques are presented in a summary section.

Conclusion

4.1 Discussion

This thesis introduces multiple ideas and methodologies regarding representation learning for general purpose applications and medical applications. Specifically, the proposed learning algorithms focus on the representation part of the learning for disentanglement, explainability, fairness, and descriptiveness among others. The main contributions of the thesis can be summarized as follows:

- Proposing an adversarial disentanglement framework that focuses on producing better disentanglement and reconstruction fidelity by virtue of an adversary that learns to distinguish generated images from real images. The framework works well on medical datasets that contain finer details compared to toy datasets used for developing disentanglement applications.
- Introducing a disentangled fair representation learning framework that utilizes concepts of independence of generative factors to disentangle sensitive information from target information in the representation. The result is a disentangled representation that is agnostic to sensitive information yet performs well on target tasks.
- Utilizing multi-scale concepts and metric embedding learning techniques to develop a method for microaneurysms segmentation in fundus images. The method improves segmentation of microaneurysms and improves the detection of early diabetic retinopathy grades.
- Proposing a solution to the scarcity of ultra-widefield fundus images by using techniques of transfer learning to utilize the abundant knowledge from narrow field fundus imaging in building a good grading and detection system of diabetic retinopathy in ultra-widefield fundus images.

Regarding disentanglement of representation. The proposed method disentangles the components of a representation by enforcing independence constraints on components which leads to factorized representations. The method is evaluated on skin lesion medical dataset and performed well both in reconstruction performance and disentanglement performance. The method is compared with other state of the art methods on disentanglement, reconstruction and downstream performance and showed that adding an adversarial reconstruction constraints improves not only the reconstruction of the data but also the disentanglement and downstream task performance.

On the topic of disentangled and fair representations, the learning process is improved by separating the sensitive and target information rather than purging sensitive information from an omni-representation. Treating sensitive and target information as two generative factors of the data proved to enhance the mitigation of sensitive information leakage into the target representation. Moreover, the proposed constraints provide a good trade-off between sensitive and target performance which is better than or comparable to other state-of-the-art approaches on 7 datasets. The utility of fairness is studied by applying the proposed fairness method on medical datasets. It has been concluded that performing fairness for medical dataset should be done with caution with regards to the selected sensitive attributes. In some applications, sensitive information is not deciding factor of diagnosis which makes selecting such an information as sensitive information a desirable practice. However, selecting sensitive attributes that correlate with a diagnosis would negatively affect the performance on the prediction of the diagnosis which might be acceptable for general purpose applications but not for medical applications.

Regarding diabetic retinopathy detection and grading. The proposed microaneurysms segmentation method provided good segmentation of the lesions especially in cases where annotation of other lesions is not available or annotations of all instances of the lesion of interest are not provided. The method proves that detecting microaneurysms on multiple scales rather than one improves the detection accuracy. This is because microaneurysms appear at multiple sizes in fundus images and detecting them on one scale may limit the algorithm to one resolution around the lesion where either pixel or semantic information is compromised. Moreover, adding metric loss to the method further improves the detection accuracy of microaneurysms by building more descriptive representations of microaneurysms that allow better differentiation between microaneurysms lesions and other lesions. The multi-scale segmentation method is utilized for the detection and grading of diabetic retinopathy. Using microaneurysms information in the grading and detection of diabetic retinopathy improves the detection of early stages of the disease which is valuable for screening programs.

On transfer learning, the proposed method utilized the concepts of knowledge transfer to enhance performance of diabetic retinopathy grading and detection on ultra-widefield fundus images using narrow field fundus images. The transfer is done either from tabletop narrow field fundus cameras to tabletop ultra-widefield fundus camera or from a handheld narrow field fundus camera to tabletop ultra-widefield fundus camera. In both scenarios, the performance of the detection or grading of diabetic retinopathy is improved in comparison with using only the small available datasets of ultra-widefield fundus imaging. This suggests that transfer learning from narrow to wide field fundus imaging datasets gives good results in detecting and grading diabetic retinopathy and limits the need for collecting large datasets of ultra-widefield imaging to develop automatic detection applications.

4.2 Future Work

There are multiple future direction that could be applied to the proposed contribution in this thesis. One direction is to further improve the fairness disentanglement framework by liberating the method from the need of a target task to learn a disentanglement of sensitive and meaningful information. A suggestion for this next step would be using a reconstruction

objective as a target application where reconstructions of input data are optimized based on target representation only. Moreover, the orthogonality constraints proposed in the method are hard coded based on the means of the enforced prior. Imposing such constraints in automatic manner without arriving to trivial solutions would a valuable contribution. Regarding diabetic retinopathy, a future direction would be studying the progression of the disease over time to better estimate when an intervention is required. Another direction is the utilization of other imaging modalities for the detection and progression analysis of diabetic retinopathy where more retinal information is visible. Example of these modalities are optical coherence tomography and optical coherence tomography angiography.

Appendix

List of Authored and Co-authored Publications

2020

- [83] **Mhd Hasan Sarhan**, Shadi Albarqouni, Mehmet Yigitsoy, Nassir Navab, Abouzar Eslami. “Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations”. *IET Image Processing* 14.17, 2020.
- [89] **Mhd Hasan Sarhan**, Nassir Navab, Abouzar Eslami, Shadi Albarqouni. “Fairness by Learning Orthogonal Disentangled Representations”. *European Conference on Computer Vision*, 2020.
- [90] **Mhd Hasan Sarhan**, Nassir Navab, Abouzar Eslami, Shadi Albarqouni. “On the Fairness of Privacy-Preserving Representations in Medical Applications”. *MICCAI workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 2020.
- [88] **Mhd Hasan Sarhan**, M Ali Nasser, Daniel Zapp, Mathias Maier, Chris P Lohmann, Nassir Navab, Abouzar Eslami. “Machine Learning Techniques for Ophthalmic Data Processing: A Review”. *IEEE Journal of Biomedical and Health Informatics*, 24.12, 2020.
- [85] **Mhd Hasan Sarhan**, Angelina Covita, Sophia Yu, Susan Su, Krunalkumar Ramanbhai Patel, Dhivakar Kanagaraj, Sandipan Chakroborty, Mary Durbin, Niranchana Manivannan, Abouzar Eslami. “Diabetic retinopathy detection in widefield images by using transfer learning from handheld narrow field images”. *Investigative Ophthalmology & Visual Science (IOVS)*, 61.9 2020.
- [40] Mert Kayhan, Okan Köpüklü, **Mhd Hasan Sarhan**, Mehmet Yigitsoy, Abouzar Eslami, Gerhard Rigoll. “Deep Attention Based Semi-Supervised 2D-Pose Estimation for Surgical Instruments”. *International Conference on Pattern Recognition Workshops and Challenges*, 2020.
- [85] Helge Hecht, **Mhd Hasan Sarhan**, Vlad Popovici. “Disentangled Autoencoder for Cross-Stain Feature Extraction in Pathology Image Analysis”. *Applied Sciences*, 10.18, 2020.

2019

- [84] **Mhd Hasan Sarhan**, Shadi Albarqouni, Mehmet Yigitsoy, Nassir Navab, Abouzar Eslami. “Multi-scale Microaneurysms Segmentation Using Embedding Triplet Loss”. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, Shenzhen, China.

- [86] **Mhd Hasan Sarhan**, Abouzar Eslami, Nassir Navab, Shadi Albarqouni. “Learning Interpretable Disentangled Representations using Adversarial VAEs”. *MICCAI workshop on Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, 2019, Shenzhen, China.
- [91] **Mhd Hasan Sarhan**, Patricia Sha, Mike Chen, Mary K Durbin, Mehmet Yigitsoy, Abouzar Eslami. “Deep learning for automatic diabetic retinopathy grading of ultra-widefield fundus images”. *Investigative Ophthalmology & Visual Science (IOVS)*, 60.9, 2019.
- [87] **Mhd Hasan Sarhan**, Katherine Makedonsky, Meike Mack, Mary K Durbin, Mehmet Yigitsoy, Abouzar Eslami. “Deep learning for automatic diabetic retinopathy detection under multiple image quality levels”. *Investigative Ophthalmology & Visual Science (IOVS)*, 60.11, 2019.

Abstracts of Publications not Discussed in this Thesis

Deep Attention Based Semi-Supervised 2D-Pose Estimation for Surgical Instruments

Mert Kayhan, Okan Köpüklü, Mhd Hasan Sarhan, Mehmet Yigitsoy, Abouzar Eslami, Gerhard Rigoll

For many practical problems and applications, it is not feasible to create a vast and accurately labeled dataset, which restricts the application of deep learning in many areas. Semi-supervised learning algorithms intend to improve performance by also leveraging unlabeled data. This is very valuable for 2D-pose estimation task where data labeling requires substantial time and is subject to noise. This work aims to investigate if semi-supervised learning techniques can achieve acceptable performance level that makes using these algorithms during training justifiable. To this end, a lightweight network architecture is introduced and mean teacher, virtual adversarial training and pseudo-labeling algorithms are evaluated on 2D-pose estimation for surgical instruments. For the applicability of pseudo-labelling algorithm, we propose a novel confidence measure, total variation. Experimental results show that utilization of semi-supervised learning improves the performance on unseen geometries drastically while maintaining high accuracy for seen geometries. For RMIT benchmark, our lightweight architecture outperforms state-of-the-art with supervised learning. For Endovis benchmark, pseudo-labelling algorithm improves the supervised baseline achieving the new state-of-the-art performance.

International Conference on Pattern Recognition Workshops and Challenges (2020)

Disentangled Autoencoder for Cross-Stain Feature Extraction in Pathology Image Analysis

Helge Hecht, Mhd Hasan Sarhan, Vlad Popovici

A novel deep autoencoder architecture is proposed for the analysis of histopathology images. Its purpose is to produce a disentangled latent representation in which the structure and colour information are confined to different subspaces so that stain-independent models may be learned. For this, we introduce two constraints on the representation which are implemented as a classifier and an adversarial discriminator. We show how they can be used for learning a latent representation across haematoxylin-eosin and a number of immune stains. Finally,

we demonstrate the utility of the proposed representation in the context of matching image patches for registration applications and for learning a bag of visual words for whole slide image summarization.

Journal of Applied Sciences (2020)

Deep Learning for Automatic Diabetic Retinopathy Detection under Multiple Image Quality Levels

Mhd Hasan Sarhan, Katherine Makedonsky, Meike Mack, Mary K Durbin, Mehmet Yigitsoy, Abouzar Eslami.

Purpose. Deep learning techniques are showing promising results for automatic diabetic retinopathy (DR) screening on fundus images. We analyze the effect of image quality of images collected with a handheld fundus camera on the performance of a DR screening algorithm.

Methods. This retrospective study used 257 fovea-centered fundus images from 93 subjects collected using VISUSCOUT® 100 (ZEISS, Jena, Germany) handheld fundus camera. The images are annotated positive for DR if signs of DR are visible (mild or more in International Clinical Diabetic Retinopathy disease severity scale). Image quality is assessed using a subjective 1-5 scale (1-very poor; 2-poor; 3-fair; 4-good; 5-excellent). 233 images were annotated as negative for DR (healthy) and 24 were annotated as positive for DR. We used our Diabetic Retinopathy Deep Network grading model (DRDN) that is trained on 35k publicly available fundus images for grading the severity of DR. DRDN is used to classify VISUSCOUT images for Healthy vs DR by regarding a mild or more prediction as DR. No data from the same handheld device has been used for training the model. Sensitivity (Sn), specificity (Sp) and area under the curve of receiver operating characteristics curve (AUC) were reported for all images. Sn and Sp were used to analyze the effect of image quality on DR prediction performance.

Results. DRDN shows good prediction power for unseen images from handheld camera reaching AUC of 0.98 on images from all quality profiles. No false negatives were triggered, hence, Sn=1. The model gave 22 false positives with Sp=0.905, 95% Confidence Interval (CI) [0.866, 0.943].

Conclusion. A high number of false positives (Sp=0.8, 95% CI [0.656, 0.943]) occurred in very poor quality Images. For images of poor and above quality, we observe a consistent performance of the algorithm with less effect of image quality on the Sp. Using samples from fair quality and above gives the highest results with Sp=0.926, 95% CI [0.886, 0.966] which suggests that using images of above-poor quality would decrease the number of false positives in the DR screening.

Investigative Ophthalmology & Visual Science (2019)

Fairness by Learning Orthogonal Disentangled Representations

Mhd Hasan Sarhan^{1,2}, Nassir Navab^{1,3}, Abouzar Eslami², Shadi Albarqouni^{1,4}

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴Computer Vision Lab, ETH Zurich, Zurich, Switzerland

Summary. This publication proposes a novel method for enforcing fairness on learned representations by disentangling sensitive and target information as two independent generative factors of the data. Target information represent a downstream task while sensitive information represent attributes the model is wished to be invariant towards. Separating the two representations is done by enforcing various constraints to reach a target representation that contains little information about sensitive attributes while retaining information regarding target attributes. This is useful for applications where fairness can be an issue (*e.g.* In justice system or loan granting related applications) or in cases where invariance towards a nuisance is required. The target and sensitive representations are enforced to be orthogonal as proxy for the independence constraints. Moreover, the target representation is forced to be agnostic towards sensitive information by using entropy maximization loss where the entropy of sensitive information is maximized given the target representation. Since it has been seen that more disentangled representations have better fairness scores, the components of the representations are disentangled using a Gaussian prior with a diagonal covariance matrix. The proposed framework is evaluated on visual and tabular data and proven to be better or comparable to state-of-the-art methods. The evaluation done by using two metrics, first is the performance of the target representation on the target task. This metric is used to measure how much target information is retained in the target representation after mitigating sensitive information leakage into the representation. Second metric is the performance of the target representation on the sensitive information detection task. This measures the leakage of sensitive information into the target representation, the more the target representation is informed about sensitive attributes the less fair it is. Ideally it should be as good as naively guessing the majority class in sensitive attributes. The proposed method had a good trade-off of sensitive and target performance compared to other state-of-the-art methods. This shows that separating sensitive and target information as two generative factors of the data leads to good fairness performance and trade-off without requiring adversarial training.

Contribution. The contribution of the author of this thesis to this publication consists of the main idea, implementation, experiment design and evaluation. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details in implementation as well as the revision of the publication was done jointly with the co-authors.

Copyright Statement. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer, Cham, Lecture Notes in Computer (LNCS, volume 12374), 16th European Conference on Computer Vision, August 2020, Fairness by Learning Orthogonal Disentangled Representations, Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, Shadi Albarqouni, © 2020.

The following text is reprinted with the permission of the publisher (License number 5995281073461). It is the accepted but not the published version of the paper. The license is attached after the publication text.

Fairness by Learning Orthogonal Disentangled Representations

Mhd Hasan Sarhan^{1,2}[0000–0003–0473–5461], Nassir Navab^{1,3}, Abouzar Eslami¹[0000–0001–8511–5541], and Shadi Albarqouni^{1,4}[0000–0003–2157–2211]

¹ Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

`hasan.sarhan@tum.de`

² Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴ Computer Vision Lab, ETH Zurich, Switzerland

Abstract. Learning discriminative powerful representations is a crucial step for machine learning systems. Introducing invariance against arbitrary nuisance or sensitive attributes while performing well on specific tasks is an important problem in representation learning. This is mostly approached by purging the sensitive information from learned representations. In this paper, we propose a novel disentanglement approach to invariant representation problem. We disentangle the meaningful and sensitive representations by enforcing orthogonality constraints as a proxy for independence. We explicitly enforce the meaningful representation to be agnostic to sensitive information by entropy maximization. The proposed approach is evaluated on five publicly available datasets and compared with state of the art methods for learning fairness and invariance achieving the state of the art performance on three datasets and comparable performance on the rest. Further, we perform an ablative study to evaluate the effect of each component.

Keywords: Representation learning, disentangled representation, fairness in machine learning

1 Introduction

Learning representations that are useful for downstream tasks yet robust against arbitrary nuisance factors is a challenging problem. Automated systems powered by machine learning techniques are corner stones for decision support systems such as granting loans, advertising, and medical diagnostics. Deep neural networks learn powerful representations that encapsulate the extracted variations in the data. Since these networks learn from historical data, they are prone to represent the past biases and the learnt representations might contain information that were not intended to be released. This has raised various concerns regarding fairness, bias and discrimination in statistical inference algorithms [17]. The European union has recently released their "Ethics guidelines for trustworthy

AI” report ⁵ where it is stated that unfairness and biases must be avoided.

Since a few years, the community has been investigating to learn a latent representation $\mathbf{z}$ that well describes a target observed variable $\mathbf{y}$ (e.g. Annual salary) while being robust against a sensitive attribute $\mathbf{s}$ (e.g. Gender or race). This nuisance could be independent from the target task which is termed as a domain adaptation problem. One example is the identification of faces $\mathbf{y}$ regardless of the illumination conditions $\mathbf{s}$. In the other case termed fair representation learning $\mathbf{s}$ and $\mathbf{y}$ are not independent. This could be the case with $\mathbf{y}$ being the credit risk of a person while $\mathbf{s}$ is age or gender. Such relation between these variables could be due to past biases that are inherently in the data. This independence is assumed to hold when building fair classification models. Although this assumption is over-optimistic as these factors are probably not independent, we wish to find a representation $\mathbf{z}$ that is independent from $\mathbf{s}$ which justifies the usage of such a prior belief [18]. This is mostly approached by approximations of mutual information scores between $\mathbf{z}$ and $\mathbf{s}$ and force the two variables to minimize this score either in an adversarial [25, 16] or non-adversarial [14, 18] manner. These methods while performing well on various datasets, are still limited by either convergence instability problems in case of adversarial solutions or hindered performance compared to the adversarial counterpart. Learning disentangled representations has been proven to be beneficial to learning fairer representations compared to general purpose representations [13]. We use this concept to disentangle the components of the learned representations. Moreover, we treat the $\mathbf{s}$ and $\mathbf{y}$ as separate independent generative factors and decompose the learned representation in such a way that each representation holds information related to the respective generative factor. This is achieved by enforcing orthogonality between the representations as a relaxation for the independence constraint. We hypothesize that decomposing the latent code into target code $\mathbf{z}_T$ and residual sensitive $\mathbf{z}_S$ code would be beneficial for limiting the leakage of sensitive information into $\mathbf{z}_T$ by redirecting it to $\mathbf{z}_S$ while keeping it informative about some target task that we are interested in.

We propose a framework for learning invariant fair representations by decomposing learned representations into target and residual/sensitive representations. We impose disentanglement on the components of each code and impose orthogonality constraint on the two learned representations as a proxy for independence. The learned target representation is explicitly enforced to be agnostic to sensitive information by maximizing the entropy of sensitive information in $\mathbf{z}_T$.

Our contributions are three-folds:

- Decomposition of target and sensitive data into two orthogonal representations to promote better mitigation of sensitive information leakage.
- Promote disentanglement property to split hidden generative factors of each learned code.

⁵ Ethics guidelines for trustworthy AI, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

- Enforce the target representation to be agnostic of sensitive information by maximizing the entropy.

2 Related work

Learning fair and invariant representations has a long history. Earlier strategies involved changing the examples to ensure fair representation of the all groups. This relies on the assumption that equalized opportunities in the training set would generalize to the test set. Such techniques are referred to as data massaging techniques [9, 19]. These approaches may suffer of under-utilization of data or complications on the logistics of data collection. Later, Zemel *et al.* [26] proposed a semi-supervised fair clustering technique to learn a representation space where data points are clustered such that each cluster contains similar proportions of the protected groups. One drawback is that the clustering constraint limits the power of a distributed representation. To solve this, Louizos *et al.* [14] presented the Variational Fair Autoencoder (VFAE) where a model is trained to learn a representation that is informative enough yet invariant to some nuisance variables. This invariance is approached through Maximum Mean Discrepancy (MMD) penalty. The learned sensitive-information-free representation could be later used for any subsequent processing such as classification of a target task. After the success of Generative Adversarial Networks (GANs) [7], multiple approaches leveraged this learning paradigm to produce robust invariant representations [25, 27, 5, 16]. The problem setup in these approaches is a minimax game between an encoder that learns a representation for a target task and an adversary that extracts sensitive information from the learned representation. In this case, the encoder minimizes the negative log-likelihood of the adversary while the adversary is forced to extract sensitive information alternatively. While methods relying on adversarial zeros-sum game of negative log-likelihood minimization and maximization perform well in the literature, they sometimes suffer from convergence problems and require additional regularization terms to stabilize the training. To overcome these problems, Roy *et al.* [21] posed the problem as an adversarial non-zero sum game where the encoder and discriminator have competing objectives that optimize for different metrics. This is achieved by adding an entropy loss that forces the discriminator to be un-informed about sensitive information. It is worth noting that it is argued by [18] that adversarial training for fairness and invariance is unnecessary and sometimes leads to counter productive results. Hence, they approximated the mutual information between the latent representation and sensitive information using a variational upper bound. Lastly, Creager *et al.* [3] proposed a fair representation learning model by disentanglement, their model has the advantage of flexibly changing sensitive information at test time and combine multiple sensitive attributes to achieve subgroup fairness. In their work, independence is enforced adversarially by utilizing a discriminator to distinguish simulated independent representations (fake) from learned representations (real).

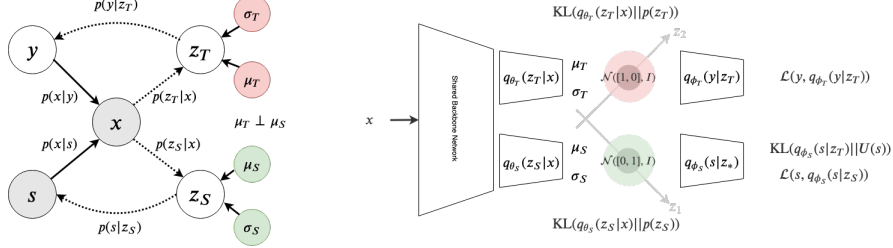


Fig. 1: Left: The graphical model of our proposed method. Right: Our framework encode the input data to intermediate target and residual (sensitive) representations, parameterized by μ and σ . Samples from the estimated posteriors are fed to the discriminators to predict the target and sensitive labels.

3 Methodology

let $\mathcal{X}$ be the dataset of individuals from all groups and $\mathbf{x} \in \mathbb{R}^D$ be an input sample. Each input is associated with a target attribute $\mathbf{y} = \{y_1, \dots, y_n\} \in \mathbb{R}^n$ with n classes, and a sensitive attribute $\mathbf{s} = \{s_1, \dots, s_m\} \in \mathbb{R}^m$ with m classes. Our goal is to learn an encoder that maps input $\mathbf{x}$ to two low-dimensional representations $\mathbf{z}_T \in \mathbb{R}^{d_T}$, $\mathbf{z}_S \in \mathbb{R}^{d_S}$. Ideally $\mathbf{z}_T$ must contain information regarding target attribute while mitigating leakage about the sensitive attribute and $\mathbf{z}_S$ contains residual information that is related to the sensitive attribute.

3.1 Fairness definition

One of the common definition of fairness that has been proposed in the literature [25, 21, 20, 1] is simply requiring the sensitive information to be statistically independent from the target. Mathematically, the prediction of a classifier $p(\mathbf{y}|\mathbf{x})$ must be independent from the sensitive information, which is expressed as follows

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, \mathbf{s}) \quad (1)$$

For example, in the German credit dataset, we need to predict the credit behaviour of the bank account holder regardless the sensitive information, such as gender, age ...etc. In other words, $p(\mathbf{y} = \text{good credit risk}|\mathbf{x}, \mathbf{s} = \text{male})$ should be equal to $p(\mathbf{y} = \text{good credit risk}|\mathbf{x}, \mathbf{s} = \text{female})$. The main objective is to learn fair data representations that are i) informative enough for the downstream task, and ii) independent from the sensitive information.

3.2 Problem Formulation

To promote the independence of the generative factors, *i.e.* target and sensitive information, we aim to maximize the log likelihood of the conditional distribution

$\log p(\mathbf{y}, \mathbf{s}|\mathbf{x})$, given the fairness assumption in Eq. 1

$$p(\mathbf{y}, \mathbf{s}|\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{s})p(\mathbf{x}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{x})} = p(\mathbf{s}|\mathbf{x})p(\mathbf{y}|\mathbf{x}) \quad (2)$$

To enforce our aforementioned conditions, we let our model $f(\cdot)$ encode the observed input data $\mathbf{x}$ into target $\mathbf{z}_T$ and residual $\mathbf{z}_S$ intermediate representations on which the independence constraints are applied,

$$p(\mathbf{y}, \mathbf{s}|\mathbf{x}) = \underbrace{p(\mathbf{y}|\mathbf{z}_T)}_{L_T} \underbrace{p(\mathbf{z}_T|\mathbf{x})}_{L_{z_T}} \underbrace{p(\mathbf{s}|\mathbf{z}_S)}_{L_S} \underbrace{p(\mathbf{z}_S|\mathbf{x})}_{L_{z_S}} \quad (3)$$

losses in Sec. 3.3 correspond to terms in Eq. 3 which are shown in the under brackets. The log likelihood is maximized given the following constraints; (i) $p(\mathbf{z}_S|\mathbf{x})$ is statistically independent from $p(\mathbf{z}_T|\mathbf{x})$, and (ii) $\mathbf{z}_T$ is agnostic to sensitive information $\mathbf{s}$. Our objective function J can be written as

$$J = -\log p(\mathbf{y}, \mathbf{s}|\mathbf{x}) \text{ s.t. } \text{MI}(\mathbf{z}_T, \mathbf{z}_S) = 0 \text{ and } \text{KL}(p(\mathbf{s}|\mathbf{z}_T), \mathcal{U}) = 0 \quad (4)$$

where $\mathcal{U}(\mathbf{s})$ is the uniform distribution.

3.3 Fairness by Learning Orthogonal and Disentangled Representations

As depicted in Fig. 1, our observed data $\mathbf{x}$ is fed to a shared encoder $f(\mathbf{x}; \theta)$, then projected into two subspaces producing our target, and residual (sensitive) representations using the encoders; $q_{\theta_T}(\mathbf{z}_T|\mathbf{x})$, and $q_{\theta_S}(\mathbf{z}_S|\mathbf{x})$, respectively, where θ is shared parameter, *i.e.* $\theta = \theta_S \cap \theta_T$. Each representation is fed to the corresponding discriminator; target discriminator, $q_{\phi_T}(\mathbf{y}|\mathbf{z}_T)$, and sensitive discriminator $q_{\phi_S}(\mathbf{s}|\mathbf{z}_S)$. Both discriminators and encoders are trained in supervised fashion to minimize the following loss,

$$\mathcal{L}_T(\theta_T, \phi_T) = \text{KL}(p(\mathbf{y}|\mathbf{x}) \parallel q_{\phi_T}(\mathbf{y}|\mathbf{z}_T)), \quad (5)$$

$$\mathcal{L}_S(\theta_S^*, \phi_S) = \text{KL}(p(\mathbf{s}|\mathbf{x}) \parallel q_{\phi_S}(\mathbf{s}|\mathbf{z}_S)), \quad (6)$$

where $\theta_S^* = \theta_S \setminus \theta$.

To ensure that our target representation does not encode any leakage of the sensitive information, we follow Roy *et al.* [21] in maximizing the entropy of the sensitive discriminator given the target representation $q_{\phi_S}(\mathbf{s}|\mathbf{z}_T)$ as

$$\mathcal{L}_E(\phi_S, \theta_T) = \text{KL}(q_{\phi_S}(\mathbf{s}|\mathbf{z}_T) \parallel \mathcal{U}(\mathbf{s})). \quad (7)$$

We relax the independence assumption by enforcing i) disentanglement property, and ii) the orthogonality of the corresponding representations.

To promote the (i) disentanglement property on the target representation, we first need to estimate the distribution $p(\mathbf{z}_T|\mathbf{x})$ and enforce some sort of independence among the latent factors,

$$p(\mathbf{z}_T|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z}_T)p(\mathbf{z}_T)}{p(\mathbf{x})}, \text{ s.t. } p(\mathbf{z}_T) = \prod_{i=1}^{N_T} p(z_T^i). \quad (8)$$

Since $p(\mathbf{z}_T|\mathbf{x})$ is intractable, we employ the Variational Inference, thanks to the re-paramterization trick [11], and let our model output the distribution parameters; $\boldsymbol{\mu}_T$, and $\boldsymbol{\sigma}_T$, and minimize the KL-divergence between posterior $q_{\theta_T}(\mathbf{z}_T|\mathbf{x})$ and prior $p(\mathbf{z}_T)$ distributions as

$$\mathcal{L}_{\mathbf{z}_T}(\theta_T) = \text{KL}(q_{\theta_T}(\mathbf{z}_T|\mathbf{x}) \parallel p(\mathbf{z}_T)), \quad (9)$$

where $p(\mathbf{z}_T) = \prod_{i=1}^{N_T} p(z_T^i) = \mathcal{N}(\mathbf{0}, I)$, and $q_{\theta_T}(\mathbf{z}_T|\mathbf{x}) = \mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}_T, \text{diag}(\boldsymbol{\sigma}_T^2))$. Similarly, we enforce the same constraints on the residual (sensitive) representation $\mathbf{z}_S$ and minimize the KL-divergence as $\mathcal{L}_{\mathbf{z}_S}(\theta_S) = \text{KL}(q_{\theta_S}(\mathbf{z}_S|\mathbf{x}) \parallel p(\mathbf{z}_S))$.

To enforce the (ii) orthogonality between the target and residual (sensitive) representations, *i.e.* $\boldsymbol{\mu}_S \perp \boldsymbol{\mu}_T$, we hard code the means of the prior distributions to orthogonal means. In this way, we implicitly enforce the weight parameters to project the representations into orthogonal subspaces. To illustrate this in 2-dimensional space, we set the prior distributions to $p(\mathbf{z}_S) = \mathcal{N}([0, 1]^T, I)$, and $p(\mathbf{z}_T) = \mathcal{N}([1, 0]^T, I)$ (*cf.* Fig. 1).

To summarize, an additional loss term is introduced to the objective function promoting both Orthogonality and Disentanglement properties, denoted *Orthogonal-Disentangled* loss,

$$\mathcal{L}_{OD}(\theta_T, \theta_S) = \mathcal{L}_{\mathbf{z}_T}(\theta_T) + \mathcal{L}_{\mathbf{z}_S}(\theta_S). \quad (10)$$

A variant of this loss without the property of orthogonality, denoted *Disentangled* loss, is also introduced for the purpose of ablative study (See Sec. 4.3).

3.4 Overall objective function

To summarize, our overall objective function is

$$\arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S) \quad (11)$$

where λ_E , and λ_{OD} are hyper-parameters to weigh the *Entropy* loss and the *Orthogonal-Disentangled* loss, respectively. A sensitivity analysis on the hyper-parameters is presented in Sec. 4.5.

4 Experiments

In this section, the performance of the learned representations using our method will be evaluated and compared against various state of the art methods in the

Algorithm 1 Learning Orthogonal Disentangled Fair Representations

Require: Maximum Epochs E_{max} , Step size t_s , λ_{OD} , λ_E , γ_{OD} , γ_E , $p(\mathbf{z}_T)$, $p(\mathbf{z}_S)$

Ensure: $\mathbf{z}_S \perp \mathbf{z}_T$

Initialize: $\theta_T, \theta_S, \phi_T, \phi_S \leftarrow \theta_T^{(0)}, \theta_S^{(0)}, \phi_T^{(0)}, \phi_S^{(0)}$

for $t = 1, 2, \dots, E_{max}$ **do**

$[\mu_T, \sigma_T] = q_{\theta_T}(\mathbf{z}_T | \mathbf{x})$

$[\mu_S, \sigma_S] = q_{\theta_S}(\mathbf{z}_S | \mathbf{x})$

sample $\mathbf{z}_T \sim \mathcal{N}(\mu_T, \text{diag}(\sigma_T^2))$

sample $\mathbf{z}_S \sim \mathcal{N}(\mu_S, \text{diag}(\sigma_S^2))$

compute $\mathcal{L}_{z_T}(\theta_T) = \text{KL}(q_{\theta_T}(\mathbf{z}_T | \mathbf{x}) \parallel p(\mathbf{z}_T))$

compute $\mathcal{L}_{z_S}(\theta_S) = \text{KL}(q_{\theta_S}(\mathbf{z}_S | \mathbf{x}) \parallel p(\mathbf{z}_S))$

compute $\mathcal{L}_T(\theta_T, \phi_T) = -\sum p(\mathbf{y} | \mathbf{x}) \log[q_{\phi_T}(\mathbf{y} | \mathbf{z}_T)]$

compute $\mathcal{L}_S(\theta_S^*, \phi_S) = -\sum p(\mathbf{s} | \mathbf{x}) \log[q_{\phi_S}(\mathbf{s} | \mathbf{z}_S)]$

compute $\mathcal{L}_E(\phi_S, \theta_T) = \sum q_{\phi_S}(\mathbf{s} | \mathbf{z}_T) \log[q_{\phi_S}(\mathbf{s} | \mathbf{z}_T)]$

update $\lambda_{OD} \leftarrow \lambda_{OD} \gamma_{OD}^{t/t_s}$

update $\lambda_E \leftarrow \lambda_E \gamma_E^{t/t_s}$

$\mathcal{L}_{OD}(\theta_T, \theta_S) = \mathcal{L}_{z_T}(\theta_T) + \mathcal{L}_{z_S}(\theta_S)$

$J(\theta_T, \theta_S, \phi_T, \phi_S) = \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S)$

update $\theta_T, \theta_S, \phi_T, \phi_S \leftarrow \arg \min J(\theta_T, \theta_S, \phi_T, \phi_S)$

end for

return $\theta_T, \theta_S, \phi_T, \phi_S$

domain. First, we present the experimental setup by describing the five datasets used for validation, the model implementation details for each dataset, and design of the experiments. We then compare the model performance with state of the art fair representation models on the datasets. We perform an ablative study to monitor the effect of each added component on the overall performance. We then evaluate the models qualitatively by showing t-SNE projections of the learned representations. Lastly, we perform a sensitivity analysis to study the effect of hyper-parameters on the training.

4.1 Experimental Setup

Tabular data: For evaluating fair classification, we use two datasets from the UCI repository [4], namely, the German and the Adult datasets. The German credit dataset consists of 1000 samples each with 20 attributes, and the target task is to classify a bank account holder having good or bad credit risk. The sensitive attribute is the gender of the bank account holder. The adult dataset contains 45,222 samples each with 14 attributes. The target task is a binary classification of annual income being more or less than \$50,000 and again gender is the sensitive attribute.

Visual data: To examine the model learned invariance on visual data, we have used the application of illumination invariant face classification. Ideally, we want the representation to contain information about the subject’s identity without

holding information regarding illumination direction. For this purpose, the extended YaleB dataset is used [6]. The dataset contains the face images of 38 subjects under five different light source direction conditions (upper right, lower right, lower left, upper left, and front). The target task is the identification of the subject while the light source condition is considered the sensitive attribute.

CIFAR data: Following Roy *et al.* [21], we have created a binary target task from CIFAR-10 dataset [12]. The original dataset contains 10 classes we refer to as fine classes, we divide the 10 classes into two categories living and non-living classes and refer to this split as coarse classes. It is expected that living objects have common visual proprieties that differ from non-living ones. The target task is the classification of the coarse classes while not revealing information about the fine classes. With a similar concept, we divide the 100 fine classes of CIFAR-100 dataset into 20 coarse classes that cluster similar concepts into one category. For example, the coarse class 'aquatic mammals' contains the fine classes 'beaver', 'dolphin', 'otter', 'seal', and 'whale'. For the full details of the split, the reader is referred to [21] or the supplementary materials of this manuscript. The target task for CIFAR-100 is the classification of the coarse classes while mitigating information leakage regarding the sensitive fine classes.

Implementation details: For the Adult and German datasets, we follow the setup appeared in [21] by having a 1-hidden-layer neural network as encoder, the discriminator has two hidden layer and the target predictor is a logistic regression layer. Each hidden layer contains 64 units. The size of the representation is 2. The learning rate for all components is 10^{-3} and weight decay is 5×10^{-4} . For the Extended YaleB dataset, we use an experimental setup similar to Xie *et al.* [25] and Louizos *et al.* [14] by using the same train/test split strategy. We used $38 \times 5 = 190$ samples for training and 1096 for testing. The model setup is similar to [25, 21], the encoder consisted of one layer, target predictor is one linear layer and the discriminator is neural network with two hidden layers each contains 100 units. The parameters are trained using Adam optimizer with a learning rate of 10^{-4} and weight decay of 5×10^{-2} . Similar to [21], we employed ResNet-18 [8] architecture for training the encoder on the two CIFAR datasets. For the discriminator and target classifiers, we employed a neural network with two hidden layers (256 and 128 neurons). For the encoder, we set the learning rate to 10^{-4} and weight decay to 10^{-2} . For the target and discriminator networks, the learning rate and weight decay were set to 10^{-2} and 10^{-3} , respectively. Adam optimizer [10] is used in all experiments.

Experiments design: We address two questions in the experiments. First, is *how much information about the sensitive attributes is retained in the learned representation $\mathbf{z}_T$* ? Ideally, $\mathbf{z}_T$ would not contain any sensitive attribute information. This is evaluated by training a classifier with the same architecture as the discriminator network on sensitive attributes classification task. The closer the accuracy to a naive majority label predictor, the better the model is. This classifier is trained with $\mathbf{z}_T$ as input after the encoder, target, and discriminator had

Table 1: Results on CIFAR-10 and CIFAR-100 datasets.

	CIFAR-10		CIFAR-100	
	Target Acc. $\uparrow$	Sensitive Acc. $\downarrow$	Target Acc. $\uparrow$	Sensitive Acc. $\downarrow$
Baseline	0.9775	0.2344	0.7199	0.3069
Xie et al. [25] (trade-off #1)	0.9752	0.2083	0.7132	0.1543
Roy et al. [21] (trade-off #1)	0.9778	0.2344	0.7117	0.1688
Xie et al. [25] (trade-off #2)	0.9735	0.2064	0.7040	0.1484
Roy et al. [21] (trade-off #2)	0.9679	0.2114	0.7050	0.1643
Ours	0.9725	0.1907	0.7074	0.1447

been trained and freezed. Second, is *how well the learned representation $\mathbf{z}_T$ performs in identifying target attributes?*. To this end, we train a classifier similar to the target on the learned representation $\mathbf{z}_T$ to detect the target attributes. We also visualize the representations $\mathbf{z}_T$ and $\mathbf{z}_S$ by using their t-SNE projections to show how the learned representations describe target attributes while being agnostic to the sensitive information.

4.2 Comparison with state of the art

We compare the proposed approach against various state of the art methods on the five presented datasets. We first train the model with Algorithm 1 while changing hyper-parameters between runs. We choose the best performing model in terms of the trade-off between target and sensitive classification accuracy based on $\mathbf{z}_T$. We then compare it with various state of the art methods for sensitive information leakage and retaining target information.

CIFAR datasets: We compare the proposed approach with two other state of the art methods on the CIFAR-10 and CIFAR-100 datasets, namely Xie *et al.* [25] and Roy *et al.* [21]. We examine two different trade-off points of both approaches. The first trade-off point is the one with best target accuracy reported by the model while the second trade-off point is the one with the target accuracy closest to ours for a more fair comparison. The lower the target accuracy in the trade-off the better (lower) the sensitive accuracy is. We can see when the target accuracies are comparable, our model performs better in preventing sensitive information leakage to the representation $\mathbf{z}_T$. Hence, the proposed method has a better trade-off on the target and sensitive accuracy for both CIFAR-10 and CIFAR-100 datasets. However, the peak target performance is comparable but lower than the peak target performance of the studied methods.

Extended YaleB dataset: For the illumination invariant classification task on the extended YaleB dataset, the proposed method is compared with the logistic regression baseline (LR), variational fair autoencoder VFAE [14], Xie *et al.* [25] and Roy *et al.* [21]. The results are shown in Fig. 2 on the right hand side. The proposed model performs best on the target attribute classification while having

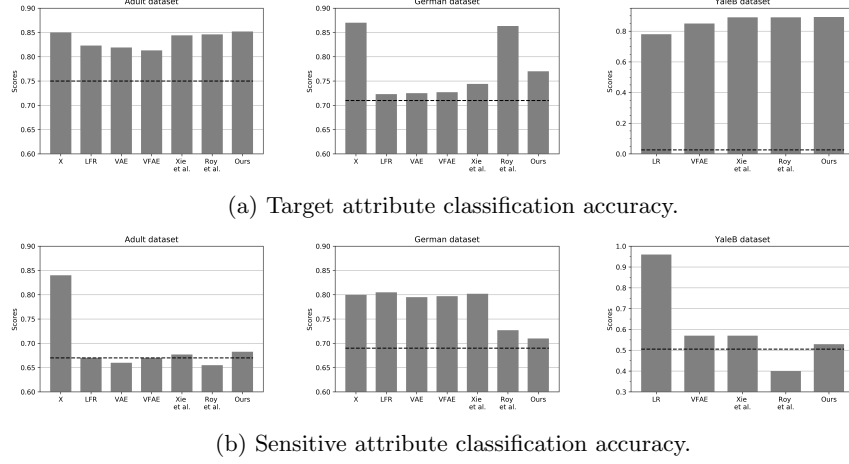


Fig. 2: Results on Adult, German, and extended YaleB datasets. The dashed black line represent a naive majority classifier that predicts the majority label.

the closest performance to the majority classification line (dashed line in Fig. 2). The majority line is the trivial baseline of predicting the majority label. The closer the sensitive accuracy to the majority line the better the model is in hiding sensitive information from $\mathbf{z}_T$. This means the learned representation is powerful at identifying subject in the images regardless of illumination conditions. To assess this visually, refer to Sec. 4.4 for qualitative analysis.

Tabular datasets: On the Adult and German datasets, we compare with LFR [26], vanilla VAE [11], variational fair autoencoder [14], Xie *et al.* [25] and Roy *et al.* [21]. The results of these comparisons are shown in Fig. 2. On the German dataset, we observe a very good performance in hiding sensitive information with 71% accuracy compared to 72.7% in [21]. On the target task, the model performs well compared to other models except for [21] which does marginally better than the rest. On the Adult dataset, our proposed model performs better than the aforementioned models on the target task while leaking slightly more information compared to other methods and the majority line at 67%. Our method has 68.26% sensitive accuracy while LFR, VAE, vFAE, Xie *et al.*, and Roy *et al.* have 67%, 66%, 67%, 67.7%, and 65.5% sensitive accuracy, respectively.

Generally, we observe that the proposed model performs well on all datasets with state of the art performance on visual datasets (CIFAR-10, CIFAR-100, YaleB). This suggests that such a model could lead to more fair/invariant representation without large sacrifices on downstream tasks.

4.3 Ablative study

In this section, we evaluate the contributions provided in the paper by eliminating parts of the loss function and study how each part affects the training in terms of target and sensitive accuracy. To this end, we used the best performing models after hyper-parameter search when training for all contributions for each dataset. The models are trained with the same settings and architectures described in Sec. 4.1. We compare five different variations for each model alongside the baseline classifier:

1. **Baseline:** Training a deterministic classifier for the target task and evaluate the information leakage about the sensitive attribute.
2. **Entropy w/o KL:** Entropy loss $\mathcal{L}_E$ is incorporated (Equation 7) in the loss while $\mathcal{L}_{OD}$ is not included (Equation 10).
3. **KL Orth. w/o Entropy:** Entropy loss $\mathcal{L}_E$ is not used (Equation 7) while $\mathcal{L}_{OD}$ is used for target and sensitive representations with orthogonal means (Equation 10).
4. **w/o Entropy w/o KL:** Neither entropy loss nor KL divergence are used in the loss. This case is similar to multi-task learning with the tasks being the classification of target and sensitive attributes.
5. **Entropy + KL w/o Orth.:** Entropy loss $\mathcal{L}_E$ is used and *disentangled loss* is used with similar means. Hence, there might be some disentanglement of generative factors in the components of each latent code but no constraints are applied to force disentanglement of the two representations.
6. **Entropy + KL Orth.:** All contributions are included.

The results of the ablative study are shown in Figure 3.

- For the *sensitive class accuracy*, it is desirable to have a lower accuracy in distinguishing sensitive attributes. Compared to the baseline, we observe that adding entropy loss and orthogonality constraints on the representations lowers the discriminative power of the learned representation regarding sensitive information. This is valid on all studied datasets except for CIFAR-10 where orthogonality constraint without entropy produced better representations for hiding sensitive information with a small drop (0.26%) on the target task performance. In the rest of the cases, having either entropy loss or KL loss only does not bring noticeable performance gains compared to a multi-task learning paradigm. This could be attributed to the fact that orthogonality on its own does not enforce independence of random variables and another constraint is needed to encourage independent latent variables (*i.e.* entropy loss).
- Comparing baseline with **w/o Entropy w/o KL** case answers the important question "*Does multi-task learning with no constraints on representations bring any added value in mitigating sensitive information leakage?*". In three out of the five studied datasets, it is the case. We can see lower accuracy in identifying sensitive information by using the learned target representation as input to a classifier while having no constraints on the relationship

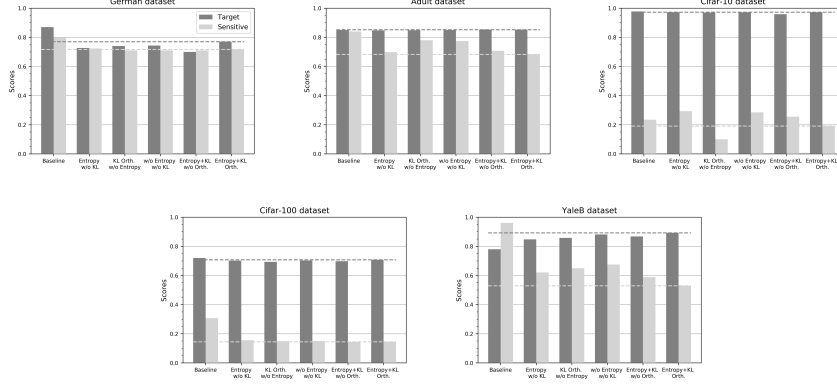


Fig. 3: Ablative study. Dark gray and light gray dashed lines represent the accuracy results on the target and sensitive task respectively for the "Entropy + KL Orth." model.

between the sensitive and target representations during the training process of the encoder. Simply, adding an auxiliary classifier to the target classifier and force it to learn information about sensitive attributes hides some sensitive data from the target classifier.

- Regarding **target accuracy**, the proposed model does not suffer from large drops in target performance when disentangling target from sensitive information. This could be seen by comparing target accuracy between the baseline and **Entropy+KL Orth.** columns. The largest drop in target performance compared to no privacy baseline is seen on the German dataset. This could be because of the very high dependence between gender and granting good or bad credit to a subject in the dataset and the small amount of subjects in the dataset.

4.4 Qualitative analysis

We visualize the learned embeddings using t-SNE [15] projections for the extended YaleB and CIFAR-10 datasets (*cf.* Fig. 4). We use the image space, $\mathbf{z}_T$, $\mathbf{z}_S$ as inputs to the projection to visualize what type of information is held within each representation. We also show the label of each image with regards to the target task to make it easier to investigate the clusters. For the extended YaleB, we see that, using the image space $\mathbf{x}$, the images are clustered mostly depending on their illumination conditions. However, when using $\mathbf{z}_T$, the images are not clustered according lighting conditions but rather, mostly based on the subject identity. Moreover, the visualization of representation $\mathbf{z}_S$ shows that the representation contains information about the sensitive class. For the CIFAR-10 dataset, using the image space basically clusters the images on the dominant

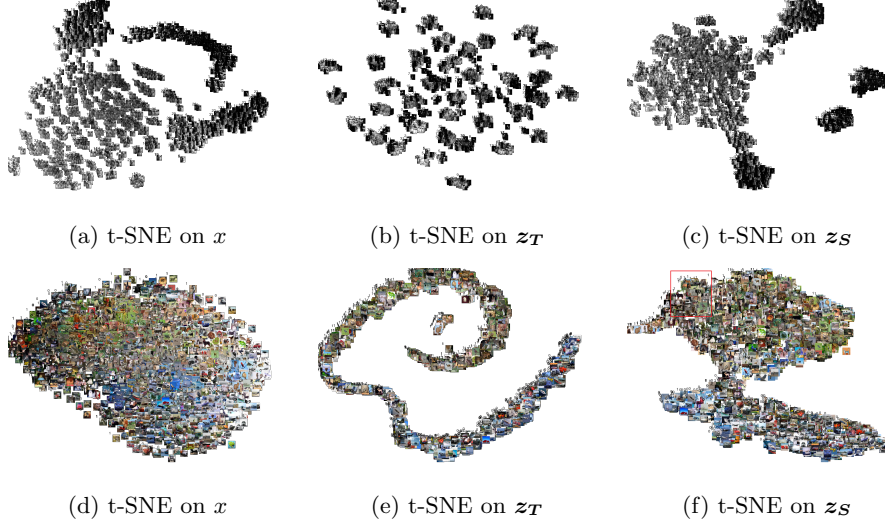


Fig. 4: t-SNE visualization of the extended YaleB faces (top) and CIFAR-10 (bottom) images. Figure is better seen in color and high resolution.

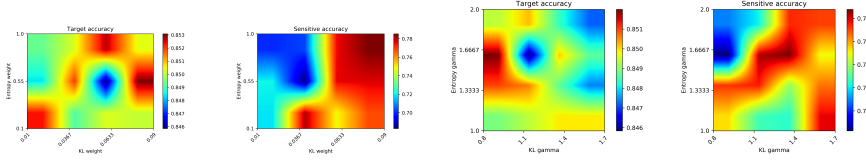


Fig. 5: Sensitivity analysis on the Adult dataset

color. When using z_T , it is clear that the target information is separated where the right side represent the non-living objects, and the left to inside part represents the living objects. What should be observed in z_T , is that within each target class, the fine classes are mixed and indistinguishable as we see cars, boats and trucks mixed in the right hand side of the figure, for example. The representation z_S has some information about the target class and also has the residual information about the fine classes as we see in the annotated red rectangle. A group of horses images are clustered together, then few dogs' images are clustered under it, then followed by birds. This shows that z_S has captured some sensitive information while z_T is more agnostic to the sensitive fine classes.

4.5 Sensitivity analysis

To analyze the effect of hyper-parameters choices on the sensitive and target accuracy, we show heatmaps of how the performance changes when the studied hyper-parameters are changed. The investigated hyper-parameters are KL

weight (λ_{OD}), Entropy Weight (λ_E), KL gamma (γ_{OD}), and Entropy gamma (γ_E). We show the results on the Adult dataset. We can see that the sensitive accuracy is sensitive to λ_{OD} more than λ_E as changes in λ_E do not induce much change on the sensitive accuracy. A similar trend is not visible on the target accuracy. Regarding the choice of γ_{OD} and γ_E , we can see that the sensitive leakage is highly affected by these hyper-parameters and the results vary when changed. However, a more robust performance is observed on the target classification task.

5 Conclusions and future work

In this work, we have proposed a novel model for learning invariant representations by decomposing the learned codes into sensitive and target representation. We imposed orthogonality and disentanglement constraints on the representations and forced the target representation to be uninformative of the sensitive information by maximizing sensitive entropy. The proposed approach is evaluated on five datasets and compared with the state of the art models. The results show that our proposed model performs better than state of the art models on three datasets and performed comparably on the other two. We observe better hiding of sensitive information while affecting the target accuracy minimally. This goes in line with our hypothesis that decomposing the two representations and enforcing orthogonality could solve the information leakage problem by redirecting the information into the sensitive representation. One current limitation of this work is that it requires a target task to learn the disentanglement which could be avoided by learning the reconstruction as an auxiliary task similar to other privacy-preserving applications [24]. A direction worth investigating is replacing the pre-definition of the orthogonal sub-spaces priority by learning orthogonality intrinsically with low-rank constraints on the learned representations [22]. Another direction for future work could be focusing on the disentanglement part of the framework. The current disentanglement of factors of generation in the learned representations could be improved by using other disentanglement frameworks [2, 23] that are capable of better disentanglement.

Acknowledgments

S.A. is supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF).

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairml-book.org (2019), <http://www.fairmlbook.org>
2. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. pp. 2610–2620 (2018)

3. Creager, E., Madras, D., Jacobsen, J.H., Weis, M.A., Swersky, K., Pitassi, T., Zemel, R.: Flexibly fair representation learning by disentanglement. arXiv preprint arXiv:1906.02589 (2019)
4. Dua, D., Graff, C.: UCI machine learning repository (2017)
5. Edwards, H., Storkey, A.: Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 (2015)
6. Georgiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(6), 643–660 (2001)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in neural information processing systems*. pp. 2672–2680 (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: *European Conference on Computer Vision*. pp. 630–645. Springer (2016)
9. Kamiran, F., Calders, T.: Classifying without discriminating. In: *2009 2nd International Conference on Computer, Control and Communication*. pp. 1–6. IEEE (2009)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
13. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: *Advances in Neural Information Processing Systems*. pp. 14584–14597 (2019)
14. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair auto-encoder. arXiv preprint arXiv:1511.00830 (2015)
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(Nov), 2579–2605 (2008)
16. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309 (2018)
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
18. Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., Ver Steeg, G.: Invariant representations without adversarial training. In: *Advances in Neural Information Processing Systems*. pp. 9084–9093 (2018)
19. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 560–568 (2008)
20. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8227–8236 (2019)
21. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: A maximum entropy approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2586–2594 (2019)
22. Sanyal, A., Kanade, V., Torr, P.H., Dokania, P.K.: Robustness via deep low-rank representations. arXiv preprint arXiv:1804.07090 (2018)
23. Sarhan, M.H., Eslami, A., Navab, N., Albarqouni, S.: Learning interpretable disentangled representations using adversarial vaes. In: *Domain Adaptation and Rep-*

- resentation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 37–44. Springer (2019)
24. Xiao, T., Tsai, Y.H., Sohn, K., Chandraker, M., Yang, M.H.: Adversarial learning of privacy-preserving and task-oriented representations. arXiv preprint arXiv:1911.10143 (2019)
 25. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems. pp. 585–596 (2017)
 26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)
 27. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)

Fairness by Learning Orthogonal Disentangled Representations (supplementary material)

Mhd Hasan Sarhan^{1,2}[0000–0003–0473–5461], Nassir Navab^{1,3}, Abouzar Eslami¹[0000–0001–8511–5541], and Shadi Albarqouni^{1,4}[0000–0003–2157–2211]

¹ Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

`hasan.sarhan@tum.de`

² Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴ Computer Vision Lab, ETH Zurich, Switzerland

1 Results in numbers

Table 1: Results on all datasets. Bold face numbers on target task are the highest. Bold numbers in sensitive task are the closest to the majority label classifier.

	CIFAR-10		CIFAR-100		YaleB		Adult		German	
	T	S	T	S	T	S	T	S	T	S
Majority	0.6000	0.1000	0.0500	0.0100	0.0260	0.5056	0.7500	0.6700	0.7100	0.6900
Baseline	0.9775	0.2344	0.7199	0.3069	0.7800	0.9600	0.8500	0.8400	0.8700	0.8000
LFR [7]	-	-	-	-	-	-	0.8230	0.6700	0.7230	0.8050
VAE [3]	-	-	-	-	-	-	0.8190	0.6600	0.7250	0.7950
VFAE [4]	-	-	-	-	0.8500	0.5700	0.8130	0.6700	0.7270	0.7970
Xie et al. [6] (trade-off #1)	0.9752	0.2083	0.7132	0.1543	0.8900	0.5700	0.8440	0.6770	0.7440	0.8020
Roy et al. [5] (trade-off #1)	0.9778	0.2344	0.7117	0.1688	0.8900	0.4000	0.8460	0.6550	0.8633	0.7270
Xie et al. [6] (trade-off #2)	0.9735	0.2064	0.704	0.1484	-	-	-	-	-	-
Roy et al. [5] (trade-off #2)	0.9679	0.2114	0.705	0.1643	-	-	-	-	-	-
Ours	0.9725	0.1907	0.7074	0.1447	0.8923	0.5292	0.8520	0.6826	0.7700	0.7100

2 Biased categories accuracy

Biased categories accuracy reported in [6] which represents the target accuracy of the biased sensitive group. It is used as an indicator of how well the model performs on minority classes. The closer the biased accuracy to the overall accuracy, the better the model is in terms of fairness. The results are shown in Table. 2

3 Implementation details

Extended YaleB dataset: For the Extended YaleB dataset, we use an experimental setup similar to Xie *et al.* [6] and Louizos *et al.* [4] by using the same

Table 2: Biased categories accuracy on three datasets

	Entropy + KL Orth.		w/o Entropy w/o KL	
Dataset	Overall	Biased	Overall	Biased
Adult	0.8520	0.8157	0.8494	0.8108
German	0.77	0.7241	0.7433	0.6897
Yale	0.8923	0.5867	0.8650	0.4933

train/test split strategy. We used $38 \times 5 = 190$ samples for training and 1096 for testing. The model setup is similar to [6, 5], the encoder consisted of one layer, target predictor is one linear layer and the discriminator is neural network with two hidden layers each contains 100 units. The parameters are trained using Adam optimizer with a learning rate of 10^{-4} and weight decay of 5×10^{-2} . For the extended YaleB dataset we set $\lambda_{OD} = 0.037$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.1$, $\gamma_E = 2.0$, $d_T = d_S = 100$.

CIFAR datasets: Similar to [5], we employed ResNet-18 [1] architecture for training the encoder on the two CIFAR datasets. For the discriminator and target classifiers, we employed a neural network with two hidden layers (256 and 128 neurons). For the encoder, we set the learning rate to 10^{-4} and weight decay to 10^{-2} . For the target and discriminator networks, the learning rate and weight decay were set to 10^{-2} and 10^{-3} , respectively. Adam optimizer [2] is used in all experiments. For the CIFAR-10 dataset we set $\lambda_{OD} = 0.063$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.7$, $\gamma_E = 1$, $d_T = d_S = 128$. For the CIFAR-100 dataset we set $\lambda_{OD} = 0.0325$, $\lambda_E = 0.1$, $\gamma_{OD} = 1.2$, $\gamma_E = 1.67$, $d_T = d_S = 128$.

Tabular datasets: For the Adult and German datasets, we follow the setup appeared in [5] by having a 1-hidden-layer neural network as encoder, the discriminator has two hidden layer and the target predictor is a logistic regression layer. Each hidden layer contains 64 units. The size of the representation is 2. The learning rate for all components is 10^{-3} and weight decay is 5×10^{-4} . For the Adult dataset we set $\lambda_{OD} = 0.037$, $\lambda_E = 0.55$, $\gamma_{OD} = 0.8$, $\gamma_E = 1.66$, $d_T = d_S = 2$. For the German dataset we set $\lambda_{OD} = 0.01$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.4$, $\gamma_E = 2.0$, $d_T = d_S = 2$.

4 Sensitivity analysis heatmaps

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645. Springer (2016)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

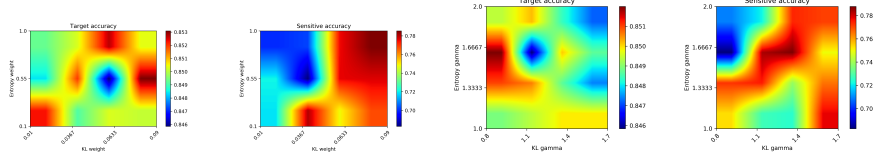


Fig. 1: Sensitivity analysis on the Adult dataset

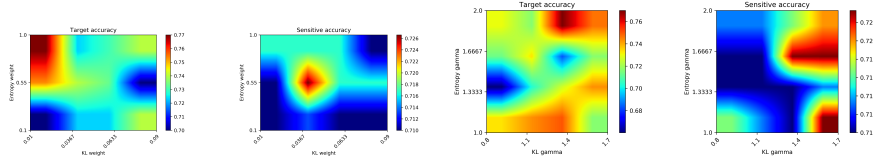


Fig. 2: Sensitivity analysis on the German dataset

3. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
4. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. arXiv preprint arXiv:1511.00830 (2015)
5. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: A maximum entropy approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2586–2594 (2019)
6. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems. pp. 585–596 (2017)
7. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)

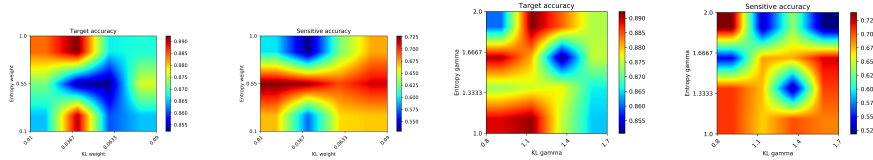


Fig. 3: Sensitivity analysis on the extended YaleB dataset

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 24, 2025

License Number 5995281073461

License date Mar 24, 2025

Licensed Content Publisher Springer Nature

Licensed Content Publication Springer eBook

Licensed Content Title Fairness by Learning Orthogonal
Disentangled Representations

Licensed Content Author Mhd Hasan Sarhan, Nassir Navab, Abouzar
Eslami et al

Licensed Content Date Jan 1, 2020

Type of Use Thesis/Dissertation

Requestor type academic/university or research institute

Format electronic

Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Learning Transferable Descriptive Disentangled Representations
Institution name	Technical University of Munich
Expected presentation date	Mar 2025
The Requesting Person / Organization to Appear on the License	Mohammad Hasan Sarhan
Requestor Location	Mohammad Hasan Sarhan Boltzmannstraße 3 Garching bei München, Bayern 85748 Germany
Billing Type	Invoice
Total	0.00 EUR

On the Fairness of Privacy-Preserving Representations in Medical Applications

Mhd Hasan Sarhan^{1,2}, Nassir Navab^{1,3}, Abouzar Eslami², Shadi Albarqouni^{1,4}

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴Computer Vision Lab, ETH Zurich, Zurich, Switzerland

Summary. This publication proposes the use of a novel fair representation learning framework in medical applications. The work focuses on discussing use cases where fairness is applied to medical datasets in order to build better understanding of scenarios where fairness is beneficial or not in such applications. The method relies on a multi-task approach to separate sensitive and target task representations while enforcing entropy maximization constraints on the sensitive prediction given the target representation. The method is evaluated on insurance claims dataset and a diagnostic dataset. Evaluation is done on target and sensitive information prediction performance using the target representation. The higher the target performance the better and the closer the sensitive performance to random uninformed classifier the better. When removing sensitive information from diagnostic data to predict a pathology, two cases are observed. Either sensitive information is important for the disease diagnosis or not. When the sensitive information is crucial for determining the disease, the removal of such information is harmful to the predictive model and may compromise the diagnostic performance of the model. Hence, when designing fairness for medical applications, it is important to identify if selected sensitive information stemming from societal beliefs for examples are important predictors of the studied disease.

Contribution. The contribution of the author of this thesis to this publication consists of the main idea, implementation, experiment design and evaluation. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details in implementation as well as the revision of the publication was done jointly with the co-authors.

Copyright Statement. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer, Cham, Lecture Notes in Computer (LNCS, volume 12444), Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020 Proceedings, On the Fairness of Privacy-Preserving Representations in Medical Applications, Mhd Hasan Sarhan, Nassir Navab,

Abouzar Eslami, Shadi Albarqouni, © 2020.

The following text is reprinted with the permission of the publisher (License number 5995550368019). The license is attached after the publication text.



On the Fairness of Privacy-Preserving Representations in Medical Applications

Mhd Hasan Sarhan^{1,2} , Nassir Navab^{1,3}, Abouzar Eslami² ,
and Shadi Albarqouni^{1,4}

¹ Computer Aided Medical Procedures, Technical University of Munich,
Munich, Germany

hasan.sarhan@tum.de

² Translational Research Lab, Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴ Computer Vision Lab, ETH Zurich, Zürich, Switzerland

Abstract. Representation learning is an important part of any machine learning model. Learning privacy-preserving discriminative representations that are invariant against nuisance factors is an open question. This is done by removing sensitive information from the learned representation. Such privacy-preserving representations are believed to be beneficial to some medical and federated learning applications. In this paper, a framework for learning invariant fair representations by decomposing the learned representation into target and sensitive codes is proposed. An entropy maximization constraint is imposed on the target code to be invariant to sensitive information. The proposed model is evaluated on three applications derived from two medical datasets for autism detection and healthcare insurance. We compare with two methods and achieve state of the art performance in sensitive information leakage trade-off. A discussion regarding the difficulties of applying fair representation learning to medical data and when it is desirable is presented.

Keywords: Fair representations · Invariant representations · Privacy-preserving representations · ABIDE dataset

1 Introduction

With the success of machine learning algorithms in learning descriptive representations, it became necessary that such representations are robust against arbitrary noisy signals. Such representations are defined to hold meaningful information about a target task of interest while being invariant to attributes believed to be noisy or sensitive. Deep neural networks have shown tremendous performance in learning descriptive representations. Since these networks learn from human collected data, they might suffer from learning inherent biases and leak information that is not intended to be released. Such systems are becoming

corner stones of various decision support systems including medical ones. This has raised various concerns about fairness, privacy and discrimination in medical applications and statistical models [2, 10]. Privacy-preserving representations and federated learning present a possible solution [14]. In most literature [15, 16, 18], a fair representation is treated as a representation $\mathbf{z}$ that is well informed about a target attribute $\mathbf{y}$ (good bank credit) while being robust against a nuisance sensitive attribute $\mathbf{s}$ (age or gender). The assumption when building such models is that sensitive $\mathbf{s}$ and target $\mathbf{y}$ variables are independent. This assumption does not always hold and it is highly application specific. When $\mathbf{s}$ and $\mathbf{y}$ are not independent, which is the case in most fairness applications, there is a trade-off between target accuracy and sensitive information leakage (such as salary prediction with gender as a sensitive attribute) [15, 18]. In medical applications, it is crucial to distinguish these two cases as such fairness might hinder the performance of the downstream task. If a diagnostic decision is dependent on the subject’s race, it might be unfair to build a diagnostic model that ignores or hides such information. We show in our experiments an example of both cases.

To achieve fairness, earlier approaches relied on data-massaging techniques [5, 13]. Such approaches make the assumption that equal opportunities in training data would generalize to test data. The data points are rearranged to enforce a fair representation of all groups. Such approaches are difficult to ensure especially in medical applications where data is scarce and must be utilized. Later, a fair clustering approach is proposed to learn a representation where examples are clustered to have equal number of examples from all sensitive attributes in each cluster [19]. This approach hinders the representations distribution property with clustering constraints which leads to a larger target/sensitive trade-off. Later approaches rely on approximating mutual information between representation $\mathbf{z}$ and sensitive attribute $\mathbf{s}$ and minimizing it in an either adversarial [9, 15, 18] or non-adversarial [7, 11, 16] manner. To approach invariance non-adversarially, Louizos *et al.* [7] have proposed the variational fair autoencoder (VFAE) where a representation is learned to be representative of the data while robust to a nuisance. This representation is then used for a target task. The robustness is approached by Maximum Mean Discrepancy (MMD) penalty. In [11], the authors use a variational upper bound to approximate mutual information between $\mathbf{z}$ and $\mathbf{s}$. Other approaches, tackle the problem in adversarial manner by posing a zero-sum minimax game between an encoder that learns the desired representation and an adversarial discriminator that learns to distinguish the sensitive information in the representation [18]. Such zero-sum optimization problems might need additional regularization terms to stabilize the training [15]. A non-zero sum game approach is proposed to mitigate this problem by optimizing an entropy loss that enforces the discriminator to be agnostic to sensitive information [15]. While adversarial approaches have been successful in learning fair representations, they suffer from convergence related problems and sometimes they could hinder the performance [11].

In this work, we propose a non-adversarial framework for learning fair representations by imposing an entropy maximization constraint on a decomposed

latent representation that encodes target information in one split and residual information related sensitive attributes in the other. The representations are learned in a multi-task-like framework with entropy loss to enforce the target representation to be agnostic to sensitive information. We evaluate this framework on two medical datasets and three applications related to fairness. We discuss how such approaches could be applied to medical data and how fairness approaches could be beneficial or harmful in the healthcare.

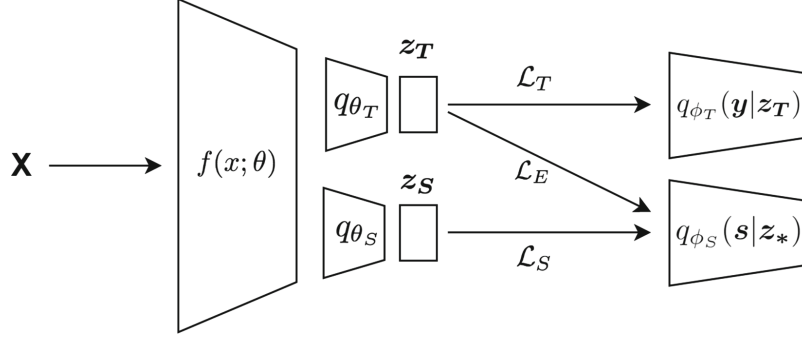


Fig. 1. The proposed fairness framework. The input is split into target and residual sensitive representations and then fed into two discriminators to predict the target and sensitive attributes.

2 Method

let $\mathcal{X}$ be a dataset of subjects and $\mathbf{x} \in \mathbb{R}^D$ be an input sample. An input sample $\mathbf{x}$ is associated with a target attribute (class) $\mathbf{y} = \{y_1, \dots, y_n\} \in \mathbb{R}^n$, and a sensitive class $\mathbf{s} = \{s_1, \dots, s_m\}$. We aim to learn a representation $\mathbf{z}_T$ that contains information about the target attributes while limiting the information leakage about sensitive attributes. An encoder is trained to map the input $\mathbf{x}$ to two representations $\mathbf{z}_T \in \mathbb{R}^d$, $\mathbf{z}_S \in \mathbb{R}^d$ where $\mathbf{z}_T$ contains target information while $\mathbf{z}_S$ carry residual information partly related to the sensitive attributes.

An input sample $\mathbf{x}$ is given to a shared encoder $f(\mathbf{x}; \theta)$, then projected into two representations $\mathbf{z}_T$ and $\mathbf{z}_S$ (see Fig. 1). These subspaces are acquired by the two encoders $q_{\theta_T}(\mathbf{z}_T|\mathbf{x})$, and $q_{\theta_S}(\mathbf{z}_S|\mathbf{x})$ parameterized by θ_T and θ_S respectively. The first part of the proposed loss is analogous to multi-task problem loss. Each of the two representations is fed to a network to learn a task. The target code $\mathbf{z}_T$ is fed to a target discriminator $q_{\phi_T}(\mathbf{y}|\mathbf{z}_T)$ to learn the target downstream task. The residual sensitive code $\mathbf{z}_S$ is fed to a sensitive discriminator $q_{\phi_S}(\mathbf{s}|\mathbf{z}_S)$ to learn to distinguish sensitive attributes. All these networks are trained in a supervised fashion to minimize the following objective functions,

$$\mathcal{L}_T(\theta, \theta_T, \phi_T) = - \sum p(\mathbf{y}|\mathbf{x}) \log[q_{\phi_T}(\mathbf{y}|\mathbf{z}_T)], \quad (1)$$

$$\mathcal{L}_S(\theta_S, \phi_S) = - \sum p(\mathbf{s}|\mathbf{x}) \log[q_{\phi_S}(\mathbf{s}|\mathbf{z}_S)], \quad (2)$$

where $p(\mathbf{y}|\mathbf{x})$ is the target attribute $\mathbf{y}$ of input $\mathbf{x}$, $p(\mathbf{s}|\mathbf{x})$ is the sensitive attribute $\mathbf{s}$ of input $\mathbf{x}$, and θ , θ_T , θ_S , ϕ_T , ϕ_S are the network parameters (weights). Only the parameters of the sensitive discriminator (ϕ_S) and the sensitive encoder (θ_S) are optimized when minimizing $\mathcal{L}_S$. The parameters of the backbone shared encoder are not considered. Since we assume independence of the two attributes (target and sensitive), the losses $\mathcal{L}_S$ and $\mathcal{L}_T$ in this case have competing objectives.

This on its own, would not prevent leakage of sensitive information to $\mathbf{z}_T$. Inspired by adversarial entropy maximization proposed by Roy *et al.* [15], we maximize the entropy of the sensitive discriminator given the target representation as input. In the proposed framework, it is not required to maximize entropy in an adversarial manner since the target and sensitive codes are split and the entropy maximization could be applied only to the target representation as

$$\mathcal{L}_E(\theta, \theta_T, \phi_S) = \sum q_{\phi_S}(\mathbf{s}|\mathbf{z}_T) \log[q_{\phi_S}(\mathbf{s}|\mathbf{z}_T)], \quad (3)$$

The entropy loss $\mathcal{L}_E$ enforces the target representation to be uninformative of the sensitive task. This is a desired feature of a fair representation since we favour representations that mitigate sensitive information leakage.

The overall objective function of the proposed framework can be written as

$$\arg \min_{\theta, \theta_T, \theta_S, \phi_T, \phi_S} \mathcal{L}_T(\theta, \theta_T, \phi_T) + \mathcal{L}_S(\theta_S, \phi_S) + \lambda_E \mathcal{L}_E(\theta, \theta_T, \phi_S) \quad (4)$$

where λ_E is a hyper-parameter to weigh the *Entropy* loss.

3 Experiments

In this section, we evaluate the effectiveness of the learned fair representations in medical applications by evaluating on two publicly available datasets for three tasks. The results are compared with a baseline model and two state-of-the-art models in learning fair representations in an adversarial fashion. We also discuss the quantitative results of the learned representations by showing the t-SNE projections of the learned representations compared to the baseline.

3.1 Experimental Setup

Datasets: To evaluate fair classification in medical applications we chose two publicly available datasets to investigate. First, we use the tabular *Heritage Health* dataset¹. The dataset contains physicians records and insurance claims of over 60k patients. The target downstream task is to predict a comorbidity indicator that estimates the risk of patient death in the next several years (the Charlson Index). The sensitive attribute in this case is the sex of the patient, partially similar to [17]. We study in this case if sex should be overlooked while analyzing insurance claims of re-visitations of patients to the hospitals.

¹ Heritage Health dataset, www.heritagehealthprize.com.

The second dataset we use is the *ABIDE* dataset [4]. The dataset contains fMRI scans and demographic data of 1112 subjects. Similar to [1, 12], we utilize the same set of 871 subjects that met the quality criteria. The dataset contains 403 subjects with Autism Spectrum Disorder (ASD) and 468 healthy ones. We follow the preprocessing steps as in [1], namely the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [3]. The target downstream task is the detection of ASD subjects. We experiment with two different potential sensitive attributes namely sex and age. We binarize age into young and old subjects. We consider patients over 20 years old to be in one class.

Implementation Details: For the health heritage dataset, we use an encoder with three hidden layers (64, 64, 40). The target classifier and discriminator are networks with three hidden layers (10, 10, 10). The size of the representation is 20. The learning rate for all components is 10^{-5} and weight decay is 5×10^{-4} . The dataset is split into train\validation splits (80%\20%) and all models are trained on the same split for fair comparison. $\lambda_E = 1.5$.

For the ABIDE dataset, we use an encoder with two hidden layers (3000, 1500, 200) with dropout probability of 0.3. The target and discriminator are 2-hidden-layered networks (40, 30). The learning rate for all components is 10^{-5} and eight decay is 5×10^{-4} . The baseline models for both datasets are the encoder and the target models combined to solve one task which is target attribute identification. Hence, no residual sensitive representation $\mathbf{z}_S$ is learned in this scenario. Following [12], we use a 10-fold stratified cross-validation for evaluation. All models are trained on the same splits for fair comparison. $\lambda_E = 2$. We use Adam optimizer [6] for all models. Moreover, we use the same encoder and discriminators architectures for training the state-of-the-art methods with the same hyper-parameters. We set $\alpha = 0.3$ for the compared methods [15, 18].

Experiments Design: In the experiments we aim to evaluate the learned representations in terms of 1) how much sensitive information is retained in the target representation $\mathbf{z}_T$. To this end, we use the learned representation $\mathbf{z}_T$ as an input to as simple logistic regression model trained to identify sensitive attributes from target representations. Ideally, this model is not able to distinguish sensitive attributes and has an accuracy similar to naively guessing the majority class as output. It is also important to evaluate 2) how well the learned representation $\mathbf{z}_T$ describes the target attributes. We train a logistic regression model to identify target attributes from the target representation $\mathbf{z}_T$. Ideally, this model should perform on par or better than a baseline model.

3.2 ABIDE Dataset

For the ABIDE dataset we chose two sensitive attributes for the experiments. First we use age as a sensitive attribute. It is believed that age is a redundant phenotypic variable that does not contribute positively to the prediction task of ASD [12]. This goes inline with the definition of fair representations where the sensitive information is believed to be independent from the target task and could

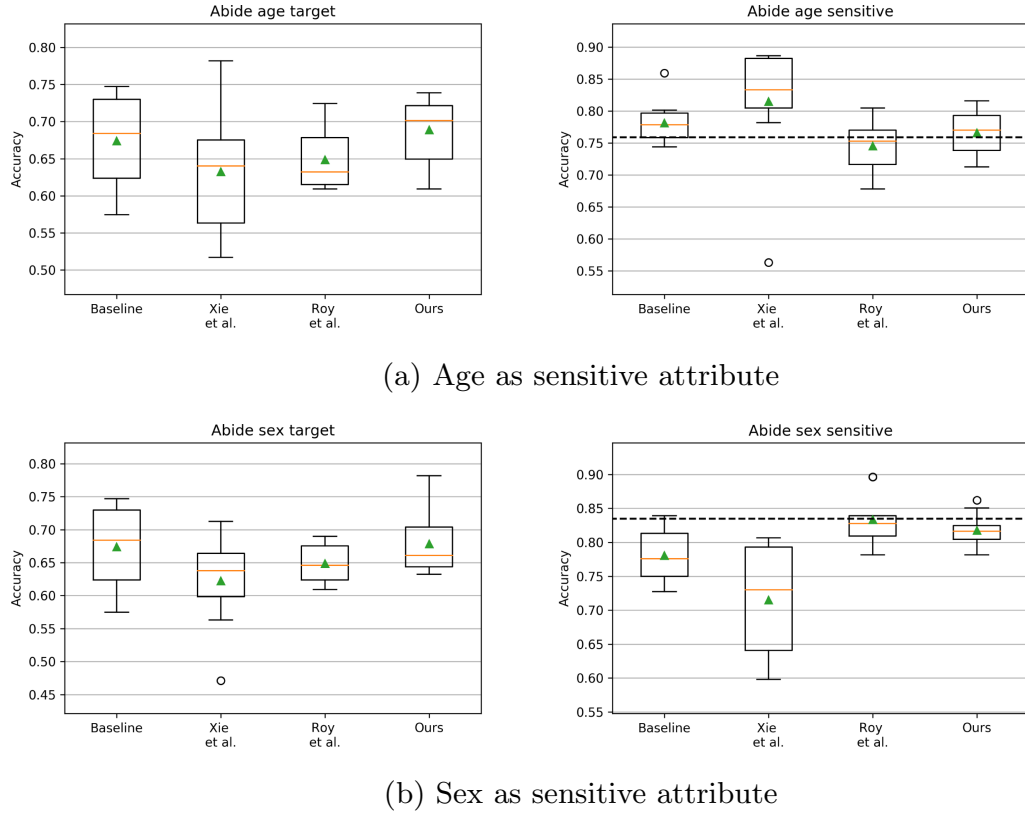


Fig. 2. Results on the ABIDE dataset. The majority classifier is represented by a dashed black line.

act as noise signal that disturbs the target classification. We try with this experiment to remove any information about patients' age from the pre-processed imaging input by learning a representation that is free from this information. The phenotypic age information itself is not provided to the network as part of the input. The results for age are shown in Fig. 2a. Regarding the sensitive information leakage mitigation, the proposed approach well hides the sensitive information in the representation. This can be seen by the mean accuracy on the sensitive class (76.57%) which is the closest to the majority classifier (75.88%). The mean and median target accuracy (69%, 70.11%) are well improved over the baseline (67.38%, 68.4%) and close to the recent state-of-the-art mean accuracy on this dataset [12] (70%) using population graph convolutional neural networks which includes phenotypic information into their model. As suggested by [12] and validated in our experiment, not all phenotypic information are beneficial to the models and some might have negative effect. By removing the age information from the data representation, we obtain better classification for the ASD detection target task. We do not see a similar trend considering sex as the sensitive attribute Fig. 2b. When hiding sex related information from the representation, we monitor a good performance in mitigating sensitive information leakage. However, the target accuracy does not have a similar trend and it is even lowered in median value (66.09%) compared to the baseline (68.4%).

This could indicate that sex and ASD are not independent variables which suggests that there should be a trade-off between sensitive information leakage and target performance.

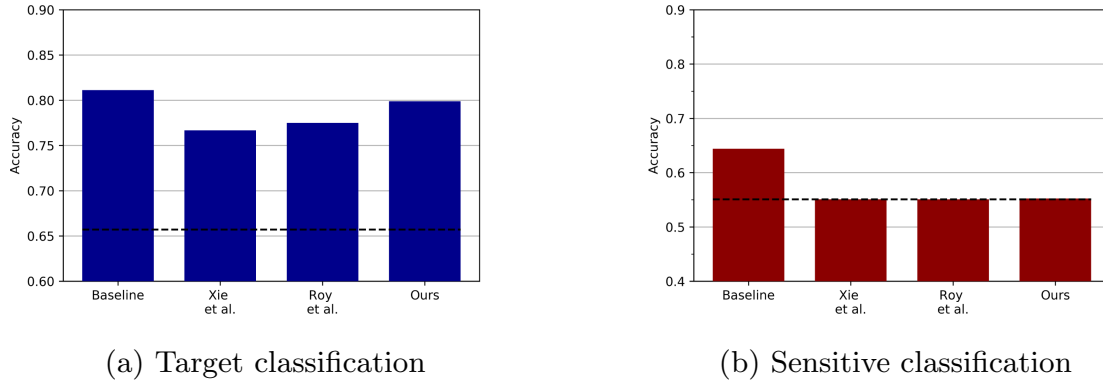


Fig. 3. Results on the health dataset. The majority classifier is represented by a dashed black line.

3.3 Health Dataset

The results on the health dataset are shown in Fig. 3. The proposed model better preserves the sensitive information compared to the baseline classifier, where the baseline has an accuracy of 64.38% compared to 55.22% which is closer to the majority classifier line at 55.09%. The results of the proposed model are similar to other fair models in terms of hiding sensitive information. However, the proposed model has a better trade-off for the target accuracy. This is shown by the higher target accuracy of the model compared to other fairness models while keeping on par sensitive performance. This suggests that it is possible to make an informed decision if a patient will return to the hospital or not regardless of the subject's sex.

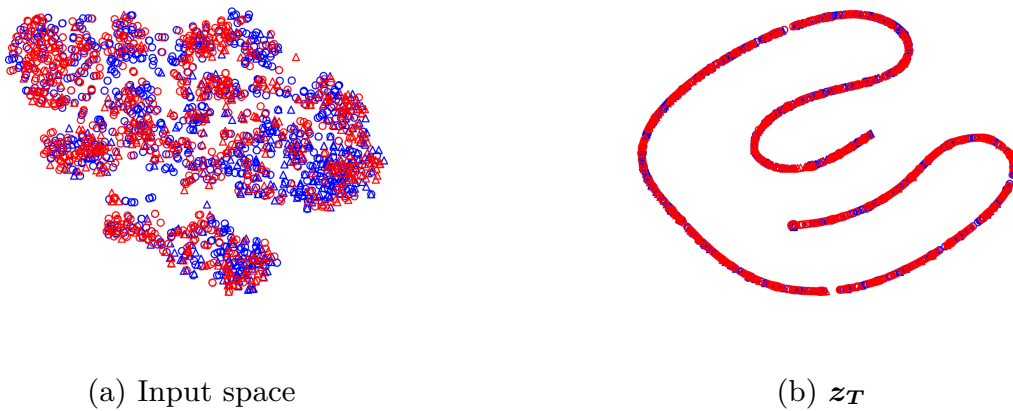


Fig. 4. t-SNE projections of the input space and z_T on the health dataset

Qualitative Analysis: The learned embeddings are visualized with t-SNE [8] projections on the health dataset in Fig. 4 (Better seen in high resolution). We visualize the input space and $\mathbf{z}_T$ projections to intuitively understand what information is held in each representation. The color represents the sensitive information (sex) and the shape represents the target information (Charlson Index). We can see in the input space there is no separation between target information as the shapes are mixed. The sensitive information is indistinguishable to some extent as we observe some small clusters with more dominance of one color. When projecting the learned embeddings, it is clear that all sensitive information is mixed. The target information is split as could be seen at the two ends of the projections where each end is dominated by one shape.

4 Discussion and Conclusion

In this paper, a model for learning privacy-preserving invariant representations is proposed. The model learns invariance by decomposing representations into target and residual representations and enforcing entropy maximization constraints to mitigate sensitive information leakage. The model is evaluated on two healthcare-related datasets to answer the question of the effectiveness of using fair representations in medical applications. One dataset is the tabular insurance claims health dataset, and the other is the ABIDE dataset for detecting autism spectrum disorder from fMRI images. We distinguish different cases for using fairness in medical applications, these cases are mainly relying on our belief of fairness in these situations and how much the selected sensitive attributes contribute to diagnosis decision. In the health dataset, the target task is not a diagnostic decision but rather, an expectation of the subject’s hospital readmission. This information is important for insurance companies for the subjects’ future plans. In such a scenario, some historic biases might be included in the retrospectively collected data. Such biases could be learned by a machine learning model and the model would have similar biases cause by target and sensitive attributes that are not independent. The proposed fairness model mitigates the leakage of sensitive information at the cost of a little drop in target performance. This suggests that the two variables cannot be forced to be independent (at least by the proposed and compared models). On the other hand, the ABIDE dataset is a diagnostic dataset which changes the way fairness is approached. It is important in such cases for the selected sensitive information to be irrelevant for the diagnosis (i.e. target and sensitive attributes are independent). The age at the scan time is selected as a sensitive attribute since it is reported in the literature that including this phenotypic information in the classification drops the diagnostic performance of the model [12]. This suggests that such an attribute is irrelevant to the diagnosis and might act as a nuisance. We validate this by showing that target performance improves when removing this attribute that is believed to be clinically irrelevant to the case at hand. However, this is not the case when the attribute is clinically relevant to the diagnosis. When trying to hide information about the subject’s sex, the target performance drops which could be

because the two variables are not independent. These experiments show that using privacy-preserving models for medical applications should be done with clinical relevance in mind as hiding clinically important information for the sake of security might compromise the downstream task performance.

Acknowledgments. S.A. is supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF).

References

1. Abraham, A., et al.: Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage* **147**, 736–745 (2017)
2. Chen, I.Y., Szolovits, P., Ghassemi, M.: Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* **21**(2), 167–179 (2019)
3. Craddock, C., et al.: Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* **42** (2013)
4. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**(6), 659–667 (2014)
5. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication, pp. 1–6. IEEE (2009)
6. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
7. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. arXiv preprint [arXiv:1511.00830](https://arxiv.org/abs/1511.00830) (2015)
8. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)
9. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. arXiv preprint [arXiv:1802.06309](https://arxiv.org/abs/1802.06309) (2018)
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint [arXiv:1908.09635](https://arxiv.org/abs/1908.09635) (2019)
11. Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., Ver Steeg, G.: Invariant representations without adversarial training. In: *Advances in Neural Information Processing Systems*, pp. 9084–9093 (2018)
12. Parisot, S., et al.: Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease. *Med. Image Anal.* **48**, 117–130 (2018)
13. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568 (2008)
14. Rieke, N., et al.: The future of digital health with federated learning. arXiv preprint [arXiv:2003.08119](https://arxiv.org/abs/2003.08119) (2020)
15. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: a maximum entropy approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2586–2594 (2019)
16. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. arXiv preprint [arXiv:2003.05707](https://arxiv.org/abs/2003.05707) (2020)

17. Song, J., Kalluri, P., Grover, A., Zhao, S., Ermon, S.: Learning controllable fair representations. arXiv preprint [arXiv:1812.04218](https://arxiv.org/abs/1812.04218) (2018)
18. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems, pp. 585–596 (2017)
19. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333 (2013)

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 24, 2025

This Agreement between Mohammad Hasan Sarhan ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5995550368019
License date	Mar 24, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	On the Fairness of Privacy-Preserving Representations in Medical Applications
Licensed Content Author	Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami et al
Licensed Content Date	Jan 1, 2020
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute

Format	electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Learning Transferable Descriptive Disentangled Representations
Institution name	Technical University of Munich
Expected presentation date	Mar 2025
The Requesting Person / Organization to Appear on the License	Mohammad Hasan Sarhan
Requestor Location	Mohammad Hasan Sarhan Boltzmannstraße 3 Garching bei München, Bayern 85748 Germany
Billing Type	Invoice
Total	0.00 EUR

Learning Interpretable Disentangled Representations using Adversarial VAEs

Mhd Hasan Sarhan^{1,2}, Abouzar Eslami², Nassir Navab^{1,3}, Shadi Albarqouni¹

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Summary. This publication introduces an unsupervised disentanglement framework using adversarial variational autoencoders that achieves good disentanglement, reconstruction, and downstream performance. The method utilizes a variational autoencoder structure to enforce a factorized Gaussian distribution on the learned representation. This factorization acts as a proxy for disentanglement as it forces different generative factors of the data to be independently represented in different components of the representation. Total correlation constraints are applied to the representation to minimize mutual information between components. The adversarial training is added to the reconstruction objective where an adversary learns to distinguish generated fake images from images sampled from the dataset to promote more realistic generation of images. The challenge of such applications in medical scenarios is the lack of data labeled with generative factors for the evaluation of disentanglement on such datasets. This is partially resolved in the proposed work by utilizing a segmentation dataset of skin lesions to extract some generative factors of the lesions. These extracted generative factors are then used in the evaluation process. The framework performed better the state-of-the-art methods in disentangling factors of variation and retaining fidelity of reconstruction. Moreover, the representation learned by the method is used for a downstream task of skin lesion classification and performed better than representation of other compared methods. In conclusion, learning disentanglement in an adversarial total correlation variational autoencoders improves the disentanglement and reconstruction fidelity of the model and provides representations that work well on downstream tasks of medical applications.

Contribution. The contribution of the author of this thesis to this publication consists of the main idea, implementation, experiment design and evaluation. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details in implementation as well as the revision of the publication was done jointly with the co-authors.

Copyright Statement. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer, Cham, Lecture Notes in Computer (LNCS, volume 11795), First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in

Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019 Proceedings, Learning Interpretable Disentangled Representations Using Adversarial VAEs, Mhd Hasan Sarhan, Abouzar Eslami, Nassir Navab, Shadi Albarqouni, © 2019.

The following text is reprinted with the permission of the publisher (License number 5995550403277). It is the accepted but not the published version of the paper. The license is attached after the publication text.

Learning Interpretable Disentangled Representations using Adversarial VAEs

Mhd Hasan Sarhan^{1,2}[0000–0003–0473–5461], Abouzar Eslami²[0000–0001–8511–5541], Nassir Navab^{1,3}, and Shadi Albarqouni¹[0000–0003–2157–2211]

¹ Computer Aided Medical Procedures, Technical University of Munich, Germany

² Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. Learning Interpretable representation in medical applications is becoming essential for adopting data-driven models into clinical practice. It has been recently shown that learning a disentangled feature representation is important for a more compact and explainable representation of the data. In this paper, we introduce a novel adversarial variational autoencoder with a total correlation constraint to enforce independence on the latent representation while preserving the reconstruction fidelity. Our proposed method is validated on a publicly available dataset showing that the learned disentangled representation is not only interpretable, but also superior to the state-of-the-art methods. We report a relative improvement of 81.50% in terms of disentanglement, 11.60% in clustering, and 2% in supervised classification with a few amount of labeled data.

Keywords: Deep Learning · Disentangled Representation · Interpretability

1 Introduction

Data-driven models with the help of Deep Learning (DL) are affecting wide areas of scientific research and the medical domain is no exception in this matter. However, in healthcare, developing a machine learning algorithm with expert level performance is important but not enough for the adoption of the algorithm when the issues of trust and explainability are not taken into consideration [14]. Explainability of a model is approached either by 1) explicitly learning it by model design or 2) after model design such as using gradient-based localization [15].

Approaching explainability by model design could be facilitated in a supervised manner as in decision trees and rule-based systems or in an unsupervised manner as in Variational Autoencoder (VAE) [9] or β -Variational Autoencoder (β -VAE) [6]. In the latter, a lower dimensional representation of the data is learned and utilized for analyzing the data. The rest of the paper discusses this type of explainability. Deep learning models extract features from data in order

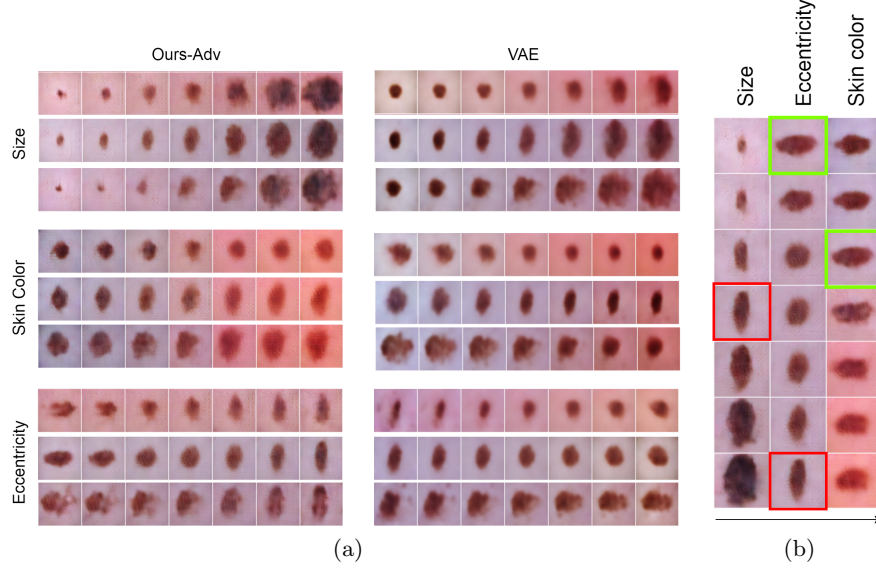


Fig. 1: Comparison of our model to VAE on examples for traversal over the representation components. Traversal is done between $[-3, 3]$ (a) Examples of traversal for three images form ISIC 2018. Each row shows reconstructions of latent traversals across one latent dimension; (b) Example of a smooth transition over the manifold by changing multiple latent dimensions to go from small lesion on pale skin (top left image) to bigger horizontal lesion on red skin (bottom right image). Each column represents one dimension of change. The colored squares represent the image of the previous column from which the traversal has started on the current dimension.

to represent it in a compressed high-level representation that suits the application. The quality of this representation is crucial for the model performance and it is argued that disentangled representations would be helpful for having better control and interpretability over the data [1, 6]. A disentangled representation can be defined as a representation where one latent unit represents one generative factor of variation in the data while being invariant to other generative factors [1]. For example, a model trained on a dataset of faces would learn disentangled latent units that represent independent ground truth generative factors such as hair color, pose, lighting or skin color. Disentangling as many explanatory factors as possible is important for a more compact, explainable, transferable, abstract representation of the data [1].

Most of the previous work regarding disentanglement relied on information about the number or nature of the ground truth generative factors [7, 10]. In medical applications, the data is complex and a priori knowledge about the gen-

erative factors is mostly unavailable. Recently, multiple models for unsupervised disentangled feature learning were proposed [3, 6, 8, 2]. β -VAE [6] is proposed as a modification on VAE [9] where the parameter β is used to introduce more emphasis on the KL-Divergence part of the VAE objective. This enforces the posterior to match the factorized Gaussian prior which constraints the bottleneck representation to be factorized while still reconstructing the data. Higher β values encourage more disentangled representations with a trade-off on the reconstruction error. In β -Total Correlation VAE (β -TCVAE) [2], the training is focused on the total correlation part of KL term which is responsible for the factorized representation. This lowers the trade-off on the reconstruction fidelity proposed by β -VAE. β -TCVAE is validated on examples from a controlled environment with clear factors of generation. This does not represent the complexity of medical data and should be addressed.

Contributions: In this work, we propose a framework for learning disentangled representations in medical imaging in an unsupervised manner. To our knowledge, this is the first work that analyzes the strength of unsupervised disentangled feature representations in medical imaging and proposes a framework that is well suited to medical applications. We propose a novel residual adversarial VAE with total correlation constraint. This enhances the fidelity of the reconstruction and captures more details that describe better the underlying generative factors.

2 Methodology

We utilize deep generative disentangled representation learning to learn the distribution of a medical imaging dataset. We then use the learned representation to generate images while controlling some generative factors. We first show how disentanglement is approached with β -VAE as a motivation for incorporating β -TCVAE. We then present our contributions to the disentanglement framework by utilizing adversarial loss with residual blocks to enhance the disentanglement and reduce the compromise on the reconstruction. We hypothesize that using adversarial loss with residual blocks in a disentanglement framework would result in higher quality representations with more disentanglement in the feature space.

Background: Let $x_n \in \mathcal{X}, n = 1, \dots, N$ be a set of images generated by combinations of K ground-truth generative factors $V = (v_1, \dots, v_K)$. Our aim is to build an unsupervised generative model that utilizes only the images in $\mathcal{X}$ to learn the joint distribution of the images and the set of latent generative factors $z \sim q_\phi(z|x) \in \mathbb{R}^d$ allowing us to have better control and interpretability of the latent space. It is worth mentioning the latent generative factors capture both disentangled and entangled factors. To realize our aim, we follow the concept of β -VAE in learning a posterior distribution that could be used to generate images from $\mathcal{X}$. The posterior representation is approximated by $q_\phi(z|x)$. The model is built such that the generative factors V are represented by the posterior bottleneck in a disentangled fashion.

In β -VAE, implicit independence is enforced on the posterior to encourage a disentangled representation. This is done by constraining the posterior to match a prior $q(z)$. The prior is set to be an isotropic unit Gaussian ($p(z) = \mathcal{N}(0, I)$). Adding extra pressure on the posterior to match $p(z)$ constraints the capacity of the bottleneck and pushes it to be factorized [6]. Thus, the objective function for β -VAE is as follows

$$\arg \min_{\phi, \theta} \left[\underbrace{-\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)]}_{\text{reconstruction loss } \mathcal{L}_{rec}} + \beta D_{KL}(q_\phi(z|x) || p(z)) \right] \quad (1)$$

where θ and ϕ are trainable weights of encoder and decoder respectively, D_{KL} is the Kullback-Leibler divergence. For enforcing disentanglement, values of $\beta > 1$ are typically chosen. Using this formula enhances the disentanglement at the cost of reconstruction fidelity. It is suggested by [2] that the total correlation term within D_{KL} is responsible for the factorized disentangled representation. Hence, focusing the training on the total correlation would result in better disentanglement while having less effect on the reconstruction. The objective function changes such as D_{KL} is decomposed and β is multiplied by the total correlation term as follows

$$\arg \min_{\phi, \theta} \left[-\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + \underbrace{I_q(z, x) + \beta D_{KL}(q_\phi(z) || \prod_j q_\phi(z_j)) + \sum_j D_{KL}(q_\phi(z_j) || p(z_j))}_{D_{KL}(q_\phi(z|x) || p(z)) \text{ decomposition } (\mathcal{L}_{prior})} \right] \quad (2)$$

The term $D_{KL}(q(z) || \prod_j q(z_j))$ is the total correlation (TC) which is a generalization of mutual information to more than two variables. Penalizing TC forces independence in the represented factors. We use β -TCVAE for its good results on disentanglement on various datasets while having better reconstruction than other disentanglement models and for the parameter-less approximation of $q(z)$. For more details about the D_{KL} decomposition and the approximation of $q(z)$ the reader is referred to [2].

Proposed approach: To enhance the fidelity of the reconstructions and improve the generative factors captured by z , we propose using a discriminator network with β -TCVAE disentanglement model. The discriminator is trained to decide whether an input image is generated synthetically or sampled from the real data distribution. We employ adversarial loss scheme for the training. The discriminator in this scenario has to learn implicitly a rich similarity metric based on features extracted from the images rather than relying only on pixel-wise similarity. This does not only improve generated images visually, but also learns a richer representation in the code z [11]. This is because the pixel-wise loss acts as a *content* loss while the discriminator loss acts as a *style* loss [4]. Moreover, we incorporate residual blocks rather than convolutional layers applied in [2]. This is because residual blocks have shown a better flow of the gradients.

This limits the problems related to vanishing/exploding gradients [5] and is being used in state-of-the-art Generative Adversarial Nets (GANs) literature [17] for more stable training. We denote $Dis(\cdot; \psi)$ to the discriminator network described by trainable parameters ψ , x is a real image sampled from $p(x)$ and $\hat{x}$ is the reconstructed image from $p_\theta(x|z)$. The final objective is

$$\begin{aligned} \arg \min_{\phi, \theta} [\mathcal{L}_{gen}] &= \arg \min_{\phi, \theta} [\mathcal{L}_{rec} + \mathcal{L}_{prior} - \log(Dis(\hat{x}))] \\ \arg \min_{\psi} [\mathcal{L}_{disc}] &= \arg \min_{\psi} [-\log(Dis(x)) - \log(1 - Dis(\hat{x}))] \end{aligned} \quad (3)$$

The model is trained by alternating between $\mathcal{L}_{gen}$ and $\mathcal{L}_{disc}$ optimization. We use pixel-wise l_2 -distance between x and $\hat{x}$ as $\mathcal{L}_{rec}$.

3 Experiments

Experimental validation evaluates the proposed framework in two main experiments: First, we compare our proposed method disentanglement performance to state-of-the-art methods in learning both entangled and disentangled representations. We also utilize the learned representations in two use-cases, namely, unsupervised clustering and supervised classification with a few amounts of labels. In the second experiment, we evaluate the results visually and analyze the interpretable learned representation.

Dataset: We opt for the publicly available Skin Lesion dataset from ISIC 2018 Challenge [16] to perform our validations. To train our model, we utilize the dataset of Task 3 which consists of 10k RGB images with 7 types of skin lesions capturing 7 pathological generative factors. To evaluate the model against ground-truth generative factors, i.e. *eccentricity*, *orientation*, and *size*, we utilize the dataset of Task 2 which consists of 2k images with pixel-wise segmentation. Note that all images are down-sampled to $64 \times 64px$.

Evaluation metrics: To quantitatively evaluate the disentanglement quality, we report the Mutual Information Gap (MIG) metric as proposed and suggested in [2]. As opposed to the disentanglement metric in [6], MIG takes axis-alignment (one v_k is captured by one z_j) into consideration, and it is unbiased to hyper-parameters opposite to [6, 8]. MIG measures the mutual information (MI) between z_j and the known generative factor v_k , then the difference between the two highest MIs of a generative factor is calculated, and normalized then by the entropy of v_k . The average MIG is then reported as the final MIG score. The generative factors are set as follows:

1. **MIG Pathologies (MIG_p):** The ground truth classes are used as generative factors in one vs. all fashion. For instance, $K = 7$ for the Skin Lesion dataset. Each generative factor has two possible values in this scenario.

2. **MIG Handcrafted Factors** (MIG_{hf}): In addition, we handcrafted a few generative factors which are easily visible in the image space, e.g. geometric and morphological changes. To do so, the segmentation masks given in Task 2 are utilized. The handcrafted factors are *eccentricity*, *orientation*, and *size* (i.e $K = 3$). Each generative factor has two possible values.

In addition, we report the Peak signal-to-noise ratio (PSNR), Normalized Mutual Information (NMI), and Accuracy (ACC) to evaluate the reconstruction error, clustering, and classification, respectively.

Baselines: We compare the proposed model to two representation learning models. The first is VAE [9] model which does not take disentanglement into account explicitly. The second model is β -TCVAE [2] which adds constraints on the representation to disentangle the components. Further, We employ two variations of our proposed method with bottleneck residual blocks [5]; 1) without the adversarial loss in Equation 3 denoted as *Ours-resnet*; and 2) with the adversarial loss denoted as *Ours-adv*.

Implementation details: We implement the same architecture appeared in the CelebA experiments in [2] for both VAE and β -TCVAE. For our proposed method, we replace the convolutional layers with bottleneck residual blocks for both *Ours-resnet* and *Ours-adv*, while the additional discriminator network in *Ours-adv* has the same architecture of the encoder except for the last layer which has a single output. All models are trained using Adam optimizer for 100K iterations with a minibatch size of 256, and a learning rate of $1e - 4$. β and d are set to 6 and 32, respectively. Note that we employ leakyReLU in our *Ours-adv* which has been successfully applied in the adversarial training literature.

Comparison with state-of-the-art: We compare our method with the recent state-of-the-art methods by reporting the evaluation metrics (cf Table.1). We notice improvements over the β -TCVAE in terms of disentanglement with a relative improvement of 81.6% and 161.8% on MIG_p and MIG_{hf} , respectively. For reconstruction error, it is expected that VAE would be superior to other models because there is no extra focus on the prior constraining part of the loss function which allows reconstruction error to optimize better. However, we notice an improvement on PSNR compared to β -TCVAE model which compromises reconstruction error for disentanglement. This experiment shows that adding the bottleneck residual blocks together with adversarial training not only improves the disentanglement, but also improves the reconstruction quality.

Use-cases: In order to show that the disentangled representation is rather capturing some meaningful generative factors, which might be relevant to the task at hand. We design two use-cases in both unsupervised and supervised paradigms. For the clustering use-case, we utilize the learned representations to fit a Gaussian Mixture Model (GMM) with 7 components and assign a label to each data point. NMI is then calculated between assigned labels and ground-truth labels.

Table 1: Comparison of various representation learning models.

	$MIG_p\%$	$MIG_{hf}\%$	PSNR	NMI%	ACC%
VAE	5.23	2.74	22.91	9.12	67.88
β -TCVAE	6.92	3.53	20.79	10.66	68.61
Ours-resnet	11.61	5.89	19.42	9.89	69.19
Ours-Adv	12.57	9.24	21.18	11.86	70.02

We report an average of 10 realizations. Regarding the classification use-case, we utilize the learned representations of a few amounts of labeled data to train a multi-layer perceptron (MLP) on 10% of the data and evaluate it on the remaining 90% of the data. 10-fold stratified cross-validation is performed.

The model gives a relative improvement of 11.6% and 2% on the NMI and ACC, respectively. This could be attributed to the quality of the learned representation where features responsible for the pathologies are captured by disentanglement models as generative factors.

Interpretability: We qualitatively examine the interpretability of the learned representations by manipulating the latent code. For instance, Fig. 1a shows a comparison of the traversal between the proposed model and VAE. We notice that the dimension responsible for changing skin color has some entanglement with eccentricity and size in the case of VAE. In contrast, we can see in our proposed model that the size and eccentricity are barely changed when the skin color dimension is changed. For eccentricity, we notice in the case of VAE that fewer variations are captured such as the absence of the horizontal elliptic lesions that are captured with the proposed approach.

In Fig. 1b, we show the possibility of generating images with specific features by smoothly moving over the manifold of the representations. We show the transition of a small lesion on pale skin to a big horizontal lesion on reddish skin by changing multiple latent dimensions responsible for each feature. Having this control over the representation does not only give the ability to generate images with specific known features, but also gives an interpretable representation of the data which can be utilized in many applications.

4 Discussion

In this paper, we introduce a novel adversarial VAE with a total correlation constraint to enforce disentanglement on the latent representation while preserving the reconstruction fidelity. The proposed framework is evaluated on skin lesions dataset and shows improvements over other state-of-the-art methods in terms of disentanglement. The learned representations have shown remarkable performance in both unsupervised clustering and supervised classification. One interesting direction to investigate is the usage of few labels to enhance the disentanglement and allow for better strategies for model selection as suggested in [12, 13].

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1798–1828 (2013)
2. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: *Advances in Neural Information Processing Systems*. pp. 2615–2625 (2018)
3. Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: *Advances in neural information processing systems*. pp. 2172–2180 (2016)
4. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: *European conference on computer vision*. pp. 630–645. Springer (2016)
6. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: *International Conference on Learning Representations* (2017)
7. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: *International Conference on Artificial Neural Networks*. pp. 44–51. Springer (2011)
8. Kim, H., Mnih, A.: Disentangling by factorising. *arXiv preprint arXiv:1802.05983* (2018)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
10. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. In: *Advances in neural information processing systems*. pp. 2539–2547 (2015)
11. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* (2015)
12. Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359* (2018)
13. Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., Bachem, O.: Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258* (2019)
14. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* **19**(6), 1236–1246 (2017)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 618–626 (2017)
16. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**, 180161 (2018)
17. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* (2018)

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 24, 2025

License Number 5995550403277

License date Mar 24, 2025

Licensed Content Publisher Springer Nature

Licensed Content Publication Springer eBook

Licensed Content Title Learning Interpretable Disentangled
Representations Using Adversarial VAEs

Licensed Content Author Mhd Hasan Sarhan, Abouzar Eslami, Nassir
Navab et al

Licensed Content Date Jan 1, 2019

Type of Use Thesis/Dissertation

Requestor type academic/university or research institute

Format electronic

Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Learning Transferable Descriptive Disentangled Representations
Institution name	Technical University of Munich
Expected presentation date	Mar 2025
The Requesting Person / Organization to Appear on the License	Mohammad Hasan Sarhan
Requestor Location	Mohammad Hasan Sarhan Boltzmannstraße 3 Garching bei München, Bayern 85748 Germany
Billing Type	Invoice
Total	0.00 EUR

Multi-scale Microaneurysms Segmentation Using Embedding Triplet Loss

Mhd Hasan Sarhan^{1,2}, Shadi Albarqouni¹, Mehmet Yigitsoy², Nassir Navab^{1,3}, Abouzar Eslami²

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Summary. This publication proposes a framework for enhancing the segmentation of microaneurysms by applying region proposals on multiple scales and refining the proposals with metric learning trained classifier. The rationale behind using multiple scales for proposing the lesions is two fold, first microaneurysms appear in different sizes and detecting them on multiple scales would improve the detection process, and second different scales has different semantic and pixel resolutions which leads to better view around lesions without losing the spacial resolution. Using embedding metric loss is beneficial in this scenario to learn a descriptive representation of microaneurysms that can well differentiate patches of healthy and microaneurysms content. The framework allow for good segmentation of the lesion without segmentation information about other lesions and without the need to have each instance of a lesion to be annotated in all images. The method is evaluated on a microaneurysms segmentation dataset and proved to perform well compared with naively using fully convolutional networks on one scales. Moreover the contribution of the multi-scale effect and the usage of metric learning are studied separately to understand the effect of each contribution on the overall performance. The model that uses multi-scale region proposals with a classification refinement network training with triplet embedding loss performed better than vanilla models and each contribution separately.

Contribution. The contribution of the author of this thesis to this publication consists of the main idea, implementation, experiment design and evaluation. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details in implementation as well as the revision of the publication was done jointly with the co-authors.

Copyright Statement. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer, Cham, Lecture Notes in Computer (LNCS, volume 11764), 22nd MICCAI Conference, Shenzhen, China, October 13–17, 2019 Proceedings, Multi-scale Microaneurysms Segmentation Using Embedding Triplet Loss, Mhd Hasan Sarhan, Shadi Albarqouni, Mehmet Yigitsoy, Nassir Navab, Abouzar Eslami, © 2019.

The following text is reprinted with the permission of the publisher (License number

5995550387656). It is the accepted but not the published version of the paper. The license is attached after the publication text.

Multi-scale Microaneurysms Segmentation Using Embedding Triplet Loss

Mhd Hasan Sarhan^{1,2}[0000–0003–0473–5461], Shadi Albarqouni¹[0000–0003–2157–2211], Mehmet Yigitsoy²[0000–0001–6598–0933], Nassir Navab^{1,3}, and Abouzar Eslami²[0000–0001–8511–5541]

¹ Computer Aided Medical Procedures, Technical University of Munich, Germany

² Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. Deep learning techniques are recently being used in fundus image analysis and diabetic retinopathy detection. Microaneurysms are an important indicator of diabetic retinopathy progression. We introduce a two-stage deep learning approach for microaneurysms segmentation using multiple scales of the input with selective sampling and embedding triplet loss. The model first segments on two scales and then the segmentations are refined with a classification model. To enhance the discriminative power of the classification model, we incorporate triplet embedding loss with a selective sampling routine. The model is evaluated quantitatively to assess the segmentation performance and qualitatively to analyze the model predictions. This approach introduces a 30.29% relative improvement over the fully convolutional neural network.

Keywords: Deep Learning · Segmentation · Ophthalmology

1 Introduction

Diabetic retinopathy (DR) is the leading cause of vision impairment and blindness for middle-aged groups [11]. DR early detection is important for the treatment planning. Severity of DR falls into one of five levels (none, mild, moderate, severe, or proliferative) [1]. Microaneurysms are considered as the first signs for detecting early stages of DR. Hence, detecting these lesions is important for Computer Aided Diagnosis systems. Microaneurysms are abnormalities in the microvascular structure and appear as small red dots in color fundus images. Screening programs use colored fundus images of the retina for their rich information and ease of access. Detecting microaneurysms in colored fundus images is a challenging task due to the small size of the lesion which makes up less than 1% of the entire image [3], and the low contrast between microaneurysms and background.

Microaneurysms are the strongest determinant for DR since they are the first lesion that appears during the early stages. Various approaches for microaneurysms detection using deep learning are proposed [7, 14, 10]. These methods are patch-wise approaches and use deep architectures to extract representative

features. These features could be added to a set of hand-crafted features [14] and passed to a classification model or used solely in an end-to-end network [7], [10]. Deep learning techniques in the literature of microaneurysms detection use random patches selection, hence, they are prone to be biased towards the over-sampled class. Moreover, no work in the microaneurysms segmentation context has leveraged the embedding space of the input patches to impose an additional constraint on the learning process.

Contributions: In this work, a multi-scale patch-wise approach for segmenting microaneurysms in retinal fundus images is proposed. The main contributions of this work are 1) fusing segmentation on multiple scales for microaneurysms detection, and 2) using embedding triplet loss [16] with selective sampling [6] to increase the descriptiveness of the feature representation while focusing the training on informative examples. The model is agnostic to other lesions (i.e. the model differentiates between healthy and microaneurysm patches regardless of information about other lesions). Being agnostic to other lesions is important in such cases as it may be difficult to obtain an annotated dataset with all DR lesions annotated.

2 Methodology

Our proposed microaneurysms segmentation framework, depicted in Fig. 1, consists of two stages; the *hypothesis generation network* (**HGN**), where multi-scale fully convolutional networks (FCNs) are employed to propose a region of interest (ROI), and *patch-wise refinement network* (**PRN**), where extracted patches around ROIs are passed to the classifier. In the next sections we introduce the details of the applied method. First, we go through the fully convolutional hypothesis generation networks, the reasoning behind having multiple scales, and the details of the loss function used for optimization. The second section is dedicated for the PRN. In which, the motive behind this network is explained and the details of triplet loss and selective sampling are presented.

Hypothesis Generation Network (HGN): High-resolution fundus images where a microaneurysm covers a very small part of the image are examined to segment microaneurysm. Using a zoomed-in patch would allow for high spatial accuracy on account of losing semantic information, whilst a zoomed-out patch would have a richer semantic representation on the account of losing spatial resolution [4]. As a trade-off, we use equally sized patches on two scales of the image to build two HGNs, one for each scale.

HGN is a fully convolutional neural network trained on patches of size 256×256 extracted from the fundus images. Two HGNs are trained for two different scales of the fundus image (1x, 0.5x). This allows the extraction of scale-related features while at the same time preserve full resolution image information. The architecture used is the full resolution residual network type A [15] for its good results in segmentation.

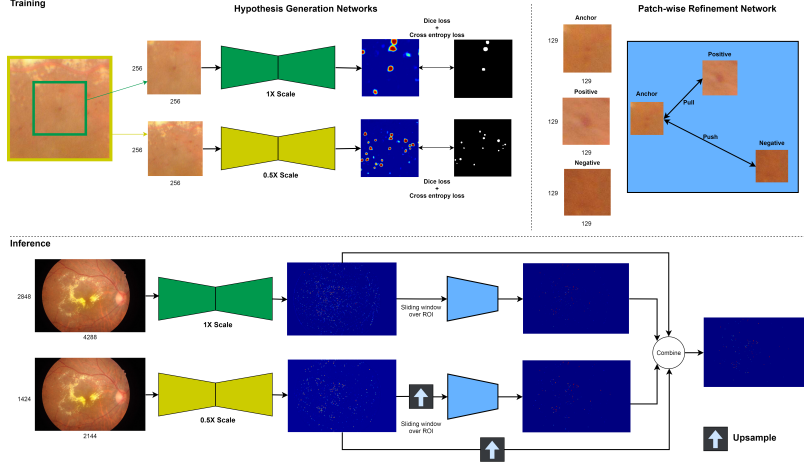


Fig. 1: Pipeline for the multi-scale microaneurysms segmentation framework. Top part shows the pipeline in the training mode where each model is trained separately. Bottom part shows the pipeline at inference time where the image is used rather than patches for the hypothesis generation networks.

To select the training patches, we define images that contain no signs of DR as healthy (negative) images and images with microaneurysms as lesion (positive) images. Healthy pixels are extracted only from healthy patients' scans and lesion pixels are extracted from DR patients at the microaneurysms locations. As a loss function, weighted cross entropy loss is used to compensate for the imbalance negative and positive patches. Moreover, dice loss is optimized to enhance the spatial overlap between a segmentation map output and the gold standard segmentation. We use a differentiable approximation of the dice loss as in [12].

Patch-wise Refinement Network PRN is a classification network that is used as on top of the HGN. The input of the network is an image patch and the output is the probability of the patch center pixel being a microaneurysm or healthy. The segmentation maps of the HGN are used as regions of interest for the PRN. The architecture of classification networks allows for receptive fields larger than fully convolutional networks that consume more memory because of the decoder part and skip connections. The larger receptive field allows for feature maps that incorporate more spatial information about the image which enriches the extracted features. The architecture employed for this network is an adopted version of the Resnet-50 [8]. One downsampling step is omitted from the original architecture because the input image size in our case is smaller than what is expected in the Resnet-50 scenario. In the training phase, patches are extracted from images in the same manner of extracting 1x resolution patches for HGN. The only difference is the size of PRN patches is 129×129 .

To extract discriminative features in PRN we propose the utilization of triplet loss [16]. Triplet loss is applied on the embedding of a patch around pixel x into a d -dimensional feature space. The aim of triplet loss is to make similar patches closer to each other in the embedding space while pushing dissimilar patches away from each other in the embedding space using a predefined distance measure. We found the feature representation of the last convolution layer after the global average pooling (GAP) as a good representation in the embedding space due to its high descriptive power while having a compact representation. The optimization of triplet loss requires three input patches namely the anchor patch x^a , the positive patch x^p and the negative patch x^n . The goal is to make the embedding of the positive patch closer to the anchor patch than the embedding of the negative patch. The loss is defined as

$$\mathcal{L}_{triplet} = \sum_i^N [d(f(x_i^a), f(x_i^p)) - d(f(x_i^a), f(x_i^n)) + a]_+ \quad (1)$$

where a is a margin to enforce a distance between positive and negative pairs, $d(.,.) \in \mathbb{R}^1$ is the distance measure in the embedding space, and N is the number of all possible triplets. As a distance measure, angular cosine distance is utilized as it shows better performance on high dimensional representations when training deep networks [13]. In addition to triplet loss, cross entropy loss for patches is optimized.

Generating all triplets, in this case, would be computationally prohibitive. Moreover, the imbalance in the dataset is high. To counter these problems, we use selective sampling [6]. This approach of training proved to enhance the results in training scenarios where data from different classes are not balanced. In our use case, the healthy class is over-represented. In selective sampling, patches with higher loss have a higher probability of being picked for the next epoch as they are considered representative samples.

3 Experiments

3.1 Experimental Setup

Dataset For our evaluations of the segmentation pipeline, we use the IDRiD⁴ publicly available dataset. All images are captured with the same device that has 50-degree field of view and have size of 4288×2848 pixels. Before patch extraction, the published train dataset is split into two parts: training, and validation sets. The validation set is used for monitoring the training. Table 1 shows the dataset splits.

Implementation details We employ contrast enhancement following the formula $I_{pre}(x, y) = 4I(x, y) - 4G_\sigma * I(x, y) + 1024/30$. To train HGN, we define a mini-batch of size 10 and consider each epoch to be 1000 mini-patches. The learning rate for the full-scale network is $1e - 6$ and for the half-scale network is $1e - 5$.

⁴ <https://idrid.grand-challenge.org/>

Table 1: IDRiD dataset splits

	Healthy		Microaneurysms	
	Images	Patches	Images	Patches
Train set	80	6M	44	~500K
Validation set	9	~6M	10	~132K
Test set	27	-	45	-

PRN is trained with mini-batches of triplets. The size of a mini-batch is 90×3 patches. We sample 90×2 patches from the pool of lesion patches randomly with uniform distribution, and 90×1 samples from the healthy patches pool with selective sampling. This neural network has a Siamese structure [2], this means that each part of the triplet’s three parts is run through identical versions of the network and the gradients are combined at the output to update the weights of the network. In addition to triplet loss, cross entropy loss for pairs is optimized. To this end, we optimize the cross-entropy loss between the anchor and the negative pair. Every 1000 mini-batches is considered as an epoch. We run selective sampling routine every 10 epoch, this is because of the big number of training patches. which takes a significant amount of time to evaluate. Learning rate is set to $1e - 5$ and decreased by a factor of 10 after 20 epochs. The optimization of the losses is done using Adam optimizer [9].

3.2 Multi-scale effect

In this evaluation, we study the effect of using multiple scales. To this end, two HGNs are trained, one for the full resolution image and one for the downsampled image by a factor of two. The evaluation is done on the publicly published test set images. We compare results from each scale with the results of combining the two scales in two different ways. First the output of the half scale is upsampled using linear interpolation, then the prediction maps of the two scales are combined either with pixelwise arithmetic or geometric averaging. The results of this evaluation are presented in Table. 2. FCN 1x, FCN 0.5x represent the evaluation on the prediction map of the full scale and half scale HGNs, respectively. FCN geometric and FCN arithmetic refer to the results of combining the two scales with geometric and arithmetic averaging, respectively. The results show that combining the two scales gives better performance either way. We notice a higher recall from the half scale network but lowest precision, this reflects that the model is very sensitive to microaneurysms and generates a high number of false positives that drops down the overall performance.

3.3 Patch-wise refinement and triplet loss effect

We evaluate the effect of 1) using a classification network to refine the classifications of HGNs and 2) using triplet loss in the classification network to refine HGNs results (i.e PRN). To evaluate the classification network, we utilize patches

from the image in a sliding window fashion and use the classification probability of each point to obtain segmentation maps. It is worth noting that sliding window does not go over all the image, but only the parts higher than a preset probability threshold (0.5 in our case) from HGN. Two segmentation maps will be obtained by sliding over the image masked with two HGNs outputs. Two prediction maps from two HGNs and two prediction maps from refining HGNs results with the classification networks combined as shown in Figure 1.

We first demonstrate the effect of incorporating a classification network to refine the results of HGNs. To this end, we use an edited version of PRN that uses only cross entropy loss without the triplet embedding optimization. This network is denoted as *cls*. Using the classification network on top of the fully convolutional networks enhances the results of the overall segmentation. The larger receptive field allows for more descriptive representations which in turn could suppress false positives that are triggered by HGNs. The effect of utilizing triplet embedding loss is then evaluated by training a PRN using triplets from the training set. We set the margin value a from Equation 1 to 0.5. At test time, this network is utilized in a sliding window fashion similar to *cls*.

Using triplet loss in a multi-scale approach with geometric averaging has an overall 30.29% PR AUC improvement over the baseline fully convolutional neural network trained with weighted cross entropy. The improvement when incorporating triplet loss could be attributed to the quality of the learned representations where lesion patches are forced to be close to each other with a certain margin of difference from healthy ones.

Our results come in 4th place in the IDRiD challenge outdated leaderboard based on the metric used on the released test set. The challenge submission is currently closed. iFLYTEK-MIG used Mask-RCNN to segment 3 lesions at the same time. VRT used a U-net to segment four lesions all together. PATech used a patch-wise approach with false positives bootstrapping on lesions simultaneously. We notice that in all these models, information about lesions other than microaneurysms is utilized. This makes the disambiguation between lesion types (e.g. hemorrhages and microaneurysms) learned inherently in the model but has the drawback of requiring full annotation of multiple lesion types. The proposed model does not require information from other lesions to be trained.

3.4 Visual evaluation

We study the misclassifications of the model by visually examining samples of the results. In Figure. 2 an example of a segmentation is presented. From the example, we notice that false positives mostly lay in the area around hemorrhages or on top of a blood vessel where a higher intensity occur. False negatives are more difficult to be detected because they sometimes appear very close to hemorrhage and blend in or the contrast in the image is low enough to lose the microaneurysm. In the top left example, we see a false negative example where the microaneurysm is misclassified because of very light edges and irregular shape that leans towards hemorrhage. In the other false negative examples (in cyan), the cases are very difficult to be distinguished and variability between

Table 2: Ablation test for emphasizing each part of the pipeline

	AUC PR	F1-score	Precision	Recall
HGN 1x - baseline	0.3374	0.3618	0.2970	0.4626
HGN 0.5x	0.3411	0.4001	0.4380	0.3682
HGN geometric	0.3622	0.3866	0.5115	0.3108
HGN arithmetic	0.3701	0.4156	0.4741	0.3701
<i>cls</i> geometric	0.3895	0.4153	0.5402	0.3374
<i>cls</i> arithmetic	0.3905	0.4368	0.4973	0.3895
PRN arithmetic	0.3978	0.4323	0.54051	0.3602
PRN geometric	0.4196	0.38477	0.61128	0.2807
IDRiD iFLYTEK-MIG	0.5017	-	-	-
IDRiD VRT	0.4951	-	-	-
IDRiD PATech	0.4740	-	-	-

raters may occur in such cases. The bottom right example shows a false positive example where a darker area around the bright exudates appears similar to microaneurysm. The variability in illumination parameters of the capturing device has also a significant effect on the training and may lead to a bias towards a certain image appearance. It is important to note that images in the IDRiD dataset are compressed with a lossy compression which leads to big jumps in intensity values next to each other.

4 Discussion and Conclusion

We hypothesize that using multiple fully convolutional networks for multiple scales of the inputs enhances the segmentation of small objects similar to microaneurysms because it gives a better trade-off between semantic and spatial accuracy. Embedding loss is employed mainly in learning image descriptors [17]. We use the triplet embedding loss in our model to treat deeper layers of the classification network as a local descriptor of the keypoint represented by the healthy or microaneurysm patch. The classification performance increases by adding this additional constraint on the features created by the network.

The segmentation results could be used in report generation for the doctors or in future studies to do big data analysis of populations. Microaneurysms turnover is also an important factor in the progression analysis of DR [5] and could be studied better with reliable models for Microaneurysms segmentation.

References

1. American Academy of Ophthalmology. international clinical diabetic retinopathy disease severity scale detailed table. <http://www.icoph.org/downloads/Diabetic-Retinopathy-Detail.pdf>, accessed: 10.09.2018
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a” siamese” time delay neural network. In: Advances in Neural Information Processing Systems. pp. 737–744 (1994)



Fig. 2: An example of a segmented microaneurysms in a fundus image. Green is for true positives, red is for false positives, and cyan is for false negatives.

3. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* **124**(7), 962–969 (2017)
4. Ghiasi, G., Fowlkes, C.C.: Laplacian pyramid reconstruction and refinement for semantic segmentation. In: *European Conference on Computer Vision*. pp. 519–534. Springer (2016)
5. Goatman, K.A., Cree, M.J., Olson, J.A., Forrester, J.V., Sharp, P.F.: Automated measurement of microaneurysm turnover. *Investigative ophthalmology & visual science* **44**(12), 5335–5341 (2003)
6. van Grinsven, M.J., van Ginneken, B., Hoyng, C.B., Theelen, T., Sánchez, C.I.: Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging* **35**(5), 1273–1284 (2016)
7. Haloi, M.: Improved microaneurysm detection using deep neural networks. *arXiv preprint arXiv:1505.04424* (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
10. Lam, C., Yu, C., Huang, L., Rubin, D.: Retinal lesion detection with deep learning using image patches. *Investigative ophthalmology & visual science* **59**(1), 590–596 (2018)
11. Lee, R., Wong, T.Y., Sabanayagam, C.: Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision* **2**(1), 17 (2015)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *3D Vision (3DV), 2016 Fourth International Conference on*. pp. 565–571. IEEE (2016)

13. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
14. Orlando, J.I., Prokofyeva, E., del Fresno, M., Blaschko, M.B.: Learning to detect red lesions in fundus photographs: An ensemble approach based on deep learning. arXiv preprint arXiv:1706.03008 (2017)
15. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Fullresolution residual networks for semantic segmentation in street scenes. arXiv preprint (2017)
16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
17. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3109–3118 (2015)

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 24, 2025

License Number 5995550387656

License date Mar 24, 2025

Licensed Content Publisher Springer Nature

Licensed Content Publication Springer eBook

Licensed Content Title Multi-scale Microaneurysms Segmentation
Using Embedding Triplet Loss

Licensed Content Author Mhd Hasan Sarhan, Shadi Albarqouni, Mehmet
Yigitsoy et al

Licensed Content Date Jan 1, 2019

Type of Use Thesis/Dissertation

Requestor type academic/university or research institute

Format electronic

Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Learning Transferable Descriptive Disentangled Representations
Institution name	Technical University of Munich
Expected presentation date	Mar 2025
The Requesting Person / Organization to Appear on the License	Mohammad Hasan Sarhan
Requestor Location	Mohammad Hasan Sarhan Boltzmannstraße 3 Garching bei München, Bayern 85748 Germany
Billing Type	Invoice
Total	0.00 EUR

Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations

Mhd Hasan Sarhan^{1,2}, Shadi Albarqouni¹, Mehmet Yigitsoy², Nassir Navab^{1,3}, Abouzar Eslami²

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Summary. This publication proposes utilizing the multi-scale microaneurysms segmentation framework proposed by the author for the detection of diabetic retinopathy and referable diabetic retinopathy. Segmentation embedding from multiple scales of the data are extracted. Such representation are assumed to hold information about microaneurysms locations, numbers, and shapes that are beneficial for detecting early signs of diabetic retinopathy as the presence of microaneurysms is crucial for decision making. A vanilla classification network is trained on the task of grading diabetic retinopathy. The representation of data learned by the classification network is utilized along with the multi-scale representations from the segmentation network to build a new representation suitable for the grading task. The representations are concatenated and passed to random forest classifier to learn the grading task. The performance of the detection of diabetic retinopathy and referable diabetic retinopathy is evaluated on both the vanilla classification network, and the combined random forest. The performance on the combined representation model is superior to the vanilla classifier. This suggests that using microaneurysms segmentation information in the detection of diabetic retinopathy and referable diabetic retinopathy improves the detection performance.

Contribution. The contribution of the author of this thesis to this publication consists of the main idea, implementation, experiment design and evaluation. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details in implementation as well as the revision of the publication was done jointly with the co-authors.

Copyright Statement. © 2020 IET. Reprinted, with permission, from Mhd Hasan Sarhan, Shadi Albarqouni, Mehmet Yigitsoy, Nassir Navab, Abouzar Eslami, Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations, IET Image Processing, December 2020.

The following text is reprinted with the permission of the publisher (License number 5995011446167). The license is attached after the publication text.

Microaneurysms Segmentation and Diabetic Retinopathy Detection by Learning Discriminative Representations

M.H. Sarhan^{1,2} S. Albarqouni¹ M. Yigitsoy² N. Navab^{1,3} E. Abouzar²

¹ Computer Aided Medical Procedures, Technical University of Munich, Germany

² Carl Zeiss Meditec AG, Munich, Germany

³ Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

* E-mail: hasan.sarhan@tum.de

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Abstract: Deep learning techniques are recently being used in fundus image analysis and diabetic retinopathy detection. Microaneurysms are an important indicator of diabetic retinopathy progression. We introduce a two-stage deep learning approach for microaneurysms segmentation using multiple scales of the input with selective sampling and embedding triplet loss. The proposed approach facilitates a region proposal fully convolutional neural network trained on segmented patches and a patch-wise refinement network for improving the results suggested by first stage hypothesis. To enhance the discriminative power of the second stage refinement network, we use triplet embedding loss with a selective sampling routine that dynamically assigns sampling probabilities to the oversampled class patches. This approach introduces a 23.5% relative improvement over the vanilla fully convolutional neural network on the IDRiD segmentation dataset. The proposed segmentation is incorporated in a classification model to solve two downstream tasks for diabetic retinopathy detection and referable diabetic retinopathy detection. The classification tasks are trained on the Kaggle diabetic retinopathy challenge dataset and evaluated on the Messidor data. We show that adding the segmentation enhances the classification performance and achieves comparable performance to state of the art models.

1 Introduction

The importance of diabetic retinopathy (DR) detection is increasing as it is the leading cause of vision impairment and blindness for middle aged groups [25]. To prevent blindness caused by DR, screening programs with early detection and referral of the disease are a widely adopted strategy. The treatment planning of DR relies on the detection of early signs of the disease. According to the International Clinical Diabetic Retinopathy Disease Severity scale [1], DR severity falls into one of five levels (none, mild, moderate, severe, or proliferative). Microaneurysms are an indicator of the early stages of DR [44], this makes the detection of microaneurysms an important pathway towards Computer Aided Diagnosis systems for timely DR detection. Colored fundus images are widely used in DR screening programs because they are rich in information and easy to access. Microaneurysms are visible in color fundus images as small red dots that represent abnormalities in the microvascular structure of the retina. Because of their small size which constitutes less than 1% of a fundus image [12], it is a challenging task to detect these small red dots reliably. Moreover, the low contrast between microaneurysms and background imposes more difficulties on the detection of these abnormalities. Figure. 1 shows an example of a color fundus image with two highlighted microaneurysms, one highlighted hemorrhage, and one crossing of two vessels that could be confusing to automatic microaneurysms detection systems.

Background: We first describe common approaches for detecting DR in fundus images and then discuss various approaches for the detection of microaneurysms in fundus images. For DR detection, one approach is to detect multiple lesions and anatomies of the fundus (e.g. red lesions, blood vessels or optic disk) then fuse this information into a prediction model to detect DR or referable DR (more than mild). In [36], lesions are detected using feature banks and results are fused with a k-nearest neighbour classifier. Abramoff *et al.* [2, 3] use deep learning for extracting the lesions and anatomy features then utilize random forests to fuse the results into a binary decision. Gulshan *et al.* [17] propose an end-to-end model to detect

referable DR directly from a fundus image using an ensemble of 10 Inception-V3 models and over 100k images annotated by 3-7 annotators. The model is trained for referable DR, severe or worse DR, referable Diabetic Macular Edema classification, and image gradability. Fusion of deep extracted features with image metadata (e.g. image field of view) is used to boost the classification performance [12]. In [43], the authors propose Zoom-in-net for DR grading. Zoom-in-net utilizes attention mechanisms to direct the training into focusing on pathological areas of the fundus image. This allows the model to make a more informed decision regarding DR grade. It was later proposed to use bounding boxes extra annotation for some lesions to enhance DR grading [26]. The development of reliable approaches for segmenting lesions responsible for the pathology are needed for more reliable and interpretable decision making diagnostic models.

Microaneurysms presence is a strong determinant for DR since they are the first lesion that appears during the early stages [44]. Automatic methods for detecting microaneurysms exist in the literature. Various methods use handcrafted features for the detection [8, 24, 39]. Such methods are decomposed into two parts. First, microaneurysms candidates are detected, then, handcrafted features that utilize color and shape information are extracted around the candidates. In [24], a local rotating feature is used around the candidate microaneurysms to detect true positive lesions. In [8], a gradient-based local convergence index feature is used to classify microaneurysms. The gradient of a patch is supposed to converge into a point (or more) within the candidate microaneurysm to be considered a true microaneurysm. Seoud *et al.* [39] proposed a method for red lesions (microaneurysms and hemorrhages) detection using dynamic shape features. More recently, various approaches for microaneurysms detection using deep learning techniques are proposed in the literature [18, 23, 31]. These methods operate on the patch level and use deep architectures to extract representative features. The extracted features are either fused with handcrafted features and passed to a classification model [32], or used directly in classification deep networks [18, 23] for center pixel classification.

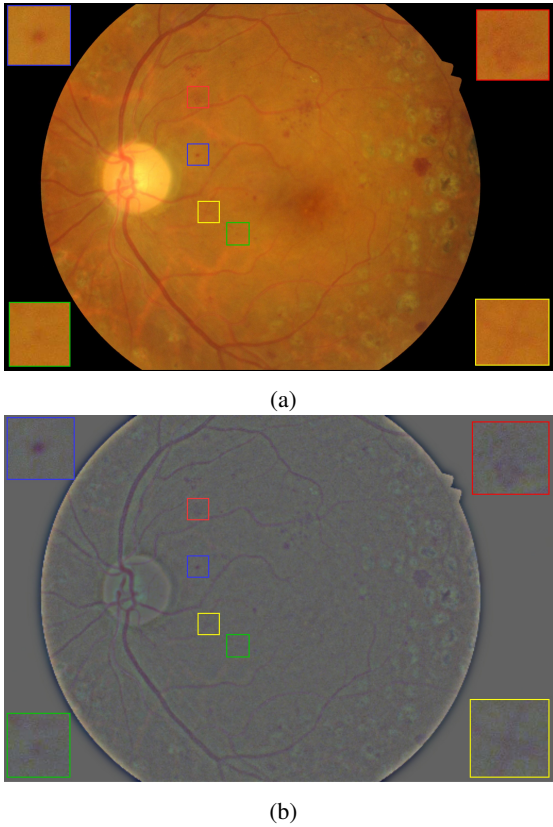


Fig. 1: Example of a fundus image. Blue and green squares contain microaneurysms, red square contains a hemorrhage, and yellow square contains vessels crossing. (a) Original fundus image; (b) preprocessed image.

Microaneurysms vary in shape and color shade and could be confused with other lesions such as hemorrhages. Learning a representation that is invariant enough to intra-class changes in microaneurysms' features (e.g. shape, color, illumination, background) while being discriminative to differentiate microaneurysms from healthy and similar lesions is a crucial part in distinguishing microaneurysms. To achieve such properties in representation learning tasks, metric learning techniques are widely adopted [11, 38]. Such techniques apply constraints on the learned image representation rather than on the prediction. These constraints are designed in such a way that similar samples are closer to each other than to dissimilar samples according to a predefined distance measure. Since these properties are desired in the discussed microaneurysms segmentation task, we utilize metric learning constraints on the learned embeddings to achieve more discriminative representations. Pairwise contrastive loss and triplet loss are examples of commonly used metric losses. Contrastive loss forces similar inputs to be close to each other (possible with a predefined margin) while triplet loss forces similar inputs to be close and dissimilar ones to be farther apart. We use triplet loss because it is widely adopted and achieves competitive results for its wide adopting and competitive results [38].

Learning from patches in medical imaging could suffer from class imbalance problems. Pathological samples are under-represented compared to healthy ones. Some approaches resolve the problem by adding balancing factors to the loss function. This could be done by assigning a higher weight to misclassifications in detecting the under-represented class [32]. Other way to counter imbalance is to oversample the under-represented class by balancing the ratios of the classes in each training batch [4]. Lastly, it is desirable to apply a selection strategy when balancing the sampling of the classes. This is important especially for the over-represented class since many samples might be uninformative or repetitive and it is important to use the more informative samples for a faster convergence and better results [41]. To counter biases towards over-represented classes

and utilize the most informative over-represented samples, we use selective sampling routines.

Contributions: In this work, we propose a framework for small lesions segmentation in retinal fundus images. We apply the framework to the task of microaneurysms segmentation. The framework incorporates patches sampled from two scales of the fundus image to balance spatial and semantic resolutions. We employ metric learning techniques to learn discriminative microaneurysms representations that are invariant to microaneurysms differences but able to differentiate microaneurysms from healthy patches. The training is balanced and focused on the most informative samples by employing over-sampling with selective sampling to ensure unbiased results. The model does not require annotations about other lesions nor does it require all instances of the lesion of interest to be annotated. This is an important property since it is difficult in some cases to annotate other lesions or even annotate all instances of a certain lesion. We show the importance of microaneurysms segmentation by using the segmentation results to create a diabetic retinopathy referral and detection model and show how it performs compared to a baseline that does not incorporate any lesion information.

Our preliminary work [37] has explored multi-scale microaneurysms segmentation with embedding metric loss and selective sampling and showed promising results. In this work we further validate efficacy of the model by utilizing it in two downstream applications namely DR and referable DR detection. We show that incorporating the segmentation information into the classification model enhances the downstream tasks. We also put the embedding triplet loss under the microscope and run a sensitivity analysis study on the effect of the margin hyperparameter on the training. A more thorough state of the art study is performed and a detailed explanation of the methodology is provided.

2 Methodology

The proposed microaneurysms segmentation framework is composed of two stages; the *hypothesis generation network (HGN)*, where fully-convolutional networks (FCNs) are utilized as a region proposal module, and *patch-wise refinement network (PRN)*, where a classifier refines the outputs of the segmentation probability maps suggested by HGN. A depiction of the framework is shown in Figure. 2.

In the next sections we introduce the details of the applied method. Firstly, we describe a simple preprocessing step for contrast enhancement to emphasize red lesions in the fundus image. Then, we define the hypothesis generation networks, the reasoning for utilizing multiple scales, and the loss function used for optimization. The third section is dedicated for the PRN, in which, the motivation behind this network is presented and the triplet loss and selective sampling are outlined.

2.1 Preprocessing

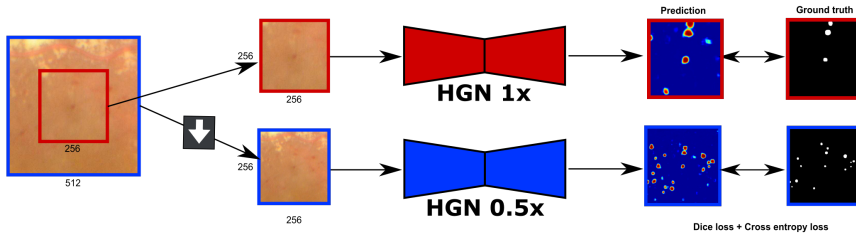
A contrast enhancement is applied to all images before patch extraction following the formula appeared in [16]. $I_{pre}(x, y) = \alpha I(x, y) + \beta G_{\sigma} * I(x, y) + \gamma$. where $*$ is the convolution operator, x, y are pixels, I is the image, I_{pre} is the contrast enhanced image and G_{σ} is a Gaussian filter with scale σ . The constants α , β , and γ control the contribution of each term and are chosen empirically.

2.2 Hypothesis Generation Network (HGN)

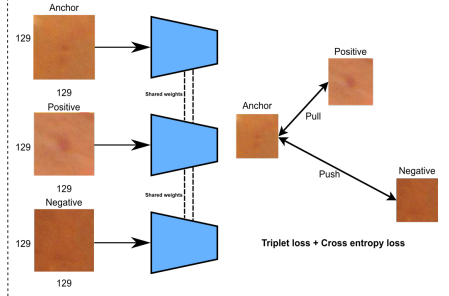
A microaneurysm covers a very small part of a high-resolution fundus image. HGN goal is obtaining microaneurysms segmentation maps corresponding to these images. Since the resolution of the images is high, operating on full resolution patches would allow for high spatial accuracy on account of losing semantic information and using downsampled patches would have a richer semantic representation while losing spatial resolution [13]. To account for both, we use equally sized patches on two scales of the image and combine

Training

Hypothesis Generation Network (HGN)



Patch-wise Refinement Network (PRN)



Inference

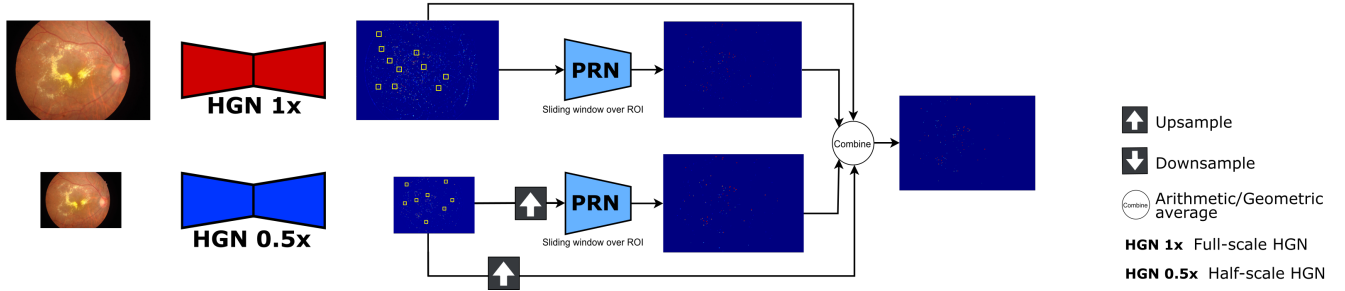


Fig. 2: The proposed microaneurysms segmentation framework is composed of two components; the *hypothesis generation network (HGN)*, where fully-convolutional networks (FCNs) are utilized as a region proposal module, and *patch-wise refinement network (PRN)*, where a classifier refines the outputs of the segmentation probability maps suggested by HGN. The training is done on patch-level while the test (inference) could be done on the image-level. The top part shows the training phase where each component is trained separately. HGN 1 $\times$, HGN 0.5 $\times$ are trained on patches sampled from the full size and downsampled version of the image respectively. HGN color corresponds to the scale. PRN is separately trained on triplets of patches using embedding triplet loss. The bottom part shows the **inference phase**. The probability maps of the HGNs predictions are used as region proposals for the PRN patch classification. The HGNs and PRN outputs are then combined with either arithmetic or geometric average to obtain the final segmentation map.

the results with the outputs of the PRN. Each of the two scales is trained with a different HGN before the results of the two scales are combined. One HGN is trained with the full scale patches and the other is trained with half scale patches that are sampled from down-sampled images. Given a preprocessed image $I_{pre} \in \mathbb{R}^{H \times W}$, HGN is a fully convolutional neural network trained either on patches $P_{1x}^{(i)} \in \mathbb{R}^{256 \times 256}$ extracted from full resolution I_{pre} or on patches $P_{0.5x}^{(i)} \in \mathbb{R}^{256 \times 256}$ extracted from I_{pre} after a downsampling step by 2. Each of the two networks is trained independently on one of the scales. This allows the extraction of scale-specific features while preserving the full resolution image information. The architecture used is the full resolution residual network type A [33] for its good results in segmentation.

Healthy training patches are extracted around pixels of healthy fundus images that contain no signs on DR. Microaneurysms patches are extracted from DR fundus images around pixel locations of microaneurysms. The full scale and half scale HGNs have exactly the same architecture but trained separately with no weight sharing. For training these HGNs the weighted cross entropy loss (Eq. 1) is used to compensate for the imbalance between background pixels and microaneurysms pixels in the dataset. This balancing is used for the HGN training while oversampling techniques with selective sampling is used for PRN training.

$$\mathcal{L}_{ce}(l, p_l) = - \sum_{x \in \Omega} [w_0 l_x \log(p_0(x)) + w_1 (1 - l_x) \log(p_1(x))] \quad (1)$$

where $\mathcal{L}_{ce}$ is the weighted cross-entropy loss, Ω is a set containing all samples x . w_0, w_1 are weights for the negative and positive classes respectively, $l_x \in \{0, 1\}$ is the label of sample x and $p_l(x)$ is the predicted probability for x to have label l . Moreover, we optimize dice loss to enhance the spatial overlap between a segmentation map output and the gold standard segmentation. We use a differentiable

approximation of the dice loss that is suitable for neural networks training as in [28]

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_{x \in \Omega} p_l(x) g_{t_l}(x)}{\sum_{x \in \Omega} p_l^2(x) + \sum_{x \in \Omega} g_{t_l}^2(x)} \quad (2)$$

Where $g_{t_l}(x)$ is the gold standard one-hot label. For regularization we use the Frobenius norm on network weights θ . The total loss for training the hypothesis generation network becomes as follows

$$\mathcal{L}_{hgn} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{dice} + \lambda_3 \|\theta\|_F^2 \quad (3)$$

where $\lambda_1, \lambda_2, \lambda_3$ are balancing operators for cross-entropy, dice, and regularization losses respectively, and $\|\theta\|_F^2$ is the Frobenius norm on network weights θ . Both full scale and half scale HGNs are trained with $\mathcal{L}_{hgn}$ loss. The output of each HGN is a probability map for each pixel being a microaneurysm or not. These probability maps are referred to as HGN segmentation maps in the reminder of the manuscript. When these probability maps are binarized by thresholding at a cutoff value, they are referred to as binary segmentation maps.

2.3 Patch-wise Refinement Network

The patch-wise refinement network is a classification network that is utilized after the HGN output is obtained. The input to this module is an image patch and the output is the probability describing whether the patch center is a microaneurysm or healthy patch.

The segmentation maps of the HGN are used as regions of interest for the PRN. This network follows standard classification deep neural networks architecture where multiple convolutional and pooling layers are stacked and followed by a fully-connected neural network for classification decision. This architecture allows for larger receptive fields compared to fully convolutional networks. This is

because the latter exhausts more memory given the decoder part and skip connections. The larger receptive field incorporates more spatial information in the feature maps which improves the extracted features. For PRN architecture, a revised version of the Resnet-50 [20] is employed. We omit one downsampling step since the patch size in our case is smaller than original Resnet design. We use Resnet-50 for its good results in large scale classification tasks [35] and the ability to extract high dimensional rich features that could be further optimized with the embedding loss. Similar to training the HGN, healthy pixels are extracted only from healthy patients scans and lesion pixels are extracted from diabetic retinopathy patients scans at the microaneurysms locations. The training patches are created around the extracted pixels with 129×129 pixels size. It is important to differentiate the training and inference routines of PRN. For the training, patches are extracted from images in similar manner to HGN patch extraction but the patch size is smaller. These patches are used for training a classification network namely PRN. This is done using triplets of inputs with embedding triplet loss. In the inference time, the network expects one input patch and returns a probability of the patch center being a microaneurysm. The patches are extracted from a test image in locations where the HGN outputs are salient. This is done to reduce the time of sliding over each pixel in test images and only focus on promising pixels.

PRN purpose is to discard false positive regions proposed by the HGN. Since microaneurysms differ in shape, size and background illumination while could be confused with other lesions and structures such as hemorrhages and vessels crossing, we aim to learn a representation that encapsulates this information for better classification of the lesion. We utilize triplet embedding loss for the purpose of learning discriminative representations that describe well microaneurysms while being invariant to intra-class differences. Triplet loss is applied to the representation being learned that maps input patches to a d -dimensional feature space. In deep neural networks, the embedding is chosen to be the feature representation of deeper layers. We found the feature representation of the last convolution layer after the global average pooling (GAP) as a good representation in the embedding space due to its high descriptive power while being compact.

To elaborate mathematically, suppose the PRN approximates a function g that maps a patch around pixel x to a binary classification decision. We decompose g to $g = h(f(x))$ where f is the embedding function that maps an input patch to d -dimensional vector (i.e. GAP output), h maps the learned representation to the desired binary classification decision. The embedding is represented as $f(x) \in \mathbb{R}^d$. The optimization of triplet loss requires three input patches namely the anchor patch x^a , the positive patch x^p and the negative patch x^n . The aim of triplet loss is to bring x^p closer to x^a while pushing x^n further in the embedding space. Microaneurysms patches are used as anchor and positive patches, while healthy patches are used as negative patches. The loss is defined as

$$\mathcal{L}_{triplet} = \sum_i^N [d(f(x_i^a), f(x_i^p)) - d(f(x_i^a), f(x_i^n)) + a]_+ \quad (4)$$

where a is a margin to enforce a distance between positive and negative pairs, $d(.,.) \in \mathbb{R}^1$ is the distance measure in the embedding space, x_i^a, x_i^p, x_i^n are anchor, positive, and negative patches of triplet i . N is the number of all possible triplets. As a distance measure, we choose the angular cosine distance as in [5] because it shows better performance on high dimensional representations when training deep networks [29]. In addition to triplet loss, cross entropy loss for patches is optimized. It is calculated similar to Eq. 1 but on the level of patches rather than pixels such as,

$$\mathcal{L}_{ce_prn} = - \sum_{x^p \in \mathcal{X}^+} \log[g(x^p)] - \sum_{x^n \in \mathcal{X}^-} \log[1 - g(x^n)] \quad (5)$$

where $\mathcal{X}^+, \mathcal{X}^-$ are the set of microaneurysms and healthy patches respectively. $g(.)$ is the prn network that maps an input patch

Algorithm 1 Selective sampling routine

- 1: Let $\mathcal{X}^+$ be a set of microaneurysms training patches and $\mathcal{X}^-$ a set of healthy training patches ▷ Initial conditions
 - 2: Train network for V epochs by randomly sampling M triplets (x^a, x^p, x^n) where $x_i^a, x_i^p \in \mathcal{X}^+, x_i^n \in \mathcal{X}^-, i = 1, 2, \dots, M$ ▷ Initial training on a random subset of the dataset
 - 3: Run feedforward pass for each $x_i \in \mathcal{X}^-$ to get a probability $g(x^n)$ ▷ Inference step to calculate probabilities
 - 4: Assign a weight to each $x_i \in \mathcal{X}^-$ such that $w_i = g(x^n)$ ▷ Weigh each patch according to network output probability
 - 5: Let $p_i \leftarrow \frac{w_i}{\sum_{x_j \in \mathcal{X}^-} w_j}$ ▷ Assign a sampling probability to each patch
 - 6: Train network for V epochs by sampling M triplets (x^a, x^p, x^n) where $x_i^a, x_i^p \in \mathcal{X}^+$ are sampled randomly and $x_i^n \in \mathcal{X}^-$ are sampled with probability p_i ▷ Training by randomly sampling positive patches and selectively sampling negative patches
 - 7: Repeat steps 3-6 until convergence
-

to a microaneurysm probability. The total loss for the patch-wise refinement network becomes as follows

$$\mathcal{L}_{prn} = \eta_1 \mathcal{L}_{ce_prn} + \eta_2 \mathcal{L}_{triplet} + \eta_3 \|\phi\|_F^2 \quad (6)$$

where $\mathcal{L}_{ce_prn}$ is the cross entropy loss, $\|\phi\|_F^2$ is the Frobenius norm on PRN weights ϕ . η_1, η_2, η_3 are balancing hyper-parameters for cross-entropy, triplet, and regularization losses respectively.

There is a high number of patches to select from. It is computationally prohibitive to consider every triplet. Moreover, the dataset is heavily imbalanced towards the healthy patches. To counter these problems, we use selective sampling. In selective sampling, patches with microaneurysms are sampled randomly from the dataset; however, healthy patches are sampled with a probability proportional to the error correlated with the patch in a previous epoch. This way, not all combinations of triplets should be considered and the training is more focused on triplets with informative healthy examples. Selective sampling is described in Algorithm 1. Let $\mathcal{X}^+$ be a set of microaneurysms training patches and $\mathcal{X}^-$ a set of healthy patches. First, M triplets of patches (x^a, x^p, x^n) are selected randomly from microaneurysms and healthy samples sets such that $x^a, x^p \in \mathcal{X}^+, x^n \in \mathcal{X}^-$. The model is trained for a predefined number of epochs V with $\mathcal{L}_{prn}$ as a loss function and the sampled triplets as training data. After that, all samples of $\mathcal{X}^-$ are fed to PRN in a forward pass and the prediction errors for all samples are calculated. Based on the prediction error, a sampling probability is assigned to each sample of $\mathcal{X}^-$ giving higher sampling probability to samples with larger prediction error. For the following training epochs, $x^n \in \mathcal{X}^-$ are sampled based on the assigned probabilities while $x^a, x^p \in \mathcal{X}^+$ are uniformly randomly sampled. The process of training for V epochs and re-assigning sampling probabilities to $\mathcal{X}^-$ patches is repeated until convergence.

3 Experiments

3.1 Experimental Setup

Datasets: To evaluate the segmentation framework shown in Figure 2, we use the publicly available Indian Diabetic Retinopathy Image Dataset (IDRiD) [27]. All images are captured with Kowa VX-10 alpha digital fundus camera with 50-degree field of view (FOV). Each image size is 4288×2848 pixels. The published dataset is split into train and test sets. For model hyper-parameters selection we split the published training set into training and validation sets. As described in Sections 2.2, 2.3, healthy patches are only extracted from healthy scans and lesion patches are extracted from diabetic eyes. Table 1 shows the dataset splits employed for our training purposes.

Table 1 IDRiD dataset splits

	Healthy		Microaneurysms	
	Images	Patches	Images	Patches
Train set	80	6M	44	~500K
Validation set	9	~6M	10	~132K
Test set	45	-	27	-

To evaluate image-level DR detection and DR referral tasks we utilize two datasets. First we use the publicly available Kaggle DR challenge dataset*. The dataset contains ~35k training and ~11k validation fundus images. Each image is graded based on the International Clinical Diabetic Retinopathy Disease Severity Scale [1] five-level DR grading. The Kaggle dataset is used for training the image-level model for its large size and the large variability in image capture conditions and devices. For the validation of the image-level classification, we use the publicly available Messidor dataset [9] because it is widely used in the literature for external validation of DR and referable DR detection. The dataset contains 1200 images graded from 0 to 3. We follow [42] to perform the investigated binary classification tasks (referable V.S. non-referable DR and normal V.S. any-level DR) where referable DR is defined as having level 2 or more DR grading.

Implementation details: We use contrast enhancement from Section 2.1 with the parameters set to $\alpha = 4$, $\beta = -4$, $\gamma = 128$, and $\sigma = 1024/30$. Figure 1 shows an example of a fundus image and how it appears after contrast enhancement. To train the HGN, we define a mini-batch of size 10 and consider each epoch to be 1000 mini-patches. We train the network until no further improvement is observed on the validation set weighted cross-entropy and dice losses. We set the weighted cross entropy parameters in Eq. 1 to $\frac{w_0}{w_1} = 80$. This ratio is decreased every two epochs by 20 until it reaches the value 20. The loss weighting parameters in Eq. 3 are set to $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 4e - 5$. The network architecture for the full and half scales is the architecture of the full resolution residual networks for semantic segmentation type A [33]. The learning rate for the full-scale network is $1e - 6$ and for the half-scale network is $1e - 5$. Batch normalization is used after convolution to reduce covariate shifts and prevent overfitting [21]. The rectified linear unit function (ReLU) is used as an activation function. Down-sampling is done by performing a convolution with a stride of 2 and up-sampling is done with the convolution transpose operation.

PRN is trained with mini-batches of triplets. The size of a mini-batch is 90×3 patches. We sample 90×2 patches from the pool of lesion patches randomly with uniform distribution, and 90×1 samples from the healthy patches pool with a selective sampling probability (step 5 in Algorithm 1). PRN has a Siamese structure [7]. In a Siamese structure, the anchor, positive, and negative patches are fed to the network to calculate the embeddings, then the triplet loss is applied to the embeddings and backpropagated to update the weights. In addition to triplet loss, cross entropy loss is optimized. We consider the anchor and negative patches as input to cross entropy to ensure even distribution of both classes. Every 1000 mini-batches are considered as an epoch. The network is trained until no further improvements in the triplet nor cross-entropy losses are observed. We run selective sampling routine every 10 epochs, this is because of the big number of training patches which takes a significant amount of time to be run through the network for inference. Learning rate is set to $1e - 5$ and decreased by a factor of 10 after 20 epochs. The hyperparameters in Equation 6 are set to $\eta_1 = \eta_2 = 1$, $\eta_3 = 1e - 4$. Variables weights in HGNs and PRN are initialized using Xavier initializer which allows signals to reach deeper layers of the network [14]. In PRN, the weights are initialized with ImageNet [10] weights. Both are optimized with Adam optimizer [22].

For the DR grading experiment we use inception-resnet-v2 [40] architecture. The network is trained on the Kaggle DR challenge

dataset for 5-level grading. Based on the 5-level grading we then derive the tasks of no DR (level 0) vs. DR signs (level 1 or more) and non-referable DR (level 0 or 1) vs. referable DR (level 2 or more). The training is performed with weighted categorical cross entropy loss where the weight of each class is the inverse frequency of the class in the train dataset. SGD optimizer with momentum is used for training. We choose an initial learning rate of 0.005 and decayed exponentially with a decay rate of 0.9 every 5 epochs. The weights are initialized from a pre-trained network on the ImageNet dataset [10].

All development is done with Python3.5 and tensorflow library for training and evaluating deep models. The machine had an NVIDIA TITAN X GPU, Intel(R) Core(TM) i7-6900K CPU and 64 GB of RAM.

Evaluation Metrics: For evaluating the segmentation performance, we use the area under the curve (AUC) of the precision-recall curve because it handles well the high data imbalance and it is used officially for evaluating IDRiD challenge dataset on the segmentation task. We use precision, recall and F1-score as an additional indicator of the model performance. The binary segmentation map is obtained by thresholding the final combined segmentation probability map at 0.5. We calculate precision and recall according to the equations $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$ where TP , FP , FN are the number of true positives, false positives and false negatives, respectively. F1-score is $F1 = 2 \frac{precision \times recall}{precision + recall}$. For the DR image-level classification we use the Receiver operating characteristic (ROC) curve to evaluate the model. In the next sections, various evaluations are performed. We evaluate the effect of using two scales for the hypothesis generation network. We also study the need for the patch-wise refinement network and the effect of training this network with embedding triplet loss. In our third experiment, we evaluate the influence of the margin parameter from Equation 4 on the model. We then experimentally show that using segmentation embeddings in the DR grading model enhances the results for both screening and referable diabetic retinopathy detection. Lastly, we interpret the results visually and explain the results.

3.2 Multi-scale effect

In this evaluation, we study the effect of utilizing multiple scales on the HGNs. Since the networks are fully convolutional, it is possible at test time to evaluate a full fundus image rather than patches. We feedforward the full image to the full-scale HGN and a downsampled version of the image to the half scale HGN. The evaluation is done on the publicly published test set images. The results of each scale are compared to results of combining both scales in two different ways. Before combining results of the scales, the output probability segmentation map of the half scale is upsampled using linear interpolation to have similar resolution as the full scale. The prediction maps from the two scales are combined either with pixelwise (1) arithmetic averaging $arithmetic = (scale_1 + scale_2)/2$ or pixelwise (2) geometric averaging $geometric = \sqrt{scale_1 \times scale_2}$. A PRN is neither trained nor evaluated in this experiment. The results of this evaluation are presented in Table. 2. HGN 1x, HGN 0.5x are the HGNs results on the full scale and half scale scenarios, respectively. HGN geometric and HGN arithmetic refer to the results of combining the two scales with geometric and arithmetic averaging, respectively. The results show that combining the scales results in a better AUC PR in both combination strategies, which goes in-line with our hypothesis that combining semantic and spatial resolutions would result-in better segmentation results. Geometric averaging gives a good trade-off of precision-recall which is shown by the better results of AUC PR. It lowers the recall more than the arithmetic one because it suppresses regions with low (close to zero) predictions coming from either scale which leads to a lower recall and a lower number of false positives. HGN 1x is referred to as baseline network since incorporating a fully convolutional neural network to solve the segmentation problem is the natural first solution to the problem.

*<https://www.kaggle.com/c/diabetic-retinopathy-detection>

Table 2 Multi-scale effect on HGNs. Evaluation is done without the use of the patch-wise refinement network.

	AUC PR	F1-score	Precision	Recall
HGN 1x - baseline	0.3374	0.3618	0.2970	0.4626
HGN 0.5x	0.3411	0.4001	0.4380	0.3682
HGN geometric	0.3622	0.3866	0.5115	0.3108
HGN arithmetic	0.3701	0.4156	0.4741	0.3701

3.3 Patch-wise refinement and triplet loss effect

In this section we perform two ablation tests to evaluate 1) the effect of adding a classification network without triplet loss to filter HGN results, and 2) the effect of integrating triplet embedding loss in the classification network as in Section. 2.3 (PRN). PRN is evaluated by extracting patches from images in a sliding window fashion. Only patches with high probability HGN output (more than 0.5) are considered as input to PRN. This drastically lowers the amount of computation required by PRN. The segmentation map of the half scale HGN is upsampled before extracting high response patches for classification. At test time, we combine the prediction map pair from the HGNs with that of the PRN as shown in Figure 2.

Classification network: The effect of incorporating a classification network to refine the results of HGNs is first demonstrated. To this end, an edited version of PRN that uses only cross-entropy loss without the triplet embedding optimization is utilized. This network is denoted as *cls*. We use selective sampling to favor informative healthy patches while training *cls*. Using *cls* to filter HGN maps enhances the results of the overall segmentation. The extracted features by *cls* are more descriptive than HGN due to the larger receptive field which results in better precision-recall trade-off.

Triplet loss: Evaluating the effect of constraining the embedding with triplet loss is performed by training a PRN with triplets sampled from the training set following to selective sampling routine (cf. Algorithm. 1). The margin a in Eq. 4 is set to 0.5. The results of adding triplet loss to the classification network are shown in Table. 3. The results show that adding triplet loss improves the results of the segmentation compared to classification network trained only with cross-entropy. This could be attributed to the more discriminative representations learned by the network when using metric learning. These representations when used for classification allow for better separation between classes. Adding PRN with triplet loss in a multi-scale fashion with geometric averaging introduces an overall 23.5% PR AUC relative improvement (calculated as $(new - old)/old$) over the baseline fully convolutional neural network trained with weighted cross entropy and dice loss (cf. Table.3). The improvement is due to multiple factors which are 1) the exploitation of information in two scales which allows spatial and semantic feature extraction, 2) the deeper classification model that enables the filtering of false prediction by HGNs and 3) the quality of the learned representations by enforcing the triplet constraint that brings lesion patches closer to each other and further from the healthy ones. Another reason that could hinder vanilla segmentation networks compared to the proposed approach is the constraint of omitting information related to other lesions. This increases the difficulty of the problem as some lesions might be correlated or have a similar appearance. Having no information about these other lesions creates a challenge to build a discriminative representation for the studied lesion (microaneurysms in our case). We hypothesize that the addition of the triplet loss to constrain the representation mitigates the negative effect of limited availability of other lesions annotation. On average it takes 2.5 seconds to segment a full fundus image of size 4288×2848 on an NVIDIA TITAN X GPU.

Comparisons: The results come in 4th place in the IDRiD challenge outdated leader board based on the metric used on the released test set. The challenge submission is currently closed. iFLYTEK-MIG team [34] segmented 3 lesions simultaneously using Mask-rcnn [19]. In VRT team [34] method, an edited U-net architecture is used to segment all 4 lesions available in the dataset. A patch-wise approach is used by PATech team [34] to classify lesions with false positives bootstrapping. All three winners models utilize information about more than one lesion simultaneously (i.e. not only

Table 3 Effect of using *cls* and PRN triplet loss on the overall performance.

	AUC PR	F1-score	Precision	Recall
HGN 1x - baseline	0.3374	0.3618	0.2970	0.4626
HGN arithmetic	0.3701	0.4156	0.4741	0.3701
<i>cls</i> geometric	0.3895	0.4153	0.5402	0.3374
<i>cls</i> arithmetic	0.3905	0.4368	0.4973	0.3895
PRN (triplet) arithmetic	0.3978	0.4323	0.54051	0.3602
PRN (triplet) geometric	0.4196	0.38477	0.61128	0.2807
IDRiD iFLYTEK-MIG	0.5017	-	-	-
IDRiD VRT	0.4951	-	-	-
IDRiD PATech	0.4740	-	-	-
IDRiD SDNU	0.4111	-	-	-

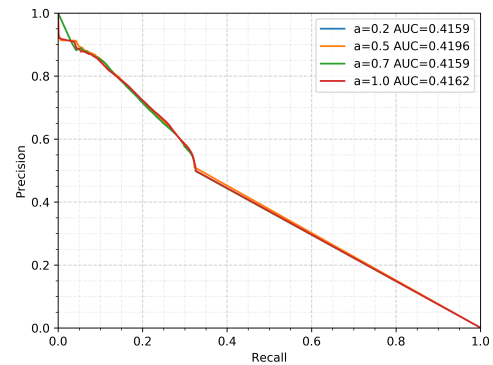


Fig. 3: Precision recall curves for various margins of triplet loss.

microaneurysms information is given to the model). This allows the models to directly learn the differences between ambiguous challenging lesions (e.g. hemorrhages and microaneurysms) which in turn enables better learning outcomes. This comes at the limitation of annotation overhead as all lesions must be annotated in the images for the model to perform well. Our proposed framework requires only information related to microaneurysms to perform well regardless of other types of lesions.

3.4 Triplet margin effect

To study the influence of the margin value a (cf. Equation. 4) in the triplet loss term on the overall model performance, we train 4 different PRNs, while keeping the HGNs fixed. Each of the PRNs is trained with a different triplet margin value. The values for training are 0.2, 0.5, 0.7, and 1.0. We plot the precision-recall curve associated with each network final segmentation and compare area under the curve values to find the best performing network. We use geometric averaging in this experiment to merge the four probability segmentation maps (two coming from HGNs and two correspondingly from PRN). Figure. 3 shows the four PR curves for the four margin values. We notice an improvement when increasing the margin from 0.2 to 0.5 then the performance drops with further increments. Intuitively, a margin of 1.0 should give the best results since it ensures the highest possible cosine distance between positive and negative samples. Since we only use negative examples from healthy images, no other lesions are fed to the network except for microaneurysms (unless it is close to microaneurysms spatially). The model does not directly learn to differentiate microaneurysms from other pathological lesions (e.g. Hemorrhage). When the distance between microaneurysms and healthy patches is emphasized, it risks that other lesions might fall in the microaneurysms cluster which could explain the lower results with more distance constraints.

3.5 Image level classification

Segmenting retinal lesions is important for various tasks such as DR progression analysis and image registration. One important aspect

is the usage of the segmentation of retinal lesions in automatic DR detection algorithms [31]. We conduct an experiment to quantitatively show that adding microaneurysms segmentation results to a DR classification model enhances the overall classification performance. To this end, we use two datasets, one to train and evaluate the model, and the other is for external validation which is finally reported. For the training, the Kaggle DR challenge dataset is used. We train the classification model on preprocessed fundus images from the training set of the Kaggle DR. We use the validation set of Kaggle DR for model selection. IDRI dataset is not included in the training of the vanilla classification model. In order to evaluate the effect of adding the segmentation features to the classification model, we use the embeddings of the trained HGNs on IDRI dataset as segmentation features and concatenate them with the classification embeddings. We consider the bottleneck layer of the HGN as an embedding representation of the image segmentation information. This forms a feature representation of an input image that represents the segmentation and classification feature extraction. We train a random forest classifier on the combined feature representation on the task of DR grading. For the evaluation, we use an external dataset that has not been used in the training process which is the Messidor dataset. We derive two clinically relevant tasks for the evaluation of the model. The first is the DR screening task which is the detection of images with any signs of DR (level 0 vs. level 1 or more). The other task is the referable DR detection (levels 0,1 vs. level 2 or more) which is an important task to detect patients who need to be referred for immediate attention. We compare the vanilla classification and the classification+segmentation models together and with state-of-the-art models. The considered models are all evaluated on the Messidor dataset similar to ours. They are trained for binary classification or segmentation and binary classification tasks as described below. IDRI is not used for training the compared models. The aim of this experiment is two-fold, 1) first, evaluation of how a vanilla classification model for DR detection improves when presented with information regarding the segmentation of lesions that represent early signs of DR. 2) Second, is to compare the results of using the segmentation information in the classification model with other state-of-the-art models that use the same dataset for evaluation but use different techniques/datasets for the training process. Table. 4 shows a comparison between various methods for screening and referable DR detection. Nandy *et al.* [30] use Gaussian mixture models to build an incremental feature extraction framework for image-level referable DR detection. Vo and Verma [42] propose the usage of hybrid color space with deep models to enhance DR grading. Sanchez *et al.* [36] fuse various methods for detection of lesions (red and bright) and anatomical landmarks (retinal vessels and optic disc) in fundus images to make an informed decision of DR presence and referral. Orlando *et al.* [32] use an ensemble of models to detect red lesions and utilize it for DR screening and referral detection. Wang *et al.* [43] incorporate attention mechanisms to state-of-the-art deep models to focus the training on salient lesions. Lastly, Lin *et al.* [26] augment the annotation of the dataset with bounding boxes of important lesions for some examples and use this annotation to enhance the DR detection. Adding microaneurysms segmentation embeddings to the classification improves the results for both screening and referable DR detection compared to the vanilla classification model. The segmentation embeddings carry information about the shape and number of microaneurysms in the fundus image which are used by the classifier to gain a performance boost.

3.6 Visual evaluation

We study the misclassifications of the model by visually examining samples of the results. In Figure. 4, two examples of segmentation are presented. From the examples, we notice that false positives mostly lay in the area around hemorrhages or on top of a blood vessel where a higher intensity occur. False negatives are more difficult to be detected because they sometimes appear very close to hemorrhage and blend in or the contrast in the image is low enough to lose the microaneurysm. In the top left example in Figure. 4a, a negative

Table 4 Diabetic retinopathy screening and need for referral on the MESSIDOR dataset. *se* values are obtained at specificity=50%

Method	Screening		Referable DR	
	AUC	Se	AUC	Se
Expert A [36]	0.920	0.9450	0.9400	0.9820
Expert B [36]	0.8650	0.9120	0.9200	0.9760
Nandy <i>et al.</i> [30]	-	-	0.9210	-
Vo and Verma (I) [42]	0.8620	-	0.8910	-
Vo and Verma (II) [42]	0.8700	-	0.8870	-
Sanchez <i>et al.</i> [36]	0.8760	0.9220	0.9100	0.9440
Orlando <i>et al.</i> [32]	0.8932	0.9109	0.9347	0.9721
Wang <i>et al.</i> [43]	0.9257	-	0.9570	-
Lin <i>et al.</i> [26]	0.935	-	0.9680	-
Ours without Segmentation	0.8400	0.8800	0.9200	0.9300
Ours with segmentation	0.8790	0.8999	0.9400	0.9843

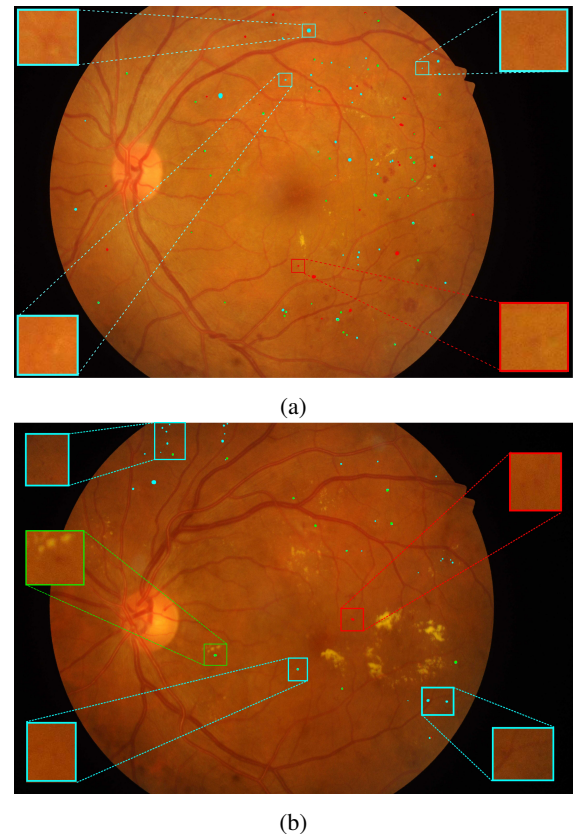


Fig. 4: Two examples of segmented microaneurysms in a fundus image. Green is for true positives, red is for false positives, and cyan is for false negatives.

example is observed. The microaneurysm is misclassified because of very light edges and irregular shape that leans towards a hemorrhage. In the other false negative examples (in cyan), the cases are very difficult to be distinguished and variability between raters may occur in such cases. The bottom right example shows a false positive example where a darker area around the bright exudates appears similar to microaneurysm. The variability in illumination parameters of the capturing device has also a significant effect on the training and may lead to a bias towards a certain image appearance. In the bottom left example in Figure. 4b, a difficult example is shown to be partially successful, the contrast between the microaneurysm and background is very low, the model was able to detect the center part of the microaneurysm (green) while failing to detect the borders that have even lower contrast.

4 Discussion and Conclusion

In this work we proposed a multi-scale microaneurysms segmentation framework with a patch-wise refinement network constraint by embedding triplet loss. Incorporating two scales enhances segmentation of small object because of the better spatial-semantic trade-off it introduces. Adding a classification network after the segmentation refines the results by omitting false positives introduced by the segmentation network. The triplet loss constraint on the embedding further improves the results by learning discriminative representation of microaneurysms patches.

The segmentation results could be used in report generation for the doctors or in future studies for big data analysis of populations. Microaneurysms turnover is also an important factor in the progression analysis of diabetic retinopathy [15] and could be studied better with reliable models for Microaneurysms segmentation. Moreover, microaneurysms segmentation improves the results of diabetic retinopathy screening and the need for referral. This is because of the difficulty of grading models to capture the fine information about microaneurysms. The proposed framework suffers in cases where microaneurysms are close to vessels or have a very low contrast from background. Moreover, in some cases hemorrhages are mistaken for microaneurysms because of their similar color and sometimes size. Training the network is done two separate steps (HGN and PRN) which might produce sub-optimal solutions. One extension of the work could be training the framework in an end-to-end fashion to optimize the losses simultaneously and produce solutions closer to the desired optimum. Finally, it should be noted that this study is not about performing an in-depth optimization of the neural networks' hyper-parameters and architecture. The search for optimal hyper-parameters is a difficult task as described in [6] and not within the scope of this work. Our work focuses on increasing the performance of small lesion segmentation using multiple scales with triplet loss on top of the classification model.

5 References

- American Academy of Ophthalmology. international clinical diabetic retinopathy disease severity scale detailed table. <http://www.icoph.org/downloads/Diabetic-Retinopathy-Detail.pdf>. Accessed: 10.09.2018.
- ABRAMOFF, M. D., LAVIN, P. T., BIRCH, M., SHAH, N., AND FOLK, J. C. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine* 1, 1 (2018), 1–8.
- ABRAMOFF, M. D., LOU, Y., ERGINAY, A., CLARIDA, W., AMELON, R., FOLK, J. C., AND NIEMEIJER, M. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology and Visual Science* 57, 13 (2016), 5200–5206.
- ALBARQOUNI, S., BAUR, C., ACHILLES, F., BELAGIANNIS, V., DEMIRCI, S., AND NAVAB, N. Agnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging* 35, 5 (2016), 1313–1321.
- BAUR, C., ALBARQOUNI, S., AND NAVAB, N. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), Springer, pp. 311–319.
- BERGSTRA, J., AND BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, Feb (2012), 281–305.
- BROMLEY, J., GUYON, I., LECUN, Y., SÄCKINGER, E., AND SHAH, R. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems* (1994), pp. 737–744.
- DASHTBOZORG, B., ZHANG, J., HUANG, F., AND TER HAAR ROMENY, B. M. Retinal microaneurysms detection using local convergence index features. *IEEE Transactions on Image Processing* 27, 7 (2018), 3300–3315.
- DECENCIÈRE, E., ZHANG, X., CAZUGUEL, G., LAY, B., COCHENER, B., TRONE, C., GAIN, P., ORDONEZ, R., MASSIN, P., ERGINAY, A., ET AL. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology* 33, 3 (2014), 231–234.
- DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), Ieee, pp. 248–255.
- FROME, A., SINGER, Y., AND MALIK, J. Image retrieval and classification using local distance functions. In *Advances in neural information processing systems* (2007), pp. 417–424.
- GARGEYA, R., AND LENG, T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 124, 7 (2017), 962–969.
- GHIASI, G., AND FOWLKES, C. C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision* (2016), Springer, pp. 519–534.
- GLOROT, X., AND BENGIO, Y. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (2010), pp. 249–256.
- GOATMAN, K. A., CREE, M. J., OLSON, J. A., FORRESTER, J. V., AND SHARP, P. F. Automated measurement of microaneurysm turnover. *Investigative ophthalmology & visual science* 44, 12 (2003), 5335–5341.
- GRAHAM, B. Kaggle diabetic retinopathy detection competition report univ. warwick, 2015.
- GULSHAN, V., PENG, L., CORAM, M., STUMPE, M. C., WU, D., NARAYANASWAMY, A., VENUGOPALAN, S., WIDNER, K., MADAMS, T., CUADROS, J., ET AL. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.
- HALOI, M. Improved microaneurysm detection using deep neural networks. *arXiv preprint arXiv:1505.04424* (2015).
- HE, K., GKIOXARI, G., DOLLÁR, P., AND GIRSHICK, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969.
- HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.
- IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
- KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- LAM, C., YU, C., HUANG, L., AND RUBIN, D. Retinal lesion detection with deep learning using image patches. *Investigative ophthalmology & visual science* 59, 1 (2018), 590–596.
- LAZAR, I., AND HAJDU, A. Retinal microaneurysm detection through local rotating cross-section profile analysis. *IEEE transactions on medical imaging* 32, 2 (2013), 400–407.
- LEE, R., WONG, T. Y., AND SABANAYAGAM, C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision* 2, 1 (2015), 17.
- LIN, Z., GUO, R., WANG, Y., WU, B., CHEN, T., WANG, W., CHEN, D. Z., AND WU, J. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), Springer, pp. 74–82.
- MERIAUDEAU, P. P. S. P. R. K. M. K. G. D. V. S. F. Indian diabetic retinopathy image dataset (idrid), 2018.
- MILLETARI, F., NAVAB, N., AND AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on* (2016), IEEE, pp. 565–571.
- NAIR, V., AND HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.
- NANDY, J., HSU, W., AND LEE, M. L. An incremental feature extraction framework for referable diabetic retinopathy detection. In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on* (2016), IEEE, pp. 908–912.
- ORLANDO, J. I., PROKOFYEVA, E., DEL FRESNO, M., AND BLASCHKO, M. B. Learning to detect red lesions in fundus photographs: An ensemble approach based on deep learning. *arXiv preprint arXiv:1706.03008* (2017).
- ORLANDO, J. I., PROKOFYEVA, E., DEL FRESNO, M., AND BLASCHKO, M. B. An ensemble deep learning based approach for red lesion detection in fundus images. *Computer methods and programs in biomedicine* 153 (2018), 115–127.
- POHLEN, T., HERMANS, A., MATHIAS, M., AND LEIBE, B. Fullresolution residual networks for semantic segmentation in street scenes. *arXiv preprint* (2017).
- PORWAL, P., PACHADE, S., KOKARE, M., DESHMUKH, G., SON, J., BAE, W., LIU, L., WANG, J., LIU, X., GAO, L., ET AL. Idrid: Diabetic retinopathy-segmentation and grading challenge. *Medical image analysis* 59 (2020), 101561.
- RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., ET AL. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- SÁNCHEZ, C. I., NIEMEIJER, M., DUMITRESCU, A. V., SUTTORP-SCHULTEN, M. S., ABRAMOFF, M. D., AND VAN GINNEKEN, B. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative ophthalmology & visual science* 52, 7 (2011), 4866–4871.
- SARHAN, M. H., ALBARQOUNI, S., YIGITSOY, M., NAVAB, N., AND ESLAMI, A. Multi-scale microaneurysms segmentation using embedding triplet loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), Springer, pp. 174–182.
- SCHROFF, F., KALENICHENKO, D., AND PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 815–823.
- SEOUD, L., HURTUT, T., CHELBI, J., CHERIET, F., AND LANGLOIS, J. P. Red lesion detection using dynamic shape features for diabetic retinopathy screening. *IEEE transactions on medical imaging* 35, 4 (2016), 1116–1126.
- SZEGEDY, C., IOFFE, S., VANHOUCHE, V., AND ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI* (2017), vol. 4, p. 12.
- VAN GRINSVEN, M. J., VAN GINNEKEN, B., HOYNG, C. B., THEELEN, T., AND SÁNCHEZ, C. I. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging* 35, 5 (2016), 1273–1284.
- VO, H. H., AND VERMA, A. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *Multimedia (ISM), 2016 IEEE*

- International Symposium on* (2016), IEEE, pp. 209–215.
- 43 WANG, Z., YIN, Y., SHI, J., FANG, W., LI, H., AND WANG, X. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), Springer, pp. 267–275.
- 44 WILKINSON, C., FERRIS, F. L., KLEIN, R. E., LEE, P. P., AGARDH, C. D., DAVIS, M., DILLS, D., KAMPIK, A., PARARAJASEGARAM, R., AND VERDAGUER, J. T. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 110, 9 (2003), 1677–1682.

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

License Number	5995011446167
License date	Mar 23, 2025
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	IET Image Processing
Licensed Content Title	Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations
Licensed Content Author	Nassir Navab, Eslami Abouzar, Mehmet Yigitsoy, et al
Licensed Content Date	Feb 24, 2021
Licensed Content Volume	14
Licensed Content Issue	17
Licensed Content Pages	8
Type of use	Dissertation/Thesis

Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Title of new work	Learning Transferable Descriptive Disentangled Representations
Institution name	Technical University of Munich
Expected presentation date	Mar 2025
The Requesting Person / Organization to Appear on the License	Mohammad Hasan Sarhan
Requestor Location	Mohammad Hasan Sarhan Boltzmannstraße 3 Garching bei München, Bayern 85748 Germany
Publisher Tax ID	EU826007151
Total	0.00 EUR

Machine Learning Techniques for Ophthalmic Data Processing: A Review

Mhd Hasan Sarhan^{1,2}, M. Ali Nasser⁴, Daniel Zapp⁴, Mathias Maier⁴, Chris P. Lohmann⁴, Nassir Navab^{1,3}, Abouzar Eslami²

¹Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

²Carl Zeiss Meditec AG, Munich, Germany

³Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

⁴Department of Ophthalmology, Klinikum rechts der Isar, Technische Universität München, Munich, Germany

Summary. This work is a review of machine learning works regarding ophthalmic data processing. There, the works regarding multiple ophthalmic related diseases and applications are presented and discussed. The publicly available datasets regarding the discussed applications are also listed and categorized. The discussed diseases are diabetic retinopathy, age-related macular degeneration, and glaucoma. The applications are retinal vessels segmentation in fundus imaging and retinal vessels segmentation in optical coherence tomography images. The review summarizes more than 60 publications and 25 publicly available datasets.

Contribution. The contribution of the author of this thesis to this publication consists of the researching involved methods and summarizing them. Writing of the manuscript was done and coordinated by the author of this thesis. Refinement of the details of disease descriptions and the revision of the publication was done jointly with the co-authors.

Copyright Statement. Reprinted, with permission, from Mhd Hasan Sarhan, M. Ali Nasser, Daniel Zapp, Mathias Maier, Chris P. Lohmann, Nassir Navab, Abouzar Eslami, Machine Learning Techniques for Ophthalmic Data Processing: A Review, IEEE Journal of Biomedical and Health Informatics, December 2020.

The following text is reprinted with the permission of the publisher (published under a Creative Commons license). The license is attached after the publication text.

Machine Learning Techniques for Ophthalmic Data Processing: A Review

Mhd Hasan Sarhan, M. Ali Nasser, Daniel Zapp, Mathias Maier, Chris P. Lohmann, Nasir Navab, Abouzar Eslami

Abstract—Machine learning and especially deep learning techniques are dominating medical image and data analysis. This article reviews machine learning approaches proposed for diagnosing ophthalmic diseases during the last four years. Three diseases are addressed in this survey, namely diabetic retinopathy, age-related macular degeneration, and glaucoma. The review covers over 60 publications and 25 public datasets and challenges related to the detection, grading, and lesion segmentation of the three considered diseases. Each section provides a summary of the public datasets and challenges related to each pathology and the current methods that have been applied to the problem. Furthermore, the recent machine learning approaches used for retinal vessels segmentation, and methods of retinal layers and fluid segmentation are reviewed. Two main imaging modalities are considered in this survey, namely color fundus imaging, and optical coherence tomography. Machine learning approaches that use eye measurements and visual field data for glaucoma detection are also included in the survey. Finally, the authors provide their views, expectations and the limitations of the future of these techniques in the clinical practice.

Index Terms—Ophthalmic Diagnostics, Deep Learning, Diabetic Retinopathy, Age-related Macular Degeneration, Glaucoma

I. INTRODUCTION

DURING the past years, machine learning has been widely employed in both ophthalmic data analysis [1], [2], and ophthalmic image analysis [3], [4], [5]. The application of machine learning for ophthalmic image analysis covered in this survey broadly falls into two categories, 1) Detection and grading of diseases where the objective is to classify the disease found in an image or a set of images [6]. Further works use multi-class classification to grade the level of severity of a given pathology rather than merely detecting it [7]; 2) Segmentation of anatomical structures from pixel-wise labels, such as segmentation of retinal vessels [3], haemorrhages

[8], drusen [9] or optic disk [10]. Analyzing images with machine learning systems involves extraction of features that represent the input data in a way suitable for the classification or segmentation task. Features are either hand-crafted, i.e. defined manually for example by an algorithm developer, or they are extracted with a predefined filter banks, or ultimately learned during training to perform suitable task-specific processing of the input data. The latter is referred to as automatic feature extraction. Recently, after the success of AlexNet in image classification [11], automatic feature extraction using Convolutional Neural Networks (CNNs) surfaced as the preferred feature extraction method in medical image analysis [12] and it is widely used for ophthalmic image analysis [13], [14], [15]. In this review, publications about both feature extraction methods (handcrafted and automatic) are discussed. Methods of Ophthalmic data Analysis for glaucoma analysis based on multiple anatomical and functional measurement are also reviewed in this paper [1], [2].

Classification and segmentation works related to three diseases namely diabetic retinopathy, age-related macular degeneration, and glaucoma are considered in the this review. Furthermore, the recent machine learning approaches used for retinal vessels segmentation, and methods of retinal layers and fluid segmentation are reviewed. Two main image modalities that are widely used to investigate ophthalmic related pathologies are considered in this survey. (1) **Fundus Color** images that show the retinal vasculature, optic disk, macula, and various abnormalities related to ophthalmic diseases. (2) **Optical Coherence Tomography** (OCT) imaging is used to acquire high-resolution cross-sectional scans of the retinal layers in the eye's posterior segment or the anterior segment structure. Only the former is considered in this review. As a subcategory of OCT, few works on **OCT Angiography** (OCTA) are considered in the survey. OCTA is used to examine the microvasculature of the retina by differentiating blood flow from static tissue in OCT over time. Fig. 1 shows fundus and OCT example images with pathologies.

Various works in the literature have been proposed to review deep learning for ophthalmology or a specific ophthalmic disease [16], [17], [18], [19]. To our knowledge, none of the existing works combine technical details with disease/application categorization and list all datasets used in the works to make the reader more

M.H. Sarhan is with the Translational Research Lab, Carl Zeiss Meditec, Munich, Germany and with the Computer Aided Medical Procedures Chair at the Technical University of Munich, Munich, Germany.

M. Ali Nasser, C.P. Lohmann, M. Maier and D. Zapp are with Department of Ophthalmology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

N. Navab is with the Computer Aided Medical Procedures chair at the Technical University of Munich, Munich, Germany, and Whiting School of Engineering, Johns Hopkins University, Baltimore, United States.

A. Eslami is with the Translational Research Lab, Carl Zeiss Meditec, Munich, Germany.

Manuscript received Dec. 09, 2019.

aware of the limitations and strengths of the discussed works. This is beneficial for readers from technical backgrounds who are interested in the ophthalmic data analysis domain and require an overview of the advances, the common diseases and applications, the available datasets, and the current limitations. To select papers in this review, we have looked into the most impactful papers in the past 4 years that appeared in ophthalmology/medical high impact journals/conferences. From these works, we branched to other works and studied them as well and kept branching in the publications graph. We have reached over 100 submissions in total and filtered some of them for the sake of capacity and impact. We included papers that have used a new methodology that has not been seen before to our knowledge. We eventually settled on more than 60 works that are presented in the manuscript.

In each of the following sections, we focus on an ophthalmic pathology and list the major machine learning publications relevant to it with a summary of the main contributions. We then select a number of these methods and discuss them in more details. The results of each method are presented alongside the used datasets in a summary table. Moreover, the public datasets and challenges relevant to the presented publications are presented in tabular format. The works since 2015 are considered in these sections except for the used public datasets which could be older.

For reporting the results of works included in the review, we have used the metrics reported by the included papers. Whenever it is feasible, we report the area under the curve (AUC) of the receiver operating characteristic (ROC) curve because it does not require a cut-off point which gives a more general idea about the classifier performance. For segmentation approaches, we report dice or intersection over union (IOU) whenever feasible. Dice and IOU measure the overlap area between the gold standard segmentation and predicted one. Accuracy, sensitivity, specificity, precision, recall or kappa scores are reported otherwise depending on the paper content. Section II focuses on diabetic retinopathy, detection of the disease and segmentation of various lesions are covered. Section III deals with Age-related Macular Degeneration (AMD) detection and segmentation tasks. Section IV presents glaucoma detection and segmentation. Section V looks at the task of retinal vessel segmentation in fundus and OCT-A images, and in section VI we focus on retinal layer and fluid segmentation in OCT scans. At the end, section VII discusses the limitations and future of machine learning in ophthalmology. A summary of the paper is finally provided in section VIII.

II. DIABETIC RETINOPATHY

Diabetic Retinopathy (DR) is the leading cause of visual impairment in middle-aged groups [21]. Screening for early detection and referral is the widely adopted strategy in practice for blindness prevention. It is still

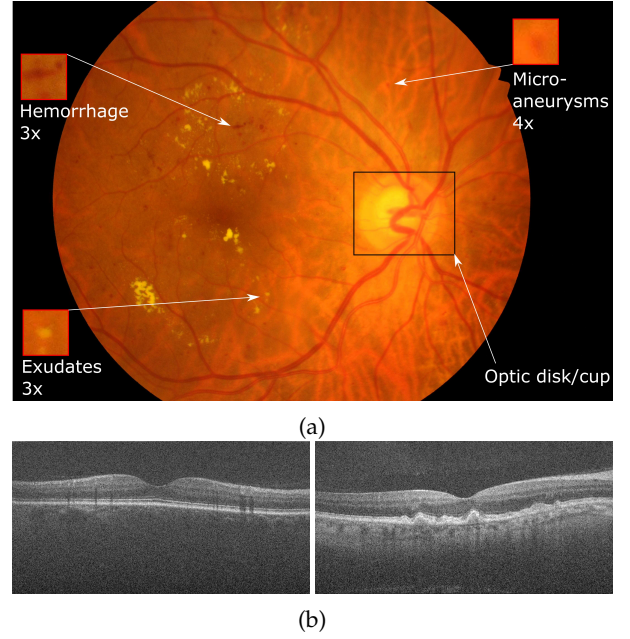


Fig. 1: (a) is an example fundus image from IDRiD dataset with retinal lesions and optic disk/cup highlighted, (b) left is a healthy OCT B-scan cross-section that shows retinal layers, and right is an OCT B-scan of patient with AMD signs [20].

challenging to implement screening programs as needed because of financial issues and the availability of human monitoring to all sites. Hence, the need for reliable automatic systems for detecting DR emerges. The capillaries start to weaken even at the early stages of DR. This causes abnormalities in the microvascular structure leading to microaneurysms (MA), which appear as small red dots on the retina. These weak capillaries may rupture causing hemorrhages (HA), which appear as larger darker red areas. Moreover, DR makes capillaries more permeable causing hard exudates (HE). When the capillary support fails, pale areas with soft edges appear as soft exudates (SE). In more severe cases, dilation appears in the intraretinal microvascular structure. New fragile small vessels start growing in response to the signals from tissues that lack nutrition which could lead to a sudden loss of vision when growing into the vitreous humour. This process is known as neovascularisation [22]. In the following subsections, literature related to the tasks of detection and grading of diabetic retinopathy, segmentation of lesions that characterize diabetic retinopathy, and attempts to use the detection and grading results for weakly supervised lesion localization are discussed. Table. I lists the publicly available datasets for DR detection and segmentation.

A. Detection and Grading of Diabetic Retinopathy

Recent approaches have applied machine learning techniques and convolutional neural networks for the task of Diabetic Retinopathy grading based on image

TABLE I: Publicly available fundus image diabetic retinopathy datasets.

Dataset	Count	Modality	Task	Notes
DiaretDB0 [23]	130: 22 normal, 108 DR	Fundus	Segmentation	Exudates, microaneurysms, haemorrhages segmentation
DiaretDB1 [24]	89: 5 normal, 84 mild signs of DR	Fundus	Segmentation	microaneurysms, haemorrhages, hard and soft exudates
Retinopathy Online Challenge [25]	100	Fundus	Segmentation	Annotations for microaneurysms
E-Ophtha [26]	381: 233 normal, 148 with red lesions	Fundus	Segmentation, Classification	Segmented small red lesions. Images with any red lesion is labeled as DR
Messidor [27]	1200: 4 DR stages	Fundus	Classification	Images are graded into 4 DR stages: healthy, mild, moderate and severe
Messidor-2 [27]	1748: 5 DR stages	Fundus	Classification	Images are graded into 5 ICDRD stages
DR2 [28]	529	Fundus	Classification	Only the presence (not the location) of different types of lesions is provided
Kaggle DR challenge [29]	88702: 5 DR stages split into train, validation, and test sets	Fundus	Classification	Images are graded into 5 ICDRD stages
IDRiD [30]	516 images split into train and test sets	Fundus	Segmentation, Classification	DR grading, pixel level annotation for four lesions and optic disc.

data. Typically, in studies presented in the literature, datasets are annotated for five grades of DR according to the International Clinical Diabetic Retinopathy Disease Severity Scale (**ICDRD**) [44] namely healthy, mild DR, moderate DR, severe DR, and proliferative DR. In this scale, DR is defined as the signs of any DR related lesions which is mild or more DR. Patients with moderate or more severe retinopathy are referred to ophthalmologists for examination and treatment. This level of DR is called referable DR.

In [31], the authors use multiple deep feature extractors for DR lesions and anatomies in fundus (e.g. optic disk). These features are then fused together to get good referable DR classification. The method is trained on a private dataset and evaluated on Messidor-2 dataset. This approach, while performing well, requires multiple steps of segmentation and feature extraction to detect DR. To mitigate the effect of sub-optimal solutions of multi-stage models, Google team [4] have used an end-to-end model consisting of an ensemble of 10 inception networks [45]. The training is done on a private dataset containing 128,175 fundus images provided by EyePACS. The images are graded by 3-7 annotators according to the ICDRD severity scale. The network is trained for a binary classification problem of detecting referable Diabetic Retinopathy. The model is evaluated on the publicly available Messidor-2 dataset and another private dataset of 9963 images provided by EyePACS that does not overlap with the training dataset. To have a more robust detection model that could generalize better to different types of retinas, the authors in [5] use a dataset of multi-ethnic populations with diabetes to develop and evaluate the model. In order to focus the training of lesions and informative areas of the retina, Wang et al. [32] utilize attention mechanisms for DR grading. This enforces the network to be more attentive to discriminative areas of the image for making grading decisions. Furthermore, the areas of focus are cropped and used in higher resolution alongside the full image to

make a more informed decision. Ultra-widefield (UWF) fundus imaging is a newer more advanced technology with the ability to show a wider area of the retina to examine the periphery in addition to the foveal area. This, in turn, allows for better staging of diabetic retinopathy [46]. Since large datasets of UWF fundus images are not yet available, utilization of narrow field fundus large datasets through transfer learning is done to improve the grading of UWF fundus images [34]. Table. II summarizes DR related works by stating the main task, the datasets used, and the results of each work. Results are not directly comparable because of utilizing different datasets and splits.

In a step to make the models more interpretable visually and have some insights about the classification results, weak localization is utilized to generate heatmaps similar to the input image size. In these heatmaps, areas that have more effect on the classification have higher values [14], [42], [35], [43]. In early works, Class Activation Map (CAM)s [47] are used to visualize the prominent areas of the image. CAMs are easy to calculate and widely used in the literature to localize areas of the image that are responsible for specific pathology (e.g. DR) [35], [42], [43]. In [35], the authors use a classification network to differentiate normal fundus images from ones with DR. This networks is then used to localize salient features in the image guiding the DR prediction. CAMs are not only used for qualitative analysis of trained models but are also evaluated quantitatively as in [43] where localization results are evaluated for sensitivity and false positives per image using an annotated dataset on pixel level. To calculate CAMs, specific network design guidelines should be taken into consideration which limits the applicability of the method. Some authors have analyzed the gradients flowing back in the neural network as in [14] where the significance of each pixel is calculated by computing the q-norm of the last layer gradient with respect to the pixel of the channel. Methods for detection and grading of DR have gone a

TABLE II: Summary: Diabetic Retinopathy. AUC is area under the curve of ROC curve. Acc is the accuracy score. Sen is the sensitivity score. FPs/I is false positives per image.

Reference	Task	Summary	Results	Dataset(s)
Gulshan et al. [4]	Classification	Ensemble of 10 inception neural networks for referable DR classification	AUC: 0.99	Messidor-2, private
Abramoff et al. [31]	Classification	Deep learning used for feature extraction and features are used as input to IDx-DR X2.1	AUC: 0.98	Messidor-2, private
Ting et al. [5]	Classification	Learning referable diabetic retinopathy classification from multi-ethnic population dataset	AUC: 0.93	Private >76k fundus images
Wang et al. [32]	Classification	Fusion of multiple attention maps to focus the training of DR grading on pathological areas	AUC: 0.95	Kaggle, Messidor
Lin et al. [33]	Classification	Using a classification model with extra bounding box annotations for some lesions to enhance the detection of DR	AUC: 0.96	Kaggle, Messidor
Sarhan et al. [34]	Classification	Grading of UWF fundus images by utilizing a large dataset of narrow field fundus and transfer learning	AUC: 0.99	Kaggle, private 154 UWF
Gargeya & Leng [35]	Classification & Weak Localization	Two level classification: GAP for visualization, decision tree ensemble for GAP features + image metadata	AUC: 0.94	Messidor-2, E-Ophtha
Yang et al. [7]	Classification & Localization	Two-stage CNN with local lesion detection CNN and a global image level grading deeper CNN	AUC: 0.95	Kaggle
Quellec et al. [14]	Classification & weak localization	Generalization of backpropagation with a sparsity constraint for training using second-order gradients	AUC: 0.95, Lesion: 0.95	Kaggle, E-Ophtha, DiaretDB1
Orlando et al. [15]	Red lesion segmentation	Use a combination of deep and hand crafted features to train a random forest on pixel level segmentation and image level classification	AUC: 0.93	E-Ophtha, DiaretDB1, Messidor
Van Grinsven et al. [8]	Haemorrhages segmentation	Introduce selective sampling for training datasets with unbalanced labels and apply it to haemorrhages segmentation	AUC: 0.97	Kaggle, Messidor-2
Abbasi-Sureshjani et al. [36]	Exudates segmentation	Residual networks with selective sampling for full image segmentation	Dice: 0.83	E-Ophtha-EX, DiaretDB1, DR2
Huang et al. [37]	Neovascularization segmentation	Extreme Learning Machine with pixel-wise filter banks as input features for the model	Acc: 0.89	Messidor-2, DiaretDB0, DiaretDB1
Levenkova et al. [38]	Segmentation in UWF	Lesion segmentation in ultra wide-field fundus imaging using a deep feature extractor and an SVM	AUC: 0.86	Private 146 UWF
Lam et al. [39]	Segmentation	Segmentation of multiple DR lesions in fundus images and comparing various neural networks architectures	Acc: 0.98	Kaggle extra annotation
Guo et al. [40]	Avascular area segmentation	Automatic quantification of retinal non-perfusion area based on OCT angiography $6 \times 6 \text{ mm}^2$ en face angiograms in DR patients	Average dice: 0.87	Private 76 healthy and 104 DR OCTA scans
Sarhan et al. [41]	Microaneurysms segmentation	Multi-scale segmentation of microaneurysms using selective sampling and embedding triplet loss	Dice: 0.43	IDRiD
Wang & Yang[42]	Weak localization	CNN with multiple max-poolings and a GAP layer to visualize class activation maps	Kappa: 0.85	Kaggle
Gondal et al. [43]	Weak localization	Weakly supervised learning to visualize lesions using image-level classification information with increased GAP size and batch normalization	Sen: 0.72 @ 2.25 FPs/I	Kaggle, DiaretDB1

long way and are now performing in competitive levels compared with human annotators. However, there are still issues with trust, domain shifts and explainability of these models. Research directions into utilizing other modalities such as UWF or studying the progression of the disease over time could be promising fields for better understanding DR.

B. Diabetic Retinopathy Segmentation

Further assessment of Diabetic Retinopathy characteristics is done by the identification and segmentation of

lesions that characterize DR such as microaneurysms, haemorrhages, hard exudates, soft exudates and neovascularization. Most current approaches for the segmentation of lesions in fundus images are not holistic and partition the image into patches for processing. These approaches incorporate high-resolution patches of the image to classify each patch into the studied lesions [8], [15], [36]. In such scenarios, the patches coming from healthy samples are plenty compared to patches containing lesions which creates an unbalanced dataset. This could be handled with selective sampling [8] where

a weighting scheme is applied to the selection of negative (over-represented) class samples for training the next epoch. This approach is combined with multi-scale segmentation and embedding triplet loss to preserve spatial and semantic information on one hand and learn a robust representation on the other [41]. In [15], the authors use a hybrid feature vector (a combination of features from a deep neural network model and features extracted in a handcrafted fashion) to detect red lesions (microaneurysms and haemorrhages) in fundus images using random forest classifier. Learning-based methods are also used for the segmentation of lesions characterizing later stages of diabetic retinopathy such as neovascularization [37]. The authors use Extreme Learning Machine (ELM) [48] for segmenting neovascularization in color fundus images and report an accuracy of 89.2%. Ultra-widefield fundus images are used in [38] to detect DR lesions automatically using a deep neural network. Detection and segmentation of DR lesions are challenging tasks especially in the case of small lesions such as microaneurysms. Models that could handle high-resolution images or work in multiple scales of the data are viable solutions to the problem.

III. AGE-RELATED MACULAR DEGENERATION

Age-related Macular Degeneration (AMD) is one of the major causes of vision impairment for the elderly. The choroid and outer layers of retina such as retinal pigment epithelium (RPE) are particularly affected. Damages to the retinal layers that contain photoreceptors could lead to vision loss. There are multiple scales to grade AMD, one commonly used is the Age-Related Eye Disease Study (AREDS) simplified scale [49]. This scale grades AMD into four stages (none, early, intermediate, and advanced/late AMD). Another fine-grained scale from AREDS provides 9 increasing grades of nonadvanced AMD. AMD could be dry or wet. In the wet form of AMD, choroidal neovascularisations (CNV) could grow into the RPE. These new vessels can leak which may cause sudden vision loss. The dry AMD represents atrophy affecting the RPE layer of the retina. In the following subsections, two tasks that have been addressed by researchers will be presented: detection of AMD and segmentation of AMD related structures. Table. III lists the publicly available datasets for AMD detection, retinal layers, and retinal fluids segmentation.

A. Detection of Age-related Macular Degeneration

The automatic detection of AMD is done either with fundus imaging or with OCT volumes and scans. Burlina et al. [60] use fundus images around the macula to extract features using a pre-trained neural network on ImageNet. The features are extracted on multiple scales and concatenated to be used as an input to a support vector machine (SVM). The model is developed for detecting intermediate AMD (level 3 in a 4-level grading scheme [49]) as it is the asymptomatic stage after which

vision loss is a serious risk. A 0.96 AUC of the ROC curve is reported in the study. To have a more fine-grained grading of AMD, an ensemble of deep neural networks is used in [61] to detect 12 grades of AMD. These grades include 9 classes based on the AREDS [49] 9-step severity scale, 3 late-stage classes, and 1 additional class for ungradable images. The method achieves a quadratic weighted kappa of 0.92. OCT scans of the eye posterior segment show the anatomical structure of the retinal layers and biomarkers related to AMD are clear in the scans (e.g. drusen) making it possible to detect AMD through OCT. Other approaches use OCT imaging to detect AMD. In [59], transfer learning is used in 3-class classification task (normal, AMD, DME). The authors used the Duke AMD-DME dataset and a pre-trained version of the Google Inception network [69]. The OCT B-scans are pre-processed to have a similar spatial resolution as the pre-trained network input with three channels by replicating the grayscale image. The results show better performance when using the pre-trained network rather than a random initialization of the network. The achieved mean accuracy for each class is 0.99, 0.89, and 0.86 for normal, AMD, and DME respectively. To solve the problem of exhaustively annotating each B-scan separately, Apostolopoulos et al. [20] propose a model that works on the volume level first to infer classification on the B-scan level. The model trained to classify healthy vs. intermediate AMD patients using OCT volumes from the Duke AMD dataset. This approach shows good performance compared to multiple baselines with a 0.997 AUC or ROC curve. For better detection of AMD, other approaches detect the presence of AMD-related abnormalities rather than directly detecting AMD. This could also contribute to the interpretability of such models. This is done in [52] where the authors use a dataset of 109k OCT B-scans to detect the presence of CNV, drusen, and DME. CNV and drusen are indicators of wet and dry AMD respectively and the detection of their presence is correlated with AMD. Transfer learning is used in this case with a pre-trained model on ImageNet. The last layers are optimized for the new task and a separate test set annotated by 6 experts is used for evaluation. The proposed network reaches a 0.99 AUC of ROC curve for the task of urgent referral. Urgent referral is defined in the study as the presence of CNV or DME. Approaches for detecting AMD perform very well in later stages but suffer in the earlier stages where abnormalities are not yet prominent. Moreover, domain shift problems may occur since many studies rely on data from AREDS. Table. IV summarizes AMD related works by stating the main task, the datasets used, and the results of each work. Results are not directly comparable because of utilizing different datasets and splits.

B. Segmentation for Age-related Macular Degeneration

Segmentation of AMD related lesions is done on both fundus images variants and OCT scans. Fundus imaging

TABLE III: Datasets for AMD, retinal layers and fluids segmentation.

Reference	Count	Modality	Task	Notes
Duke AMD OCT [50]	384: 115 no AMD, 269 intermediate AMD	OCT volumes	Classification	100 B-scans per volume (1000×500). Volumes are centered at the foveal pit
Duke AMD-DME [51]	45: 15 normal, 15 AMD, 15 Diabetic Macular Edema	OCT volumes	Classification	31-97 B-scans per volume and 512-1024 A-scans per B-scan
Zhang Lab Data [52]	109k: 51k Normal, 37k CNV, 8k drusen, 11k DME	OCT B-Scans	Classification	OCT B-scans acquired with Spectralis OCT, Heidelberg Engineering, Germany
Duke SD-OCT DME Dataset [53]	110 B-scans from 10 patients	OCT B-scans	Retinal layers and fluid segmentation	Images of subjects with Diabetic Macular Edema. B-scans are centered around the fovea and annotated for retinal layers and fluid regions
OPTIMA MICCAI 2015 Challeng [54]	30: 15 training, 15 testing	OCT volumes	Retinal layers and fluid segmentation	Two manually annotated segmentation maps of retinal cyst areas are provided
RETOUCH [55]	112: 70 training, 42 testing	OCT volumes	Retinal layers and fluid segmentation	3 fluid types segmentation maps for OCT volumes from 3 different devices

is used for segmentation of AMD-related lesion as in [65] where a random forest is utilized to segment Geographic Atrophy (GA) in AMD patients. A hand-crafted feature vector is extracted for each pixel neighborhood and a model is trained to predict if a pixel belongs to a GA. The model is tested on the full dataset and a subset that contains images with low ambiguity and the achieved dice coefficients are 0.68 ± 0.25 and 0.70 ± 0.21 , respectively. OCT is becoming more accessible and OCT cross-sections show information about retinal layers pathologies, OCT is widely used in AMD lesion segmentation. In [9], the authors use U-Net [70] architecture to segment drusen in OCT retinal images. A drusen segmentation dataset is created by utilizing a pre-existing dataset that contains segmentation maps for Bruch's membrane (BM) and the retinal pigment epithelium (RPE) layers. It is also possible to segment GA from OCT volumes without going through a retinal layer segmentation step [67]. A voting strategy is suggested to combine the segmentation from multiple networks into the final segmentation result. The method achieves an intersection over union (IoU) of $86.94\% \pm 8.75\%$. Most of the methods in the literature apply machine learning techniques on a limited dataset where the comparison is between normal and pathological cases. However, pathological cases can interfere with each other and affect the robustness of neural networks on unseen pathologies. This effect is not well studied and investigated. To account for that, DeepMind's work [63] includes 10 different retinal pathologies to be detected. Some of these pathologies are AMD-related such as CNV, GA and drusen. A dataset from Moorfield eye hospital is utilized to create a two-stage deep learning network for detecting 10 OCT conditions and give a referral decision. The first stage model is trained to segment 15 different structures and pathologies in OCT volumes. The segmentation results are then used as input to a classification network. The classification network is trained for 1) referral suggestion (urgent, semi-urgent, routine, observation), 2) diagnosis probability for multiple characteristics (normal, CNV,

macular retinal edema, full macula hole, partial macula hole, central serous retinopathy, Vitreomacular traction, GA), and 3) tissue volume of drusen and the epiretinal membrane (ERM). The reported results are comparable with expert results on these tasks with 0.99 AUC of ROC curve for the Task of referring urgent cases. The intermediate segmentation step enables easier adoption of the model in case it is needed to change OCT vendors as the second model is independent of the OCT source. Detecting lesions for AMD is an important task for detecting early signs and tracking the progression of the disease. However, there is high variability among experts when it comes to pixel-level annotation. To partly overcome this difficulty, approaches that model uncertainty are desired to model this variability.

IV. GLAUCOMA

Glaucoma affects the optic nerve progressively and it is detected commonly in three approaches [71] (1) detection of increased intraocular pressure, (2) identifying the field of abnormal vision, or (3) assessing the damage of the optic nerve by calculating the cup-to-disc ratio (CDR). Optic Disc (OD) is the area where the optic nerve leaves the eye and it can be divided into two parts, the optic cup which is in the center of the Optic Disc as a bright circular area and the peripheral region around the cup which is the neuroretinal rim. The Optic Disc changes visually when the optic nerve fibers are damaged due to glaucoma, this leads to enlargement of the cup region called cupping and it is an indicator for glaucoma suspect detection. Detection of glaucoma and segmentation of areas that characterize glaucoma are presented in the following subsections. Table. V lists the publicly available datasets for glaucoma detection, and optic disk and cup segmentation.

A. Detection of Glaucoma

Multiple measurements are used for detecting glaucoma. In [2], the authors used Visual Fields measurements to differentiate open-angle glaucoma patients

TABLE IV: Summary: Age-related macular degeneration and macular pathologies. AUC is the area under the ROC curve. Acc is the accuracy score. IOU is the intersection over union score.

Reference	Task	Summary	Results	Dataset(s)
Apostolopoulos et al. [20]	Classification	Full OCT volume classification for AMD screening using CNNs and pre training on single B-scans with volume annotation	AUC: 0.99	Duke AMD
Deng et al. [56]	Classification	Distinguish between normal, dry AMD, and wet AMD OCT scans using Gabor filtering with multiple machine learning methods (random forests, SVM, and ANN)	Acc: 0.94	Private 21 patients
Lee et al. [57]	Classification & weak localization	Use OCT scans and VGG network to classify AMD vs. normal patients with occlusion tests to localize salient areas	AUC: 0.93	Private: 48,312 normal, 52,690 AMD
Venhuizen et al. [58]	Classification	Use visual bag of words dictionary and random forest classifier to classify five severity levels of AMD using OCT scans. Levels are No AMD, Early AMD, Intermediate AMD, Advanced AMD with geographic atrophy, and Advanced AMD with choroidal neovascularization	AUC: 0.97	EUGENDA 3265 OCT volumes, Duke AMD-DME
Karri et al. [59]	Classification	Detect AMD and DME with transfer learning and multiple pre-processing steps for saturation, flattening, resizing, and noise reduction	Acc: 0.89	Duke AMD-DME
Burlina et al. [60]	Classification	A pre-trained network is used to extract features around the macula in multiple scales of fundus images. Then, these features are used as input to an SVM	AUC: 0.96	AREDS >130,000 color fundus images
Grassmann et al. [61]	Classification	An ensemble of deep neural networks is used to detect AMD fine-grained stages and image gradability using fundus images	Kappa: 0.92	AREDS >120,000 color fundus images
Kermany et al. [52]	Classification	Detection of CNV and drusen from OCT B-scans as an indicator for AMD. Transfer learning is used for this task.	AUC: 0.99	Zhang Lab Data
Treder et al. [62]	Classification	Use transfer learning to detect wet AMD in OCT B-scans	Acc: 0.96	1,112 OCT B-scans
De Fauw et al. [63]	Classification & segmentation	Detect 10 retinal pathologies and referral scores of OCT volumes based on a segmentation step of 15 retinal structures	AUC: 0.99 urgent referral	Moorfield hospital 14,884 OCT volumes
Schlegl et al. [64]	Segmentation	Segmentation model to quantify intraretinal cystoid fluid (IRC) and subretinal fluid (SRF) in 3 different pathologies	AUC IRC: 0.94, AUC SRF: 0.92	Private 1200 OCT volumes
Feeny et al. [65]	Geographic atrophy segmentation	Random forest model for pixel-wise segmentation of geographic atrophy using handcrafted features extracted from colored fundus images	Dice: 0.68	Private 143 fundus images
Zedeh et al. [9]	Drusen segmentation	Drusen segmentation using U-Net architecture. Ground truth generation from OCT volumes of segmented retinal layers	IOU: 0.82	Private >50k images
Lee et al. [66]	IRF segmentation	Segmentation of intra-retinal fluids in OCT B-scans	Dice: 0.91	Private 1,289 OCT B-scans
Ji et al. [67]	Geographic atrophy segmentation	Voting scheme for combining geographic atrophy segmentation of multiple models from OCT volumes	IOU: 0.86	Private 51 OCT volumes, 54 OCT volumes
Xue et al. [68]	Choroidal neovascularization segmentation	Unsupervised clustering-based segmentation of choroidal neovascularization based on OCT angiography $3 \times 3 \text{ mm}^2$ scans	IOU: 0.87	Private 22 OCTA scans

from healthy patients using deep feedforward neural networks. Visual Fields using Humphrey Field Analyzer from 171 pre-perimetric glaucoma patients and 108 healthy cases are acquired for analysis. The input to the neural network model is the 52 visual field values generated by the measuring machine with Pattern Standard Deviation (PSD) and Mean Deviation (MD) values. The neural network consists of two hidden layers and classifies pre-perimetric glaucoma from normal visual field. The method is compared with several other machine learning models including SVM, Random Forest, and K-Nearest Neighbors. It achieved the highest AUC of 0.926. To go beyond visual field measurements, some works use anatomical features alongside the visual field data.

One work [1] selects the best features from various measurements for 297 glaucoma eyes and 202 non-glaucoma eyes. The measurements included ocular pressure and corneal thickness measurement as general exams, OCT scan for retinal nerve fiber layer (RNFL) thickness, and visual field (VF) exam. The model achieves 0.979 AUC of ROC curve for detecting if a subject has glaucoma or not. The RNFL thickness is the anatomical information most used in the detection process. Other works have attempted to utilize fundus imagery for detecting glaucoma. Li et al. [80] have used fundus images for detecting referable glaucomatous optic neuropathy. The images are annotated by 21 trained ophthalmologists into unlikely suspects of glaucoma and certain glau-

TABLE V: Datasets for glaucoma detection and optic disc/cup segmentation

Reference	Count	Modality	Task	Notes
DRISHTI-GS [72]	50	Fundus	Segmentation	Annotation for optic disc and optic cup
DRIONS-DB [73]	110	Fundus	Segmentation	Annotation for optic disc
RIM-ONE [74]	159	Fundus	Segmentation	Annotation for optic disc
SiMES [75]	230	Fundus	Segmentation	Optic disc segmentation for subjects with parapapillary atrophy
ORIGA [76]	650: 482 No glaucoma, 168 with glaucoma	Fundus	Classification	Binary classification
SCES [77]	1676: 1630 No glaucoma, 46 with glaucoma	Fundus	Classification	Binary classification
MICCAI 2018 REFUGE [78]	1200: 1080 No glaucoma, 120 with glaucoma	Fundus	Classification and segmentation	Binary classification, optic disc and cup segmentation and fovea localization

TABLE VI: Summary of works on glaucoma detection and glaucoma-related anatomies segmentation. AUC is the area under the ROC curve. IOU is intersection over union.

Reference	Task	Summary	Results	Dataset(s)
Cheng et al. [79]	Classification	CNN with overlapping pooling layers for binary classification of glaucoma in fundus images	AUC: 0.83, 0.88	ORIGA, SCES
kim et al. [1]	Classification	Use multiple eye measurements as input for decision trees, random forests, SVMs, and KNNs for comparison	AUC: 0.97	Private 499 Cases
Asaoka et al. [2]	Classification	Use Visual Field measurements for preperimetric glaucoma detection using feedforward neural networks	AUC: 0.92	Private 279 Visual Fields
Li et al. [80]	Classification	Glaucoma suspect and glaucoma certain classification based on fundus imaging	AUC: 0.98	Private 48,116 fundus images
Zilly et al. [10]	Disk & cup segmentation	Entropy based sampling for informative points selection with convolutional filters optimization using greedy boosting	IOU: disk 0.89, cup 0.80 on RIM-One	DRISHTI-GS, RIM-ONE, Messidor
Maninis et al. [81]	Disk & vessels segmentation	Fully convolutional approach with pre-trained VGG network and supervision on multiple internal layers	0.95 precision @ 0.95 recall	DRIONS-DB, RIM-ONE
Tan et al. [82]	Disc, fovea, vessels segmentation	7-layer CNN to classify central pixel of 33×33 patches into disc, fovea, vessel, or background	IOU: 0.62	DRIVE
Srivastava et al. [83]	Disc segmentation	Pixel-wise classification of OD with distance and intensity handcrafted features and a 7-layer neural network to test optic disc segmentation robustness against parapapillary atrophy	IOU: 0.90	SiMES

coma depending on the vertical CDR, rim width, RNFL defect, and disc hemorrhage. The deep model registers a 0.98 AUC of ROC for referable glaucomatous optic neuropathy. Detection of glaucoma is a challenging task because the early signs of the disease could be either functional or anatomical. Approaches that aim for the early detection of glaucoma should consider both factors. Table. VI summarizes glaucoma-related approaches and states the main task, the used datasets, and the results of each approach. Results are not directly comparable because of different datasets and splits utilization.

B. Segmentation for Glaucoma

The authors in [10] used an entropy-based sampling scheme to select points of interest for classification. The idea behind entropy based sampling is to pick the most informative patches for the training process. The informative points are selected to have the highest entropy values where the entropy of each pixel is calculated from its $N \times N$ neighborhood. A convex hull transform is applied to the initial segmentation assuming that the disk and cup shapes are convex. The method is compared with state-of-the-art methods and achieves in-

tersection over union (IOU) values for optic disc and cup segmentation of 91.4, 85.0 on DRISHTI-GS dataset and 89.0, 80.2 on the RIM-ONE dataset. To have more global information about the image, the authors in [81] have used the full fundus image rather than image patches to segment retinal vessels and optic disk simultaneously. The authors used a pre-trained VGG network [84] and inspired by GoogLeNet inception network [69], supervision to multiple internal layers is added. The authors further added two sets of specialized layers connected to the internal layers of the network. One set is used to do full image segmentation and uses the first four blocks of VGG to have higher spatial resolution. The second set of added layers is used for optic disk segmentation by connecting to the last four layers of VGG. This approach is fully convolutional without any fully connected layers and a probability map of the same size as the original image is produced for each task. For optic disk segmentation, the evaluation is done on DRIONS-DB and RIM-ONE datasets and the model achieves dice coefficient values of 0.971 and 0.959 for optic disk segmentation on the two datasets respectively. While the segmentation of glaucoma related anatomies (such as optic disc/cup)

is achievable with good performance. It is important to note that it is clinically challenging to decide for glaucoma based on data acquired at one instance. It would be more beneficial to utilize such quantifiable segmentation systems for the tracking of glaucoma suspects to monitor any glaucoma progression symptoms.

V. RETINAL VESSELS SEGMENTATION

Retinal vessel segmentation is widely researched in the literature. Retinal vessel networks contain rich information about various retinal pathologies. Early signs of diabetic retinopathy and vascular burden from hypertension could be detected. Moreover, vision threatening diseases such as Retinal Vein Occlusion (RVO) and Retinal Artery Occlusion (RAO) are detected from abnormalities in microvascular structure. The exclusion of retinal vessels network is a preprocessing step used in various automatic detection algorithms. Therefore, reliable methods for automatic retinal vessel segmentation are needed. Retinal vessel segmentation is a challenging task for multiple reasons. Retinal vessels vary in thickness and could reach 1-pixel thickness (depending on resolution and anatomical characteristics) and vessels could have intersections and sometimes thicker vessels have centerline reflex effect where it appears as two vessels. Table. VII lists the publicly available datasets for retinal vessels segmentation.

Using Machine Learning to label central pixel of small extracted patches is one approach frequently used in the literature [88], [92], [89], [90]. Liskowski et al. [88] use patches of size 27×27 and applied a preprocessing for global contrast normalization on image level and zero-phase component analysis whitening on patch level. A deep convolutional network is then trained to classify the center pixel of the pre-processed patches. As for end-to-end direct image segmentation, holistic view of the image structure is considered which allows for a more context-aware segmentation. Moreover, a faster segmentation is possible since there is no need for a sliding window patch-wise classification. In [13] the vessel segmentation is done on image level using a convolutional neural network with sided output and a conditional random field. The sided output is used as a classifier for shallow layers outputs. Each sided output has its loss that is weighted and added to the final loss as in [95]. A Conditional Random Field is used to capture non-local correlations between pixels and takes the image and the sided outputs from the convolution layers as input. The final loss of the network is the weighted sum of the sided outputs and the output of the Conditional Random Field. The time needed for full segmentation of one image of size 565×584 is 1.3 seconds. The emergence of generative adversarial neural networks is also affecting research in ophthalmic image analysis. In order to counter the problems of data scarcity and high cost of annotations, Costa et al. [94] proposed a framework to synthesize fundus images with their corresponding vessel map segmentation. Two segmentation

networks are trained on the real and generated images. The performance of the two networks is close which indicates good generation. However, the generated images are of relatively low resolution (256×256). Hence, the segmentation results of both models are much lower than state-of-the-art. Machine learning is also used to segment microvasculature in OCTA images [91]. This work includes 80 OCTA foveal region images of 12 eyes from 6 volunteers with no sign of pathologies. The network is composed of 3 convolutional layers and 3 max-pooling layers. The input is 61×61 pixels patch of the image and the network is trained to classify the central pixel of each patch. The number of sampled patches with and without vessels is equal to avoid bias towards the over-represented class. The method achieves an accuracy of 0.83 and a dice score of 0.7 and it is compared with the human performance by asking a second annotator to segment 10 of the images. Full segmentation of OCT-A image microvasculature takes 2 minutes using the proposed method. While various approaches perform well on the retinal vessels segmentation task, it is still challenging to detect smaller vessels with low contrast and to detect vessels in cases of severe pathologies. Using OCTA is a good option for monitoring microvascular abnormalities in limited areas. Table. VIII summarizes retinal vessels segmentation works by stating the main task and the datasets used. The results of each work on retinal vessels segmentation are shown in Table. IX.

VI. RETINAL LAYERS SEGMENTATION IN OCT IMAGES

Retinal layers segmentation is valuable for analyzing OCT scans to diagnose or monitor retinal diseases. This task can be challenging due to pathologies such as cysts, subretinal fluids or drusen where the structure of the retinal layers varies significantly.

In [97], the authors use a generalized form of U-Net [70] to segment retinal layers in OCT images with AMD. AMD introduces difficulties in capturing the structure of the retina. The generalized U-Net used in this work is a deeper network than original U-Net with more downsampling layers and more feature maps per layer. The receptive field becomes 572×572 pixels rather than 140×140 pixels in the original U-Net. This is important for the network to be able to capture larger structures in the image such as fluid-filled spaces. The model achieves a central macular thickness estimation error of $14.0 \pm 22.1 \mu m$. It takes the system 5 seconds to segment a volume of 38 B-scans. In [98], authors use an encoder-decoder architecture to segment retinal layers and fluid into 10 classes (7 anatomical layers, 1 fluid, 2 above and under retina classes). The loss function uses the weighted logistic regression loss where the weights are used to balance imbalanced classes and to give higher importance to segmentation of boundary pixels. Moreover, a differentiable approximation of dice loss is added to the loss function making the total loss as a weighted sum of the two losses. The network employs convolutional kernels

TABLE VII: Publicly available datasets for retinal vessels segmentation

Reference	Count	Modality	Notes
DRIVE [85]	40: 7 mild DR signs	Fundus	Manual segmentation of the blood vessels is provided
STARE [86]	20: 10 cases with pathologies that overlap with blood vessels	Fundus	Two manually annotated segmentation maps are provided
CHASE_DB [87]	28 images	Fundus	Children Fundus images. The images have worse contrast for blood vessels, nonuniform background illumination and wide vessels that have the central vessel reflex

TABLE VIII: Summary of retinal vessels segmentation works.

Reference	Summary	Dataset(s)
Liskowski et al. [88]	CNNs with and without max-pooling applied for central pixel classification of 27×27 patches	DRIVE, STARE, CHASE_DB
Wang et al. [89]	Ensemble learning with random forests trained on feature maps extracted with a CNN (25×25 patches)	DRIVE, STARE
Maji et al. [90]	Ensemble of multiple CNNs trained on binary classification of central pixel for 31×31 patches	DRIVE
Li et al. [3]	Use artificial neural networks and denoising autoencoders to learn full patch segmentation map	DRIVE, STARE
Fu et al. [13]	A CNN with sided output and conditional random field to capture non-local correlations	DRIVE, STARE, CHASE_DB
Prentavsic et al. [91]	OCT angiography microvasculature segmentation using a 3-layer CNN to classify center pixel of 61×61 patches	Private 80 OCTA images
Leopold et al. [92]	Gabor filter preprocessing to train a CNN with batch normalization for central pixel classification of 65×65 patches	DRIVE
Son et al. [93]	Retinal vessels segmentation using generative adversarial networks where the generator has an architecture similar to U-Net and the discriminator classifies human and non-human segmentation	DRIVE, STARE
Costa et al. [94]	Automatic generation of fundus images and vessels segmentation maps pairs to account for data scarcity problems using generative adversarial neural networks	DRIVE, Messidor

TABLE IX: Retinal vessels segmentation results in terms of sensitivity (Sen), specificity (Spe), accuracy (Acc), and area under the curve of ROC (AUC).

Reference	DRIVE dataset				STARE dataset				CHASE-DB dataset			
	Sen	Spc	Acc	AUC	Sen	Spc	Acc	AUC	Sen	Spc	Acc	AUC
Li et al. [3]	75.69	98.16	95.27	97.38	77.26	98.44	96.28	98.79	75.07	97.93	95.81	97.16
Fu et al. [13]	76.03	-	95.23	-	74.12	-	95.85	-	71.30	-	94.89	-
Liskowski et al. [88]	77.63	97.68	94.95	97.20	79.67	97.54	95.66	97.85	-	-	-	-
Wang et al. [89]	81.73	97.33	97.67	94.75	81.04	97.91	98.13	97.51	-	-	-	-
Leopold et al. [92]	78.00	97.27	94.78	-	-	-	-	-	-	-	-	-
Maji et al. [90]	-	-	94.7	92.83	-	-	-	-	-	-	-	-
Son et al. [93]	-	-	-	98.03	-	-	-	98.38	-	-	-	-

of size 7×3 to keep consistency with OCT spatial resolution, this kernel size ensures the receptive field of the network to cover the whole retinal layers. The images are sliced with none overlapping width-wise stripes. This allows training a larger batch size with GPU limitations. At test time the whole B-Scan is given to the network as input and it takes 10 milliseconds to segment one B-scan. The achieved dice score for fluid segmentation which is the most challenging part in the segmentation task is 0.77. Segmenting retinal layers is a challenging task when retinal fluids and abnormalities occur. For such cases, it could be beneficial to use certainty aware approaches to model the inter-rater variability. Table. III lists the publicly available datasets for retinal layers and fluids segmentation. Table. X summarizes retinal

layers segmentation works by stating the main task, the datasets used, and the results of each work. Results are not directly comparable because of different datasets and splits utilization.

VII. LIMITATIONS AND FUTURE

Machine learning techniques are being widely researched for their good performance on challenging tasks of fine-grained grading and pixel-level segmentation. However, these techniques are hindered by various limitations. Some important limitations are the availability of data and the regulations for data usage. Unlike most computer vision applications, medical applications require datasets annotated by medical experts. This limits the amount of available annotated data because of

TABLE X: Summary of retinal layers segmentation works. Acc is the accuracy score. Average dice is the average of dice of multiple structures reported in the paper.

Reference	Summary	Results	Dataset(s)
Schlegl et al. [96]	Cystoid structures weak segmentation in OCT images using semantic descriptors extracted from clinical reports without pixel-wise labels for training	Acc: 0.81	Private 157 OCT Volumes
Venhuizen et al. [97]	Retinal layers segmentation by using a generalized form of U-Net to have a larger receptive field and deeper architecture	Dice: 0.95	EUGENDA, Duke_DME
Roy et al. [98]	Retinal layers segmentation by using U-Net like structure with batch normalization to segment retinal layers and fluid using a tailored loss function with a differentiable approximation of dice loss	Average dice: 0.90	Duke_DME
Apostolopoulos et al. [99]	Retinal layers segmentation by using dilated convolution for bigger receptive field and residual connections with MSE loss on U-Net like architecture	Dice: 0.99	Private 20 OCT volumes
Gopinath et al. [100]	Cystoid structures segmentation in OCT images using selective enhancement with generalized motion pattern on 3D OCT data	Dice: 0.71	OPTIMA, Duke_DME
Schlegl et al. [101]	Using adversarial neural networks to mark biomarkers candidates such as retinal fluids in OCT B-scans without the need for annotated data	Dice: 0.79	Private 280 healthy volumes + 10 volumes with fluids
venhuizen et al. [102]	Deep learning for segmentation and quantification of intraretinal cystoid fluid in OCT volumes on B-scan level	Dice: 0.75	EUGENDA 221 OCT volumes

the high cost of annotation. One way to resolve this in the literature is data augmentation which could be done more realistically with generative models [103] that learn the data distribution and possibly generate unseen examples [94]. Another way is utilizing semi-supervised learning techniques [104] which allow the incorporation of unlabeled data alongside labeled data into the model to enhance the overall performance on the labeled data. This is beneficial for medical applications where a large amount of unlabeled data is not used in developing the models. Another data-related limitation is the country-specific regulations on data collection and patient consent which mostly results in homogeneous datasets in terms of ethnicity and/or hardware settings for development. Models developed on such datasets are prone to generalization problems as one population data might have different characteristics that introduce a bias in the model. Hence, domain adaptation techniques should be kept in mind while developing these models. One solution to the problem is to decompose it into multiple parts as in first segment then classify [63] to allow easier domain adaptation when needed. An alternative solution is to separate shared content information from attribute (domain-specific) information using image-to-image translation via disentangled representations [105]. Although deep models perform well on quantitative measures, trust is still an issue for the adoption of these models into clinical scenarios. A deep model is seen as a black box where results are not interpretable and failure cases cannot be anticipated nor explained. Approaches such as [35], [14] visualize heat maps associated with the classification to show attention areas of the model that lead the classification. Another possibility to explore is disentangling image generative factors in image representation latent space. Disentangling allows a better understanding of the image representation in high dimensional space where each component of representation holds information about one trait in the data

[106], [107].

VIII. SUMMARY

Machine learning and deep convolutional neural networks have been the focus of most recent publications on ophthalmic data analysis outperforming legacy algorithms on different applications from detection and grading diseases such as AMD, DR, and glaucoma to the segmentation of anatomical structures such as vessels and retinal layers or intraretinal fluids. In this paper, we reviewed the recently published machine learning methods in Ophthalmology and listed the datasets publicly available for research. Classification and segmentation are the two most addressed problems in these publications. Detection and grading of eye diseases through classification leads to more efficient screening programs enabling the healthcare organizations to cover a larger population. Segmentation of vessels and retinal layer or lesions such as intra-retinal fluid, on the other hand, can be used for improving the performance in detecting diseases or as a crucial intermediate process of the clinical decision support systems in the future. Together with the rise of these methods came the concerns about patient privacy and the rights on the use of data. Interpretability is one requirement in medical applications as more than sensitivity and specificity is needed to assess the performance of machine learning methods. There is no doubt that machine learning came a long way fast and future researches and developments will make it one of the most powerful tools in the hands of experts to diagnose and treat eye diseases.

REFERENCES

- [1] S. J. Kim, K. J. Cho, and S. Oh, "Development of machine learning models for diagnosis of glaucoma," *PLoS ONE*, vol. 12, no. 5, p. e0177726, 2017.

- [2] R. Asaoka, H. Murata, A. Iwase, and M. Araie, "Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier," *Ophthalmology*, vol. 123, no. 9, pp. 1974–1980, 2016.
- [3] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 109–118, 2015.
- [4] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [5] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
- [6] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.
- [7] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 533–540.
- [8] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1273–1284, 2016.
- [9] S. G. Zadeh, M. W. Wintergerst, V. Wiens, S. Thiele, F. G. Holz, R. P. Finger, and T. Schultz, "CNNs enable accurate and fast segmentation of drusen in optical coherence tomography," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 65–73.
- [10] J. Zilly, J. M. Buhmann, and D. Mahapatra, "Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation," *Computerized Medical Imaging and Graphics*, vol. 55, pp. 28–41, 2017.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [12] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.
- [13] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu, "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2016, pp. 132–139.
- [14] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Medical Image Analysis*, vol. 39, pp. 178–193, 2017.
- [15] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, "An ensemble deep learning based approach for red lesion detection in fundus images," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 115–127, 2018.
- [16] D. S. Ting, L. Peng, A. V. Varadarajan, P. A. Keane, P. Burlina, M. F. Chiang, L. Schmetterer, L. R. Pasquale, N. M. Bressler, D. R. Webster *et al.*, "Deep learning in ophthalmology: the technical and clinical considerations," *Progress in Retinal and Eye Research*, 2019.
- [17] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.
- [18] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Progress in Retinal and Eye Research*, vol. 67, pp. 1–29, 2018.
- [19] K. B. Nielsen, M. L. Lautrup, J. K. Andersen, T. R. Savarimuthu, and J. Grauslund, "Deep learning-based algorithms in screening of diabetic retinopathy: A systematic review of diagnostic performance," *Ophthalmology Retina*, vol. 3, no. 4, pp. 294–304, 2019.
- [20] S. Apostolopoulos, C. Ciller, S. De Zanet, S. Wolf, and R. Sznitman, "Retinet: Automatic amd identification in OCT volumetric data," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 8, pp. 387–387, 2017.
- [21] R. Lee, T. Y. Wong, and C. Sabanayagam, "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss," *Eye and vision*, vol. 2, no. 1, p. 17, 2015.
- [22] T. Kauppi *et al.*, *Eye fundus image analysis for automatic detection of diabetic retinopathy*. Lappeenranta University of Technology, 2010.
- [23] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "Diaretddb0: Evaluation database and methodology for diabetic retinopathy algorithms," *Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, Finland*, vol. 73, 2006.
- [24] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietilä, "Diaretddb1 diabetic retinopathy database and evaluation protocol," in *Medical Image Understanding and Analysis, Aberystwyth, Wales, UK, 2007*, pp. 61–65.
- [25] M. Niemeijer, B. Van Ginneken, M. J. Cree, A. Mizutani, G. Quélélec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu *et al.*, "Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 185–195, 2010.
- [26] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno *et al.*, "Teleophtha: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, 2013.
- [27] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Chardon, and J.-C. Klein, "Feedback on a publicly distributed database: the messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014. [Online]. Available: <http://www.ias-iss.org/ojs/IAS/article/view/1155>
- [28] R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing bag-of-visual-words representations for lesion classification in retinal images," *PLoS ONE*, vol. 9, no. 6, p. e96814, 2014.
- [29] Kaggle, "Kaggle diabetic retinopathy detection challenge," <https://www.kaggle.com/c/diabetic-retinopathy-detection>, 2015, accessed: 26.09.2018.
- [30] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabudhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid)," 2018. [Online]. Available: <http://dx.doi.org/10.21227/H25W98>
- [31] M. D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, 2016.
- [32] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-inet: Deep mining lesions for diabetic retinopathy detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 267–275.
- [33] Z. Lin, R. Guo, Y. Wang, B. Wu, T. Chen, W. Wang, D. Z. Chen, and J. Wu, "A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2018, pp. 74–82.
- [34] M. H. Sarhan, P. Sha, M. Chen, M. K. Durbin, M. Yigitsoy, and A. Eslami, "Deep learning for automatic diabetic retinopathy grading of ultra-widefield fundus images," *Investigative Ophthalmology & Visual Science*, vol. 60, no. 9, pp. 1094–1094, 2019.
- [35] R. Gargya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, 2017.
- [36] S. Abbasi-Sureshjani, B. Dashtbozorg, B. M. ter Haar Romeny, and F. Fleuret, "Boosted exudate segmentation in retinal images using residual nets," in *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 210–218.
- [37] H. Huang, H. Ma, H. J. van Triest, Y. Wei, and W. Qian, "Automatic detection of neovascularization in retinal images using extreme learning machine," *Neurocomputing*, 2017.
- [38] A. Levenkova, A. Sowmya, M. Kalloniatis, A. Ly, and A. Ho, "Lesion detection in ultra-wide field retinal images for diabetic

- retinopathy diagnosis," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. International Society for Optics and Photonics, 2018, p. 1057531.
- [39] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Investigative Ophthalmology & Visual Science*, vol. 59, no. 1, pp. 590–596, 2018.
- [40] Y. Guo, A. Camino, J. Wang, D. Huang, T. S. Hwang, and Y. Jia, "Mednet, a neural network for automated detection of avascular area in OCT angiography," *Biomedical Optics Express*, vol. 9, no. 11, pp. 5147–5158, 2018.
- [41] M. H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, and A. Es-lami, "Multi-scale microaneurysms segmentation using embedding triplet loss," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2019, pp. 174–182.
- [42] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [43] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2069–2073.
- [44] "American Academy of Ophthalmology. International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table," <http://www.icoph.org/downloads/Diabetic-Retinopathy-Detail.pdf>, accessed: 11.06.2019.
- [45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [46] P. S. Silva, J. D. Cavallerano, J. K. Sun, J. Noble, L. M. Aiello, and L. P. Aiello, "Nonmydriatic ultrawide field retinal imaging compared with dilated standard 7-field 35-mm photography and retinal specialist examination for evaluation of diabetic retinopathy," *American Journal of Ophthalmology*, vol. 154, no. 3, pp. 549–559, 2012.
- [47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [48] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.
- [49] F. L. Ferris, M. D. Davis, T. E. Clemons, L. Y. Lee, E. Y. Chew, A. S. Lindblad, R. C. Milton, S. B. Bressler, and R. Klein, "A simplified severity scale for age-related macular degeneration: Areds report no. 18." *Archives of ophthalmology (Chicago, Ill.: 1960)*, vol. 123, no. 11, pp. 1570–1574, 2005.
- [50] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.
- [51] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomedical Optics Express*, vol. 5, no. 10, pp. 3568–3577, 2014.
- [52] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [53] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomedical Optics Express*, vol. 6, no. 4, pp. 1172–1194, 2015.
- [54] OPTIMA, "Proceeding of the miccai 2015 optima cysts segmentation challenge," <https://optima.meduniwien.ac.at/research/challenges/>, 2015, accessed: 22.04.2020.
- [55] H. Bogunović, F. Venhuizen, S. Klmscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller *et al.*, "Retouch-the retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Transactions on Medical Imaging*, 2019.
- [56] J. Deng, X. Xie, L. Terry, A. Wood, N. White, T. H. Margrain, and R. V. North, "Age-related macular degeneration detection and stage classification using choroidal OCT images," in *International Conference on Image Analysis and Recognition*. Springer, 2016, pp. 707–715.
- [57] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration optical coherence tomography images," *Ophthalmology Retina*, 2017.
- [58] F. G. Venhuizen, B. van Ginneken, F. van Asten, M. J. van Grinsven, S. Fauser, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Automated staging of age-related macular degeneration using optical coherence tomography," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 4, pp. 2318–2328, 2017.
- [59] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical Optics Express*, vol. 8, no. 2, pp. 579–592, 2017.
- [60] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmology*, vol. 135, no. 11, pp. 1170–1176, 2017.
- [61] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H. Weber, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.
- [62] M. Treder, J. L. Lauermaun, and N. Eter, "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," *Graefes Archive for Clinical and Experimental Ophthalmology*, vol. 256, no. 2, pp. 259–265, Feb 2018.
- [63] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, p. 1342, 2018.
- [64] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, 2018.
- [65] A. K. Feeny, M. Tadarati, D. E. Freund, N. M. Bressler, and P. Burlina, "Automated segmentation of geographic atrophy of the retinal epithelium via random forests in areds color fundus images," *Computers in Biology and Medicine*, vol. 65, pp. 124–136, 2015.
- [66] C. S. Lee, A. J. Tying, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomedical Optics Express*, vol. 8, no. 7, pp. 3440–3448, 2017.
- [67] Z. Ji, Q. Chen, S. Niu, T. Leng, and D. L. Rubin, "Beyond retinal layers: a deep voting model for automated geographic atrophy segmentation in SD-OCT images," *Translational Vision Science & Technology*, vol. 7, no. 1, pp. 1–1, 2018.
- [68] J. Xue, A. Camino, S. T. Bailey, X. Liu, D. Li, and Y. Jia, "Automatic quantification of choroidal neovascularization lesion area on OCT angiography based on density cell-like p systems with active membranes," *Biomedical Optics Express*, vol. 9, no. 7, pp. 3208–3219, 2018.
- [69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [70] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.
- [71] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, and T. Y. Wong, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

- [72] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, "Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation," in *IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 53–56.
- [73] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-de-la Casa, "Identification of the optic nerve head with genetic algorithms," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008.
- [74] F. Fumero, S. Alayón, J. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2011, pp. 1–6.
- [75] A. W. Foong, S.-M. Saw, J.-L. Loo, S. Shen, S.-C. Loon, M. Rosman, T. Aung, D. T. Tan, E. S. Tai, and T. Y. Wong, "Rationale and methodology for a population-based study of eye diseases in malay people: The singapore malay eye study (simes)," *Ophthalmic Epidemiology*, vol. 14, no. 1, pp. 25–35, 2007.
- [76] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2010, pp. 3065–3068.
- [77] C. C. Sng, L.-L. Foo, C.-Y. Cheng, J. C. Allen, M. He, G. Krishnaswamy, M. E. Nongpiur, D. S. Friedman, T. Y. Wong, and T. Aung, "Determinants of anterior chamber depth: the singapore chinese eye study," *Ophthalmology*, vol. 119, no. 6, pp. 1143–1150, 2012.
- [78] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, 2020.
- [79] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 715–718.
- [80] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.
- [81] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2016, pp. 140–148.
- [82] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," *Journal of Computational Science*, vol. 20, pp. 70–79, 2017.
- [83] R. Srivastava, J. Cheng, D. W. Wong, and J. Liu, "Using deep learning for robustness to parapapillary atrophy in optic disc segmentation," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2015, pp. 768–771.
- [84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [85] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.
- [86] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.
- [87] C. G. Owen, A. R. Rudnicka, R. Mullen, S. A. Barman, D. Monckos, P. H. Whincup, J. Ng, and C. Paterson, "Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program," *Investigative Ophthalmology & Visual Science*, vol. 50, no. 5, pp. 2004–2010, 2009.
- [88] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.
- [89] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, 2015.
- [90] D. Maji, A. Santara, P. Mitra, and D. Sheet, "Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images," *arXiv preprint arXiv:1603.04833*, 2016.
- [91] P. Prentašić, M. Heisler, Z. Mammo, S. Lee, A. Merkur, E. Navajas, M. F. Beg, M. Šarunić, and S. Lončarić, "Segmentation of the foveal microvasculature using deep learning networks," *Journal of Biomedical Optics*, vol. 21, no. 7, pp. 075 008–075 008, 2016.
- [92] H. A. Leopold, J. Orchard, J. Zelek, and V. Lakshminarayanan, "Use of gabor filters and deep networks in the segmentation of retinal vessel morphology," in *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XV*, vol. 10068. International Society for Optics and Photonics, 2017, p. 100680R.
- [93] J. Son, S. J. Park, and K.-H. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," *arXiv preprint arXiv:1706.09318*, 2017.
- [94] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho, "End-to-end adversarial retinal image synthesis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2017.
- [95] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.
- [96] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs, "Predicting semantic descriptions from medical images with convolutional neural networks," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 437–448.
- [97] F. G. Venhuizen, B. van Ginneken, B. Liefers, M. J. van Grinsven, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks," *Biomedical Optics Express*, vol. 8, no. 7, pp. 3292–3316, 2017.
- [98] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, no. 8, pp. 3627–3642, 2017.
- [99] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, "Pathological OCT retinal layer segmentation using branch residual u-shape networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2017, pp. 294–301.
- [100] K. Gopinath and J. Sivaswamy, "Segmentation of retinal cysts from optical coherence tomography volumes via selective enhancement," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 273–282, 2018.
- [101] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019.
- [102] F. G. Venhuizen, B. van Ginneken, B. Liefers, F. van Asten, V. Schreur, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography," *Biomedical Optics Express*, vol. 9, no. 4, pp. 1545–1569, 2018.
- [103] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [104] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.
- [105] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–51.
- [106] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." *ICLR*, vol. 2, no. 5, p. 6, 2017.
- [107] M. H. Sarhan, A. Eslami, N. Navab, and S. Albarqouni, "Learning interpretable disentangled representations using adversarial vae's," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Cham: Springer International Publishing, 2019, pp. 37–44.

Creative Commons Attribution 4.0 International

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give [appropriate credit](#) , provide a link to the license, and [indicate if changes were made](#) . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#) .

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

Bibliography

- [1] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk. “Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices”. In: *NPJ digital medicine* 1.1 (2018), pp. 1–8 (cit. on pp. 32, 34).
- [2] A. Achille and S. Soatto. “Emergence of invariance and disentanglement in deep representations”. In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 1947–1980 (cit. on p. 13).
- [3] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. “Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1313–1321 (cit. on p. 35).
- [4] American Academy of Ophthalmology. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table*. <http://www.icoph.org/downloads/Diabetic-Retinopathy-Detail.pdf>. Accessed: June 22 2021 (cit. on p. 29).
- [5] M. A. Arcones and E. Gine. “On the bootstrap of U and V statistics”. In: *The Annals of Statistics* (1992), pp. 655–674 (cit. on p. 13).
- [6] C. Baur, S. Albarqouni, and N. Navab. “Semi-supervised Deep Learning for Fully Convolutional Networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 311–319 (cit. on p. 8).
- [7] Y. Bengio, A. Courville, and P. Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828 (cit. on pp. 3, 9).
- [8] Y. Bengio and O. Delalleau. “On the expressive power of deep architectures”. In: *International conference on algorithmic learning theory*. Springer. 2011, pp. 18–36 (cit. on pp. 2, 3).
- [9] Y. Bengio, O. Delalleau, and N. Le Roux. “The curse of highly variable functions for local kernel machines”. In: *Advances in neural information processing systems* 18 (2006), p. 107 (cit. on p. 3).
- [10] A. Chatsias, T. Joyce, G. Papanastasiou, et al. “Disentangled representation learning in cardiac image analysis”. In: *Medical image analysis* 58 (2019), p. 101535 (cit. on pp. 4, 18).
- [11] I. Y. Chen, P. Szolovits, and M. Ghassemi. “Can AI help reduce disparities in general medical and mental health care?” In: *AMA journal of ethics* 21.2 (2019), pp. 167–179 (cit. on p. 24).
- [12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818 (cit. on p. 7).
- [13] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud. “Isolating sources of disentanglement in variational autoencoders”. In: *arXiv preprint arXiv:1802.04942* (2018) (cit. on pp. 4, 9, 10, 12–14, 26).

- [14] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., 2016, 2180–2188 (cit. on p. 13).
- [15] S. Chopra, R. Hadsell, and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 539–546 (cit. on p. 8).
- [16] P. Costa, A. Galdran, M. I. Meyer, et al. “End-to-end adversarial retinal image synthesis”. In: *IEEE Transactions on Medical Imaging* 37.3 (2017), pp. 781–791 (cit. on p. 31).
- [17] E. Creager, D. Madras, J.-H. Jacobsen, et al. “Flexibly fair representation learning by disentanglement”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1436–1445 (cit. on pp. 9, 17).
- [18] B. Dashtbozorg, J. Zhang, F. Huang, and B. M. ter Haar Romeny. “Retinal microaneurysms detection using local convergence index features”. In: *IEEE Transactions on Image Processing* 27.7 (2018), pp. 3300–3315 (cit. on p. 33).
- [19] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: *Nature Medicine* 24.9 (2018), p. 1342 (cit. on p. 31).
- [20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, pp. 248–255 (cit. on p. 7).
- [21] A. Di Martino, C.-G. Yan, Q. Li, et al. “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism”. In: *Molecular psychiatry* 19.6 (2014), pp. 659–667 (cit. on p. 24).
- [22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226 (cit. on p. 5).
- [23] C. Eastwood and C. K. Williams. “A framework for the quantitative evaluation of disentangled representations”. In: *International Conference on Learning Representations*. 2018 (cit. on p. 14).
- [24] R. Gargeya and T. Leng. “Automated identification of diabetic retinopathy using deep learning”. In: *Ophthalmology* 124.7 (2017), pp. 962–969 (cit. on p. 30).
- [25] L. Gatys, A. Ecker, and M. Bethge. “A Neural Algorithm of Artistic Style”. In: *Journal of Vision* 16.12 (2016), pp. 326–326 (cit. on p. 27).
- [26] K. A. Goatman, M. J. Cree, J. A. Olson, J. V. Forrester, and P. F. Sharp. “Automated measurement of microaneurysm turnover”. In: *Investigative ophthalmology & visual science* 44.12 (2003), pp. 5335–5341 (cit. on p. 37).
- [27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016 (cit. on p. 3).
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014 (cit. on pp. 15, 27).
- [29] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. “The case for process fairness in learning: Feature selection for fair decision making”. In: *NIPS Symposium on Machine Learning and the Law*. Vol. 1. 2016, p. 2 (cit. on p. 5).
- [30] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez. “Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1273–1284 (cit. on p. 36).

- [31] V. Gulshan, L. Peng, M. Coram, et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *Jama* 316.22 (2016), pp. 2402–2410 (cit. on pp. 31, 32).
- [32] M. Haloi. “Improved microaneurysm detection using deep neural networks”. In: *arXiv preprint arXiv:1505.04424* (2015) (cit. on p. 33).
- [33] M. Hardt, E. Price, and N. Srebro. “Equality of opportunity in supervised learning”. In: *arXiv preprint arXiv:1610.02413* (2016) (cit. on p. 5).
- [34] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 3).
- [35] I. Higgins, L. Matthey, A. Pal, et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: (2016) (cit. on pp. 4, 9, 10, 12, 13, 23).
- [36] M. D. Hoffman and M. J. Johnson. “Elbo surgery: yet another way to carve up the variational evidence lower bound”. In: *Workshop in Advances in Approximate Bayesian Inference, NIPS*. Vol. 1. 2016, p. 2 (cit. on p. 12).
- [37] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. “Multimodal unsupervised image-to-image translation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189 (cit. on pp. 4, 10).
- [38] M. Jaderberg, V. Mnih, W. M. Czarnecki, et al. “Reinforcement learning with unsupervised auxiliary tasks”. In: *arXiv preprint arXiv:1611.05397* (2016) (cit. on p. 3).
- [39] F. Kamiran and T. Calders. “Classifying without discriminating”. In: *2nd International Conference on Computer, Control and Communication*. IEEE. 2009, pp. 1–6 (cit. on p. 14).
- [40] M. Kayhan, O. Köpüklü, M. H. Sarhan, M. Yigitsoy, A. Eslami, and G. Rigoll. “Deep Attention Based Semi-supervised 2D-Pose Estimation for Surgical Instruments”. In: *International Conference on Pattern Recognition. ICPR Workshops and Challenges*. Vol. 12661. Lecture Notes in Computer Science. Springer, 2020, pp. 444–460 (cit. on p. 49).
- [41] H. Kim and A. Mnih. “Disentangling by factorising”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658 (cit. on pp. 4, 9, 10, 12–14).
- [42] D. P. Kingma and M. Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on pp. 9, 11).
- [43] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105 (cit. on p. 3).
- [44] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. “Building machines that learn and think like people”. In: *Behavioral and brain sciences* 40 (2017) (cit. on p. 9).
- [45] C. Lam, C. Yu, L. Huang, and D. Rubin. “Retinal lesion detection with deep learning using image patches”. In: *Investigative ophthalmology & visual science* 59.1 (2018), pp. 590–596 (cit. on p. 33).
- [46] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. “Autoencoding beyond pixels using a learned similarity metric”. In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566 (cit. on p. 26).
- [47] I. Lazar and A. Hajdu. “Retinal microaneurysm detection through local rotating cross-section profile analysis”. In: *IEEE transactions on medical imaging* 32.2 (2012), pp. 400–407 (cit. on p. 33).
- [48] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. “Diverse image-to-image translation via disentangled representations”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 35–51 (cit. on pp. 4, 10).

- [49] R. Lee, T. Y. Wong, and C. Sabanayagam. “Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss”. In: *Eye and vision* 2.1 (2015), pp. 1–25 (cit. on p. 29).
- [50] H. A. Leopold, J. Orchard, J. Zelek, and V. Lakshminarayanan. “Use of Gabor filters and deep networks in the segmentation of retinal vessel morphology”. In: *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XV*. Vol. 10068. International Society for Optics and Photonics. 2017, 100680R (cit. on p. 31).
- [51] A. Levenkova, A. Sowmya, M. Kalloniatis, A. Ly, and A. Ho. “Lesion detection in ultra-wide field retinal images for diabetic retinopathy diagnosis”. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics. 2018, p. 1057531 (cit. on p. 33).
- [52] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He. “Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs”. In: *Ophthalmology* 125.8 (2018), pp. 1199–1206 (cit. on p. 31).
- [53] Z. Li, C. Guo, D. Nie, et al. “Deep learning for detecting retinal detachment and discerning macular status using ultra-widfield fundus images”. In: *Communications biology* 3.1 (2020), pp. 1–10 (cit. on p. 33).
- [54] Z. C. Lipton. “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57 (cit. on p. 9).
- [55] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem. “On the fairness of disentangled representations”. In: *arXiv preprint arXiv:1905.13662* (2019) (cit. on pp. 9, 18, 20).
- [56] F. Locatello, S. Bauer, M. Lucic, et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124 (cit. on pp. 9, 27).
- [57] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. “Weakly-supervised disentanglement without compromises”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6348–6359 (cit. on p. 4).
- [58] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. “Disentangling Factors of Variations Using Few Labels”. In: *International Conference on Learning Representations*. 2019 (cit. on pp. 4, 27).
- [59] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. “The Variational Fair Autoencoder”. In: *International conference on learning representations*. 2016 (cit. on pp. 14, 23).
- [60] D. Madras, E. Creager, T. Pitassi, and R. Zemel. “Learning adversarially fair and transferable representations”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3384–3393 (cit. on pp. 15, 21).
- [61] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. “Adversarial autoencoders”. In: *arXiv preprint arXiv:1511.05644* (2015) (cit. on p. 13).
- [62] S. Matsuba, H. Tabuchi, H. Ohsugi, et al. “Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration”. In: *International ophthalmology* 39.6 (2019), pp. 1269–1275 (cit. on p. 33).
- [63] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. “A survey on bias and fairness in machine learning”. In: *arXiv preprint arXiv:1908.09635* (2019) (cit. on p. 4).
- [64] D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. “Invariant representations without adversarial training”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 9102–9111 (cit. on pp. 10, 17, 18).

- [65] T. Nagasawa, H. Tabuchi, H. Masumoto, et al. “Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatment-naïve proliferative diabetic retinopathy”. In: *International ophthalmology* 39.10 (2019), pp. 2153–2159 (cit. on p. 33).
- [66] V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814 (cit. on p. 35).
- [67] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon. “Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images”. In: *Scientific Reports* 11.1 (2021), pp. 1–9 (cit. on p. 33).
- [68] H. Ohsugi, H. Tabuchi, H. Enno, and N. Ishitobi. “Accuracy of deep learning, a machine-learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment”. In: *Scientific reports* 7.1 (2017), pp. 1–4 (cit. on p. 33).
- [69] C. O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016 (cit. on p. 4).
- [70] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. “Learning and transferring mid-level image representations using convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724 (cit. on p. 7).
- [71] J. I. Orlando, E. Prokofyeva, M. Del Fresno, and M. B. Blaschko. “An ensemble deep learning based approach for red lesion detection in fundus images”. In: *Computer methods and programs in biomedicine* 153 (2018), pp. 115–127 (cit. on pp. 31, 33).
- [72] R. B. Parikh, S. Teeple, and A. S. Navathe. “Addressing bias in artificial intelligence in health care”. In: *Jama* 322.24 (2019), pp. 2377–2378 (cit. on p. 24).
- [73] S. Parisot, S. I. Ktena, E. Ferrante, et al. “Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease”. In: *Medical image analysis* 48 (2018), pp. 117–130 (cit. on p. 25).
- [74] D. Pedreshi, S. Ruggieri, and F. Turini. “Discrimination-aware data mining”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 560–568 (cit. on p. 14).
- [75] P. Porwal, S. Pachade, R. Kamble, et al. *Indian Diabetic Retinopathy Image Dataset (IDRiD)*. 2018 (cit. on p. 31).
- [76] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen. “Unsupervised deformable registration for multi-modal images via disentangled representations”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 249–261 (cit. on pp. 4, 18).
- [77] K. Ridgeway. “A survey of inductive biases for factorial representation-learning”. In: *arXiv preprint arXiv:1612.05299* (2016) (cit. on p. 3).
- [78] K. Ridgeway and M. C. Mozer. “Learning Deep Disentangled Embeddings with the F-Statistic Loss”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 2018, 185–194 (cit. on p. 14).
- [79] A. G. Roy, S. Conjeti, S. P. K. Karri, et al. “ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks”. In: *Biomedical Optics Express* 8.8 (2017), pp. 3627–3642 (cit. on p. 31).
- [80] P. C. Roy and V. N. Boddeti. “Mitigating information leakage in image representations: A maximum entropy approach”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2586–2594 (cit. on pp. 5, 10, 15, 16, 18, 21, 23).
- [81] D. E. Rumelhart, G. E. Hinton, J. L. McClelland, et al. “A general framework for parallel distributed processing”. In: *Parallel distributed processing: Explorations in the microstructure of cognition* 1.45-76 (1986), p. 26 (cit. on p. 3).

- [82] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. Suttorp-Schulten, M. D. Abramoff, and B. van Ginneken. “Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data”. In: *Investigative ophthalmology & visual science* 52.7 (2011), pp. 4866–4871 (cit. on p. 32).
- [83] M. H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, and E. Abouzar. “Microaneurysms segmentation and diabetic retinopathy detection by learning discriminative representations”. In: *IET Image Processing* 14.17 (2020), pp. 4571–4578 (cit. on p. 49).
- [84] M. H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, and A. Eslami. “Multi-scale microaneurysms segmentation using embedding triplet loss”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 174–182 (cit. on p. 49).
- [85] M. H. Sarhan, A. Covita, S. Yu, et al. “Diabetic retinopathy detection in widefield images by using transfer learning from handheld narrow field images”. In: *Investigative Ophthalmology & Visual Science* 61.9 (2020), PB00108–PB00108 (cit. on p. 49).
- [86] M. H. Sarhan, A. Eslami, N. Navab, and S. Albarqouni. “Learning interpretable disentangled representations using adversarial vaes”. In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 37–44 (cit. on pp. 4, 9, 50).
- [87] M. H. Sarhan, K. Makedonsky, M. Mack, M. Durbin, M. Yigitsoy, and A. Eslami. “Deep learning for automatic diabetic retinopathy detection under multiple image quality levels”. In: *Investigative Ophthalmology & Visual Science* 60.11 (2019), PB0105–PB0105 (cit. on p. 50).
- [88] M. H. Sarhan, M. A. Nasser, D. Zapp, et al. “Machine Learning Techniques for Ophthalmic Data Processing: A Review”. In: *IEEE Journal of Biomedical and Health Informatics* 24.12 (2020), pp. 3338–3350 (cit. on p. 49).
- [89] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni. “Fairness by learning orthogonal disentangled representations”. In: *European Conference on Computer Vision*. Springer, 2020, pp. 746–761 (cit. on pp. 5, 49).
- [90] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni. “On the Fairness of Privacy-Preserving Representations in Medical Applications”. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, 2020, pp. 140–149 (cit. on pp. 18, 49).
- [91] M. H. Sarhan, P. Sha, M. Chen, M. K. Durbin, M. Yigitsoy, and A. Eslami. “Deep learning for automatic diabetic retinopathy grading of ultra-widefield fundus images”. In: *Investigative Ophthalmology & Visual Science* 60.9 (2019), pp. 1094–1094 (cit. on p. 50).
- [92] F. Schroff, D. Kalenichenko, and J. Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823 (cit. on pp. 8, 35).
- [93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626 (cit. on p. 9).
- [94] P. S. Silva, J. D. Cavallerano, J. K. Sun, J. Noble, L. M. Aiello, and L. P. Aiello. “Nonmydriatic ultrawide field retinal imaging compared with dilated standard 7-field 35-mm photography and retinal specialist examination for evaluation of diabetic retinopathy”. In: *American journal of ophthalmology* 154.3 (2012), pp. 549–559 (cit. on pp. 30, 34).
- [95] D. Silver, A. Huang, C. J. Maddison, et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489 (cit. on p. 3).
- [96] C. Szegedy, W. Liu, Y. Jia, et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 3).

- [97] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. on p. 32).
- [98] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. “A survey on deep transfer learning”. In: *International conference on artificial neural networks*. Springer. 2018, pp. 270–279 (cit. on p. 7).
- [99] L. Torrey and J. Shavlik. “Transfer learning”. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264 (cit. on p. 6).
- [100] P. Tschandl, C. Rosendahl, and H. Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Scientific data* 5.1 (2018), pp. 1–9 (cit. on pp. 15, 27).
- [101] S. Van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. “Are disentangled representations helpful for abstract visual reasoning?” In: *arXiv preprint arXiv:1905.12506* (2019) (cit. on p. 18).
- [102] C. Wang, A. Elazab, J. Wu, and Q. Hu. “Lung nodule classification using deep feature fusion in chest radiography”. In: *Computerized Medical Imaging and Graphics* 57 (2017), pp. 10–18 (cit. on p. 7).
- [103] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang. “Zoom-in-net: Deep mining lesions for diabetic retinopathy detection”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 267–275 (cit. on pp. 7, 32).
- [104] M. T. Witmer, G. Parlitsis, S. Patel, and S. Kiss. “Comparison of ultra-widefield fluorescein angiography with the Heidelberg Spectralis® noncontact ultra-widefield module versus the Optos® Optomap®”. In: *Clinical ophthalmology (Auckland, NZ)* 7 (2013), p. 389 (cit. on p. 30).
- [105] T. Xiao, Y.-H. Tsai, K. Sohn, M. Chandraker, and M.-H. Yang. “Adversarial learning of privacy-preserving and task-oriented representations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12434–12441 (cit. on p. 23).
- [106] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. “Controllable invariance through adversarial feature learning”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 585–596 (cit. on pp. 5, 15, 16, 21).
- [107] K. Xu, J. Ba, R. Kiros, et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057 (cit. on p. 9).
- [108] H.-C. Yang, F.-S. Tsai, Y.-M. Weng, C.-J. Ng, and C.-C. Lee. “A triplet-loss embedded deep regressor network for estimating blood pressure changes using prosodic features”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 6019–6023 (cit. on p. 8).
- [109] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan. “Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 255–263 (cit. on p. 18).
- [110] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. “Learning fair representations”. In: *International conference on machine learning*. PMLR. 2013, pp. 325–333 (cit. on p. 14).
- [111] R. Zhang, Y. Zheng, T. W. C. Mak, et al. “Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain”. In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 41–47 (cit. on p. 7).
- [112] S. Zhao, J. Song, and S. Ermon. “Infovae: Balancing learning and inference in variational autoencoders”. In: *Proceedings of the aai conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 5885–5892 (cit. on p. 13).

- [113] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232 (cit. on p. 11).

List of Figures

1.1	Fairness dependency graph	7
1.2	Transfer learning	8
2.1	Disentanglement VAE	12
2.2	Latent traversal	15
2.3	Adversarial fairness	16
2.4	Qualitative evaluation of fairness	19
2.5	Proposed fairness framework	22
2.6	Adversarial disentanglement	26
3.1	Fundus image example	31
3.2	Example of a widefield fundus camera	32
3.3	Microaneurysms segmentation training framework	36
3.4	Microaneurysms segmentation inference	37
3.5	Diabetic retinopathy classification	39
3.6	Example of a handheld fundus camera	41

