
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
MASTER THESIS



Interactive Joint Learning of Multi-lesion Segmentation and Classification for Diabetic Retinopathy Grading

submitted by
Hasan Md Tusfiquir Alam
Saarbrücken
June 2022

Advisor:

M.Sc. Duy Minh Ho Nguyen
Interactive Machine Learning Department
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarland Informatics Campus
Stuhlsatzenhausweg 3
Campus D3.2
66123 Saarbrücken
Germany

Reviewer 1: Prof. Dr.-Ing. Daniel Sonntag

DFKI GmbH
Saarland Informatics Campus
Stuhlsatzenhausweg 3
Campus D3.2
66123 Saarbrücken
Germany

Applied Artificial Intelligence
University of Oldenburg
Marie-Curie Str. 1 D-26129 Oldenburg
Germany

Reviewer 2: Prof. Dr. Antonio Krüger

DFKI GmbH
Saarland Informatics Campus
Stuhlsatzenhausweg 3
Campus D3.2
66123 Saarbrücken
Germany

Submitted

1st June 2022

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Declarations

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, _____
(Datum/Date)

(Unterschrift/Signature)

Erklärung

Ich erkläre hiermit, dass die vorliegende Arbeit mit der elektronischen Version übereinstimmt.

Statement

I hereby confirm the congruence of the contents of the printed data and the electronic version of the thesis.

Saarbrücken,-----
(Datum/Date)

(Unterschrift / Signature)

Acknowledgments

First and foremost, I would like to send heartfelt thanks to my advisor, Duy Minh Ho Nguyen (M.Sc.), for his guidance throughout my thesis work. I am immensely grateful for his invaluable advice, continuous support, patience, and encouragement from the time of initial failures until the time I was able to complete my thesis.

I also take this opportunity to thank Prof. Dr.-Ing. Daniel Sonntag and Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) providing me the opportunity to work on this thesis project and supporting me throughout. I am grateful for his guidelines and inspiration in the human-in-the-loop AI paradigm.

Furthermore, I would like to thank Prof. Dr. Antonio Krüger for his precious feedback on the Master's thesis seminar.

Finally, I would like to express a special appreciation to my dearest parents and lovely wife. They have always been by my side to encourage and support me unconditionally during my study time.

Saarbrücken, May 2022, Hasan Md Tusfiqur Alam

Abstract

Diabetic retinopathy (DR) is one of the most common causes of irreversible blindness in the population, and automated DR detection can support ophthalmologists in creating personalized treatments by providing DR grading and lesion regions. In this work, we investigate a joint learning framework to improve the performance of disease grading and multi-lesion segmentation using the interactive machine learning approach. In the machine learning aspect, we integrate new transfer learning mechanisms, learning invariant feature representations by aligning latent feature embedding using tools from Wasserstein distance and adversarial learning-based entropy minimization. These components permit neural networks to train efficiently under sparse training data while remaining generalized under the influences of domain shift problems. Besides, we propose innovative attention strategies at both low- and high- level concepts, allowing the DR grading network to automatically select the most significant lesion information and provide explainable properties. In terms of human interaction, we enable expert users to correct the system's predictions, which may then be used to update the system as a whole. Finally, the strategy can reduce perturbations in labels made by users using attention networks, thereby saving time and accelerating the data annotation process. The empirical experiments validate our method: we outperform common baselines of state-of-the-art systems by a significant margin. Also, our system's performance improves over time when more user feedback is fed into the network, even in a weakly-supervised form.

Contents

1	Introduction	1
1.1	Motivation and Problem Statement	1
1.2	Approach and Contribution	3
1.3	Thesis Outline	5
2	Related Works	6
2.1	Deep-Learning-Based Diabetic Retinopathy (DR) Grading	6
2.2	Explainability for Deep Learning	8
2.3	Interactive Machine Learning Process	8
2.4	Transfer learning and Domain Adaptation	9
3	Technical Background	11
3.1	Generative Adversarial Networks (GANs)	11
3.1.1	Generative Modeling in Computer Vision	11
3.1.2	Principal Mechanisms in GANs	13
3.1.3	Recent Advances of GANs	14
3.2	Self-Attention in Transformer Architecture	17
3.2.1	Fundamental Concepts of Transformer	17
3.2.2	Main Components in Transformer Architecture	19
3.2.3	Vision Transformer Architecture	21
3.3	Explainable Deep Learning	21
3.3.1	Methods for Explaining DNNs	22
3.3.2	Visualization-Based Methods: CAM and GradCam	23
3.3.2.1	CAM Method	23
3.3.2.2	GradCam Method	23
4	Methodology	25
4.1	Methodology Overview	25
4.1.1	Interactive Machine Learning Flow	25
4.1.2	Deep Network Architectures for Lesion Attributes Segmentation and DR Grading Prediction	26
4.1.2.1	Notations and Settings	27

4.1.2.2	Overview Training Procedures	28
4.2	Learning Domain-Invariant Lesion Attributes Segmentation	29
4.2.1	Task Agnostic Transfer Learning for Lesion Segmentation in Source Domain	30
4.2.2	Adversarial Learning on Predicted Segmentation Maps in Source Domain	32
4.2.3	Incorporating Domain Adaptation with Unlabeled Data in Target Domain	32
4.2.3.1	Wasserstein Distance Minimization on Feature Encoder	33
4.2.3.2	Adversarial Entropy Minimization on Target Domain	34
4.3	Integrating Lesion Features into DR Grading Networks	36
4.3.1	Integrating Lesion Features for CNN-based Methods	36
4.3.1.1	Attention Lesion Regions at Low-level Concepts	36
4.3.1.2	Attention Lesion Regions at High-level Concepts	38
4.3.2	Integrating Lesion Features for Transformer-based Methods	40
4.4	Human Interaction with Trained Systems	41
4.4.1	DR Grading Predictions with Explainable Properties	41
4.4.2	Improving System's Performance through User Feedback	42
5	Experiments and Results	44
5.1	Data Description	44
5.2	Evaluation Metric	45
5.3	Implementation Details	46
5.4	Performance of Multi-Lesion Segmentation Task	47
5.4.1	Influence of Pre-training Steps on Lesion Generator Models	47
5.4.2	Influence of Domain Adaption on Lesion Generator Models	49
5.5	Performance of Diabetic Grading Tasks with Attentive Lesion Information	52
5.5.1	Influence of Attention Mechanism on Grading Networks	52
5.5.2	DR Grading Prediction with Explainable Property	56
5.6	Performance of Trained System using User Feedback	56
6	Discussion and Future Works	62
6.1	Discussion	62
6.2	Future Works	63
	Bibliography	65

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Diabetes is a chronic health condition that is estimated to affect about one in every ten people worldwide [1]. According to Yang et al. [2], 40% to 45% of people with diabetes may develop *Diabetic Retinopathy (DR)* during their lifetime. DR is a kind of ocular disease that damages the retina’s blood vessels, and it is one of the leading causes of irreversible blindness. Although the symptoms are diagnosed only in the later stage, people suffering from this disease start to lose vision from an early stage.

To assess the complexity of DR, *International Clinical Diabetic Retinopathy Disease Severity Scale* [3] has used five grades (0-4), including *no DR, mild, moderate, severe, and proliferative* (figure 1.1). In practice, accurate DR grading is a time-consuming task. While most countries are in shortage of qualified ophthalmologists, an automated and intelligent diabetic retinopathy diagnosis system, therefore, can play an important role in supporting ophthalmologists.

Deep learning algorithms have been leveraged for different disease classification tasks, including automated diabetic retinopathy grading [3, 4, 5]. Given a large dataset of image-level grading annotations, these models automatically learn most predictive features directly from images using back-propagation optimization to predict the desired output. However, due to nonlinear multi-layer structure, Deep Neural Networks (DNNs) are black-box models, and often, predictions are non-traceable to humans [6]. Moreover, these models only use global image features during the learning procedure and ignore fine-grained lesion information. There are 4-6 types of lesions that are closely associated with the DR grading. In practice, ophthalmologists grade the severity of this disease for a patient based on the type of lesions and the regions they appear in the retina image [7]. For example, the earliest clinically visible symptoms for grade-1 (mild DR) are microaneurysms (MAs) lesions, which are local capillary dilation and appear as red dots. Grade-2 (moderate DR) contains both MAs and dot or blot-shaped hemorrhages (HEs) [8]. Therefore, integrating lesion features (medical priors) (figure 1.1) with the global image features can boost the prediction accuracy of deep neural classification

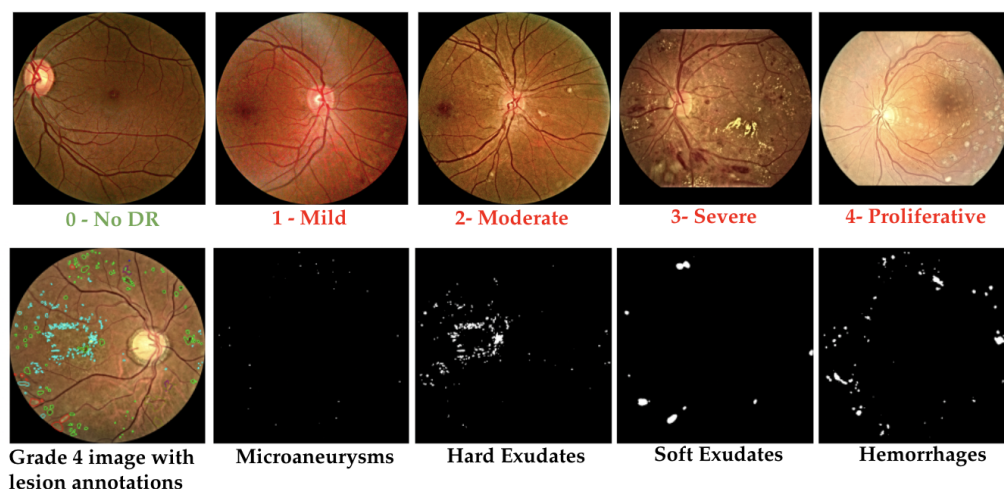


Figure 1.1: Illustration of diabetic retinopathy retinal images. Top row - examples of different stages of DR disease. Bottom row - a retinal image with lesion information annotated by domain expert (left most image). The remaining 4 images are lesion mask predicted by the Dense U-Net Segmentation model [8, 17].

models. Besides, automatic discovery of these lesion regions can assist ophthalmologists in validating and interpreting the prediction of an intelligent diagnosis system.

Detection of these lesion regions is another fundamental challenge in the medical imaging domain. It can be regarded as the multi-lesions semantic segmentation task where labels are assigned in each pixel of a clinical image to different lesion classes [9, 10, 11]. Segmentation models are used as a backbone for many computer vision systems such as autonomous driving [12] and organ localization in radiology systems [13]. However, the task formulations for lesion detection in biomedical images are not straightforward as they present significant technical challenges, including irregularities in shape, noisy and ill-defined boundaries, intra-class lesion diversities [14, 15, 16]. Though disease classification and lesion discovery are inherently correlated, they are mostly studied as independent tasks.

Recently, joint learning methods for classification and segmentation have been studied for some of the medical tasks such as breast cancer diagnosis [18], skin lesion diagnosis [19], or retinal blood vessel analysis [20]. Deep learning models are optimized together to learn disease grading and segmentation tasks simultaneously using the image and pixel-level supervision. Nevertheless, the scalability of these methods is limited in the actual deployment due to the lack of training data. The annotation of fundus images is a costly and labor-intensive task that requires manual labeling by domain experts. For instance, so far, there are only two publicly available datasets with a few hundred of annotated images for the DR related lesion discovery task [8, 21]. Self-supervised learning (SSL) schemes are methods in machine learning where the model trains itself to learn the pretext for a task and transfer the knowledge to address the final tasks. Some methods have utilized the SSL transfer learning concepts to address the lack of training data problems [22]. Moreover, in the scenario of a scarcity of training data, an interactive human-machine learning strategy can be exploited to teach an intelligent DR diagnosis system by involving *human-in-the-loop* of the learning process [23].

The performance of joint learning methods in practice also suffers from the domain-shift

problem when deployed in a cross-domain environment. In the real world, (medical) images are captured from different devices and vendors. Images captured using the same device but different parameter settings can have heterogeneous appearances [24]. For example, though retina fundus images captured across-devices share the same feature space, deep learning models still perform poorly in domain generalization and performance worsens drastically when a model inferred on data-points that are different from the source data using which the model is trained. “Domain adaptation is the ability to apply an algorithm trained in ‘source domains’ to a different but related ‘target domain’ [25]”. It is an active research topic, and some extensive studies have provided theoretical insights about constraining feature representation space to achieve a robust performance across-domain [12, 26, 27, 28]. Another drawback of these methods is explainability. Although highlighted lesion regions can provide visual instructions for the ophthalmologists to assist their diagnosis, they do not explicitly describe how these lesions contribute to the final grading predictions. “Not all lesion information is beneficial to a particular DR severity level, and even some lesion information is noise signal” [7, 29]. Therefore, the importance of each lesion region should be considered separately.

1.2 Approach and Contribution

In this thesis work, we propose a deep-learning-based explainable prediction engine for an intelligent medical diagnosis system focusing on diabetic retinopathy (DR) disease. Inspired by the clinical diagnosis behavior of ophthalmologists, we introduce a novel joint learning system of semantic lesion segmentation and disease grading classification for diabetic retinopathy addressing the challenges we discussed before. Our proposed method simultaneously predicts the disease grading class, associated lesion features, and explanation for the disease grading prediction in a visual form that correlates lesion features with the grading network’s class activation map (CAM).

Our proposed architecture includes a novel *Domain Invariant Lesion Feature Generator* to identify and distinguish important lesion regions of different semantics and an *Attention-Based Disease Grading Classifier* which incorporates the predicted lesion information of the feature generator in the learning process of the disease classification network. Moreover, our framework provides interactive explanations for its decision on a given input. It provides diverse information to the end-users with explainable predictions. In terms of Human-Machine Interaction, we aim to provide a back-end component of an Intelligent User Interface (IUI) [30] using which user/expert can explore the explainable prediction and participate in the training process with minimal data annotation effort to improve model performance in an active learning manner.

In summary, we propose an *interactive joint learning method of multi-lesion segmentation and classification for diabetic retinopathy grading* with the following contributions:

Machine Learning Aspect:

1. We construct a novel lesion attribute segmentation model by formulating a domain invariant learning scheme combining a new transfer learning scheme and domain adaption concepts. We adopt a self-supervised transfer learning method based on Task agnostic [22] to pre-train the segmentation model with limited numbers of available labeled images from the source domain. We further formulated domain adaption constraints to guide our segmentation model in learning domain-

invariant feature representations in an unsupervised manner to attain source-like performance in the target domain.

We constrain our lesion attribute segmentation model to minimize the domain distance in feature representation level by introducing Wasserstein distance [28, 31] based domain critic loss. Moreover, we bound our learning strategy to produce high-confident predictions on target domain images using entropy minimization [12] loss. Our work is the first of a kind that addresses both the challenges of lack of training data and domain adaptation problems for DR-related lesion segmentation tasks. Finally, our segmentation model is trained within adversarial learning [32] to enhance the model’s robustness and accuracy.

2. We construct an attention-based grading network to incorporate the predicted lesion features in the learning process of the disease classification task. Our attention network that identifies and exploits the most significant lesion feature regions is built on low-level and high-level concepts. We integrate the attention block within the structural architecture of the grading network, thus merging the latent features of lesions with the grading network. Our high-level concept for attention is based on explainable principles. We constrain our model to attend to important lesion regions by introducing an explanation loss that directly compares the class activation maps (CAMs) [33] with the lesion regions to guide the network in training. This thesis work demonstrates that our attention methods can be generalized to both CNN-based and Transformer-based neural architecture.
3. Finally, we compare our methods with recent baselines, and our approach outperforms state-of-the-art methods for lesion segmentation and disease grading classification tasks for diabetic retinopathy.

Human Machine Interaction Aspect:

1. Our framework provides diverse information to the end-users with explanations to support its decision. Our joint learning framework is inherently explainable. Our system can provide relevant evidence for the diagnosis by simultaneously predicting medical priors (lesion features) and disease grading. Moreover, our proposed attention-based grading model is constrained to attend and focus on lesion features during its learning process. Therefore, a correlation between the class activation map (CAM), which highlights the class discriminative regions on image input for its prediction, and predicted lesion features can provide a direct explanation to the user.
2. During our method development process, we consider the role of the users in the progressive improvement in model performance. Given various information as output, the users can provide feedback on them and, if required, re-annotate the predicted lesion features, which can be used to fine-tune the model further. We equip our framework with attention methods that make neural networks robust in the presence of noise in the data. This makes the annotation process more comfortable for users as our learning method can leverage noises to a certain threshold, and pixel-wise correction is not required. We have conducted experiments by simulating users, and our experiment results support our study.

1.3 Thesis Outline

The structure of this thesis is as follows:

- In this first chapter, we introduce the motivation for this thesis work and point out the challenges and difficulties in constructing an explainable and intelligent diagnosis system for diabetic retinopathy. We then briefly describe our approach and major contribution ending with the outline for upcoming chapters.
- Chapter 2 covers some related works for this thesis. We discuss the recent deep-learning-based state-of-the-art methods related to diabetic retinopathy grading. The following two sections discuss works related to the explainability of deep learning models and the Interactive Machine Learning process. Finally, we outline some works related to domain adaption and transfer learning techniques. We conclude each section by briefly discussing the drawback of some of these methods.
- Chapter 3 describes different deep learning concepts that have been explored in the thesis work in the formulation of our proposed methods. These technical methods include Generative Adversarial Networks (GANs), attention mechanisms, and transformer architectures. Mathematical formulation and mechanisms of different deep learning model explanation techniques are also described, which we have used to investigate the explainability of our framework.
- Chapter 4 formulates our proposed architecture and details our pipeline. The first section provides a high-level overview of our architecture within the Interactive Machine Learning (IML) framework. The following sections provide the detailed formulation for our deep learning architecture for learning domain-invariant lesion attribute segmentation and the attention mechanism in integrating lesion features for DR grading classification. The Final section details about the explainability of our architecture and the formulation of human interaction in the model learning system.
- Chapter 5 details our experiments and describes implementation setups. We conducted ablation studies for each of the components of our proposed intelligent decision support system. We benchmark our proposed methods for lesion segmentation and DR grading against state-of-the-art baselines on several publicly available datasets and demonstrate significantly better performance. We report some qualitative results of our framework on simulated user-feedback.
- Chapter 6 discusses the advantages of our proposed methods and insights into potential challenges for future investigation.

Chapter 2

Related Works

This section presents the related works for our proposed method. At first, we will discuss existing deep-learning-based methods for diabetic retinopathy grading. The following two sections include the studies related to explainable methods for deep neural networks and user-interactive machine learning methods. Studies related to transfer learning and domain adaptations are discussed in the final section.

2.1 Deep-Learning-Based Diabetic Retinopathy (DR) Grading

DR grading aim is to classify fundus images into different DR severity classes. In recent years, Convolutional Neural Networks (CNNs) architectures have performed exceedingly well and outperformed human experts in many classification tasks. Most recent works construct multi-class classifiers for DR grading leveraging some state-of-the-art models or with their own CNN architectures [3, 34, 35, 36]. These algorithms solved the DR grading as a *black box* classification task and did not consider the fine-grained DR-related lesion information in the learning process. Some researchers attempt to integrate lesion information to improve the grading performance. For instance, Yang et al. [37] propose a two-stages deep convolutional neural networks algorithm to address the DR grading task. In stage one, using pixel-level lesion annotations, a neural network is trained to learn a weighted lesion map. In stage two, a global classification network is trained using this weighted lesion map which provides imbalanced attention on different locations of the fundus image so that more severe lesions will attract more attention in training. Lin et al. [7] also describe a similar framework to automatically detect missing lesion features and integrate with global image features using a classification network for grading prediction. Another difference of this method from the previous one is that instead of the pixel-level annotations where each pixel of an image is labeled with one or more lesion types, the deep learning model was trained using patch-based annotations. In patch-level annotation, an image is split into n number of patches, and each patch is labeled with a particular lesion type. Antal and Hajdu [4] introduced an ensemble-based algorithm to detect a particular lesion and predict DR severity based on the presence or

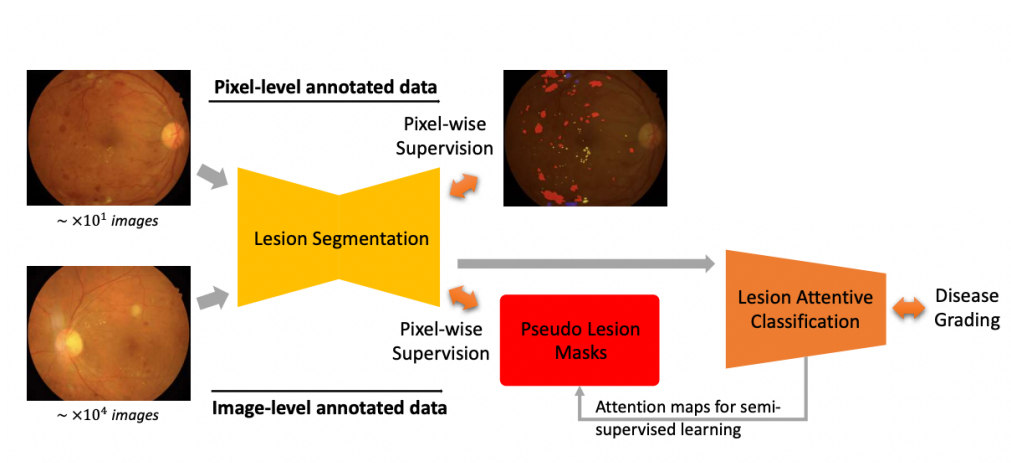


Figure 2.1: An overview of collaborative learning method of semi-supervised multi-lesion segmentation and disease severity classification introduced by Zhou et al. [38].

absence of that lesion.

Although the aforementioned methods incorporate the lesion information, they construct a one-way feature transmission, i.e., the lesion-information extractor modules and DR grading modules are trained separately, and they cannot be optimized jointly for the final tasks. Zhou et al. [38] recently proposed a collaborative learning pipeline to jointly improve the performance of the lesion segmentation and disease grading models by semi-supervised learning with an attention mechanism (figure 2.1). The intuition about this joint learning technique is that *"accurate lesion detection can make considerable contributions to classifying disease grades. Likewise, class-specific information can also benefit lesion segmentation performance"* [38]. At first, similar to the previously discussed methods, a multi-lesion map generation segmentation model is trained using a small set of pixel-level annotation data. Then, based on initially predicted lesion maps for a large scale of image-level annotated images, a lesion attentive grading model is trained. During the training process, this lesion attention model also predicts the fine-tuned lesion maps adopting weak supervision of class-specific information. These predicted fine-tuned lesion maps are used to train the previous segmentation model further. Similarly, Yang et al. [2] also introduced an attention-based learning scheme, which collaboratively learns to predict lesion maps patches and DR grading. In this method, the feature generation model and classification model are connected through an attention model and the loss functions for both of the models are optimized together in an end-to-end manner. Attention mechanisms are introduced by Vaswani et al. [39] initially to address Natural Language Processing (NLP) task. Nevertheless, since then, they have been studied for many vision tasks, including image classification [40], semantic segmentation Yu et al. [41]. In the context of neural networks (which is considered to be an effort to mimic human brain action in a simplified manner), attention is a technique that mimics cognitive attention behavior. Attention mechanisms attempt to selectively concentrate on a few relevant parts of the data during the learning process while ignoring the rest. A visual attention model learns through training data by gradient descent optimization to extract and highlight task-specific salient regions and fade out the rest of the parts of an input image. Sun et al. [29] propose a multi-head attention-based method, which combines several different attention mechanisms to direct the overall attention of a network to consider lesion region diversity and their importance separately in the final DR

grading prediction. While most of the methods mentioned above can achieve promising performance, they share the following limitations:

- Domain shift problem, i.e., the change in data distribution between an algorithms' training dataset and the data it encounters when deployed, was mostly not considered. Thus, these methods suffer significantly in the cross-domain environment.
- Another general drawback for the above-mentioned methods is the lack of explainability on how this multi-lesion information contributes to the decision-making process. Zhou et al. [38] and Yang et al. [2] predict both multi-lesion maps and diabetic retinopathy grading simultaneously as output, but they do not explain how each of these lesion maps influenced the final DR grading prediction. Not all lesion regions should have equal importance for a particular DR grade.

2.2 Explainability for Deep Learning

The decisions of Deep neural networks can be explained by highlighting the saliency regions of an input that strongly influence the output. One approach for underlining the saliency regions is to generate augmented datasets for the predictors and compare them with the network output for an input [42, 43, 44]. Other approaches include evaluating gradient signals passed from output to input during network training. For example, [45] describe the procedure of generating *class activation maps* (CAM) for convolutional neural network models using global average pooling (GAP). A class activation map for a particular category is the discriminatory image regions used by the model to identify that category. Global Average Pooling operation is performed on the last convolution layer of a CNN architecture to get a single feature vector containing the information for all the activation maps, and using this feature vector, weights for each of the activation map is computed. Finally, the class activation map is computed by combining these weighted activation maps. In order to produce explainable output with class activation maps, retraining is required to learn these activation weights. To address this issue, [46] proposed gradient-based localization (Grad-CAM), which uses gradients with respect to each of the activation maps as weights, and requires no further training. The method also combines guided-backpropagation to get a variant of Grad-CAM named Guided Grad-CAM. Different variants of CAM and Grad-CAM techniques have been used in some medical image studies for explainable output. Most recently, Nguyen et al. [47] proposed a CAM-based explainable COVID-19 detection method using CT images. Similarly, Wu et al. [48] adopt the Grad-CAM technique to perform the joint task of classification and segmentation for COVID-19 detection.

2.3 Interactive Machine Learning Process

Holzinger [23] defined the term Interactive Machine Learning (IML) as "*algorithms that can interact with agents and can optimize their learning behavior through these interactions, where the agents can also be human.*" Automated Machine Learning (AML) approaches require large numbers of training samples and suffer in performance when samples are insufficient. IML approaches can be of help, especially in the health domain, where we are often confronted with a smaller number of datasets or rare events. This learning process allows users to examine the impact of their actions and, when required, can adopt

subsequent inputs to obtain the desired behavior [49]. For building a computer-based interactive support system, the initial challenge was to represent the semantics in a machine-readable form using medical ontologies [50] so that the information is easily exchangeable between human experts and machines. RadSpeech's [51] *Mobile Dialogue System for Radiologist* provides a multi-modal interaction system for radiology image annotations. It is a user-friendly semantic search interface where users can annotate medical image regions with a specific medical, structured diagnosis using speech and pointing gestures. Prange et al. [52] present a medical decision support system inside virtual reality (VR). In this system, a doctor can visualize patients' records and clinical images, as well as therapy predictions which are computed in real-time using a pre-trained deep learning model. The aforementioned studies provide the techniques to capture user feedback (annotations on inputs) for an intelligent learning system. However, they did not consider the inclusion process of these feedbacks in an IML loop.

Sonntag et al. [53] describe the functionality and interface of an interactive decision support system for differential diagnosis of malignant skin lesions. They deploy two deep neural networks, an encoder-decoder-based segmentation network (U-Net [15]) to extract the shape, location, and features of a lesion and a CNN based network (VGG16 [54]) to classify the lesion type. They provide a web interface to introduce inter-activeness in the understanding of the prediction explanation. Based on the explanation, the user can re-annotate an input that is used to fine-tune the prediction model. Recently, Dai et al. [55] presented a real-time deep learning based interactive system for diabetic retinopathy. Their architecture has three sub-networks to perform different tasks one by one. The assessment sub-network to assess the quality of the real-time input image, the lesion-aware sub-network to highlight that performs the lesion-segmentation task, and the DR grading sub-network which predicts the DR grading severity. All of these sub-networks have some parameters sharing with the same architectural backbone (ResNet [54]), but each of these tasks was considered as independent, and the coherent relation between lesion discovery and disease classification was mostly ignored. Thus, it is hard to conclude that lesion-feature information predicted by the segmentation part has any influence on the classification result.

2.4 Transfer learning and Domain Adaptation

Transfer learning is a widely used technique in deep learning, especially for solving a new problem for which data are limited. *"It focuses on storing knowledge that is gained while solving one problem and reusing this learned knowledge during solving a related problem"* [56]. However, the performance of a deep neural model trained in a particular source domain, when transferred to a different target domain (e.g., different vendor, acquisition parameters), can drop unexpectedly due to domain shift [57]. Domain adaptation is a sub-category of transfer learning which is the ability to apply an algorithm trained in one or more *source domain* to a related *target domains* where both source and target domain have the same feature space [58]. To address the problem of limited numbers training of data and the domain shift issue, we will explore to incorporate both transfer learning techniques and domain adaptation constraints in the learning process of our task.

In medical image analysis, transfer learning is a commonly used strategy. Rather than training a network with limited training data from a target task, the network is first trained for a task with potentially larger source datasets, creating a more robust model. This pre-trained network is trained to adjust for the target task. Most of the deep learning libraries (e.g., Pytorch, Keras) provide a pre-trained model for almost all state-of-the-art

DNNs models (e.g., VGGs, ResNet, DenseNet). These models are generally trained using a large-scale ImageNet dataset [59] which contains 1000 classes with thousands of images for each class. The popular transfer learning strategy is to use these pre-trained models as initialization to construct additional task-specific components. As a norm for practitioners, this transfer learning approach has been adopted in many existing segmentation methods. However, a large-scale analysis about the benefit of this strategy has been studied recently by Cheplygina [60]; He et al. [61]. Their findings suggest that this learning strategy is not better than random initialization for most medical tasks. Medical images are significantly different from the ImageNet dataset; therefore, features learned using the ImageNet dataset are not helpful for the target medical domain task. Secondly, medical data are often imbalanced. Nguyen et al. [22] describe a novel transfer learning strategy named *Task Agnostic Transfer Learning (TATL)* motivated by dermatologists' behavior in the skincare context. In a two-stage learning step, an attribute-agnostic network is trained at first, which detects all the lesion regions irrespective of their labels. Then the knowledge from this network is transferred to a set of attribute-specific classifiers to label each particular region. Their work also provides theoretical insights and explanations on why this method works well in practice. We aim to apply this transfer learning scheme for DR grading tasks.

Similarly, domain adaption is an efficient way to address the inadequate training data problem. The aim here is to reduce the domain discrepancies between source and target domains. Networks are trained with domain adaptation constraints to be optimized to address the domain-shift problem in deployment. Tzeng et al. [26] proposed a Deep Domain Confusion (DDC) method to reduce the divergence between two distributions by minimizing the maximum mean discrepancy (MMD) loss [62]. MMD is a nonparametric metric and can be defined by "*the idea of representing distances between distributions as distances between mean embeddings of features*" [63]. In their method, a network is trained with data from multiple distributions using a loss function that consists of both task-specific loss and MMD loss. Some studies apply adversarial optimization to remove the domain discrepancy by incorporating generative adversarial networks. Generative adversarial networks (GANs) have two model models; a generator that generates an image from a distribution and a discriminator which evaluates the image [32]. The two networks compete with each other to have accurate predictions. Tzeng et al. [27] combined standard adversarial loss with the task-specific classification loss to minimize domain distances. At first, using labeled data from the source domain, a source encoder CNN is trained. Then, using GANs adaptation, a target encoder is learned such that a discriminator that observes encoded source and target examples cannot reliably predict their domain label. Shen et al. [28] describe Wasserstein distance guided representation learning (WDGRL) method to reduce the domain discrepancy by minimizing Wasserstein distance for each feature block of the encoder CNN.

Chapter 3

Technical Background

This chapter describes different deep learning concepts and frameworks that have been explored in this thesis work. In the first section, we provide the theoretical and technical insight of Generative Adversarial Networks (GANs) [32], which we have used in the formulation of our domain invariant lesion feature generator network described in the Methodology chapter in section 4.1.2. In next section we introduce the fundamental concepts of Self-Attention mechanism [39] and Vision Transformer architecture [64]. Our proposed grading network is built on the attention concept, which we have formulated in section 4.3. In the final section of this chapter, we discuss formulations for explainability in deep learning methods. We describe several explanation visualization techniques e.g. CAM [33], GradCAM [46] we have used in formulation of our proposed explanation loss described in section 4.3 of the methodology.

3.1 Generative Adversarial Networks (GANs)

3.1.1 Generative Modeling in Computer Vision

Generative modeling is an unsupervised learning task that takes training samples from a distribution as input and learns a model representing that distribution. This technique aims to generate new samples from the learned distribution, which is expected to be close to that of the original dataset. In this section, we would like to briefly introduce deep generative (likelihood-based) model as Variational Autoencoder (VAE) [65, 66] and focus on the Generative Adversarial Networks (GANs) [32].

One of the earlier versions of VAE is the Autoencoder network [67, 68]. This network consists of two components: an encoder to make network learn a compressed latent representation of data (e.g latent vectors) and a decoder to reconstruct original data (figure 3.1). Reconstruction loss (e.g L2 distance) is a performance measure of an autoencoder that forces the latent representation to capture as much information about the data as possible.

Variational Autoencoder is a variation of Autoencoder [65, 66] that models the latent

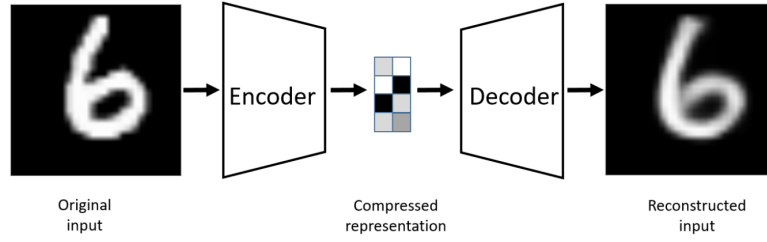


Figure 3.1: An autoencoder architecture (taken from [67]). The input data is encoded to a compressed representation and then decoded.

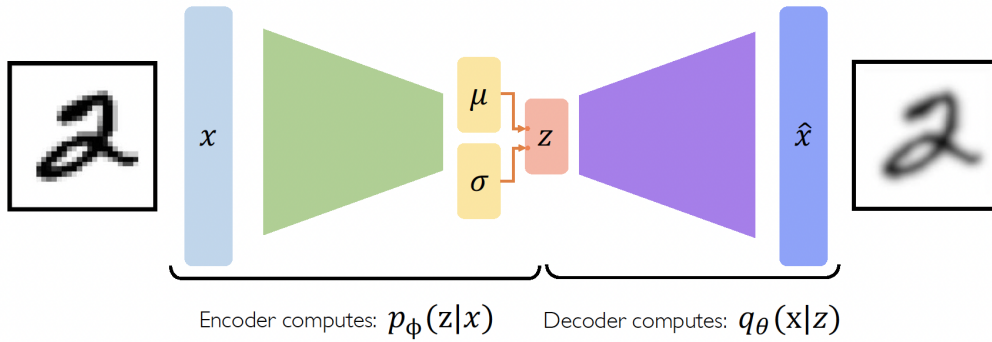


Figure 3.2: Variational Autoencoder architecture (taken from [67]). The latent space z is modeled as a distribution parameterized by μ and σ rather than a deterministic vector as in AutoEncoder. Image taken from [69].

space as a distribution for generative process instead of producing latent vectors only. This model also addresses the problem of overfitting by adding a Kullback-Leibler divergence between prior on the latent variables $p(z)$ and variational distribution $q_\phi(z)$, where ϕ represents parameters of the distribution (figure 3.2). These parameters are sampled by *reparameterization trick* and can be optimized by a neural network (decoder) during the training process.

The problem of VAE is that there are several assumptions required to make it work properly such as latent variable structure, specific likelihood forms and variational posterior. It motivates for a question whether there exists a way to model data distributions better with fewer assumptions? Goodfellow et al. in [32] proposed GANs method to address this challenge by not to explicitly model the density, instead directly generating new instances. However, it is hard to directly sample data from a complex distribution. One idea is to transform data given initial noise using a flexible function: sampling an initial noise vector from the prior $z \sim p(z)$, then deterministically transform z into the target data via an function $x = f_\theta(z)$. The process defines a valid density on the output space $p_\theta(x)$ that is intractable due to computing integral and taking partial derivatives being inefficient in a high-dimensional space. GANs implements this strategy into the Generator to create imitation of the data from noise. In the next section, we will present main mechanisms in GANs and the its current advanced versions.

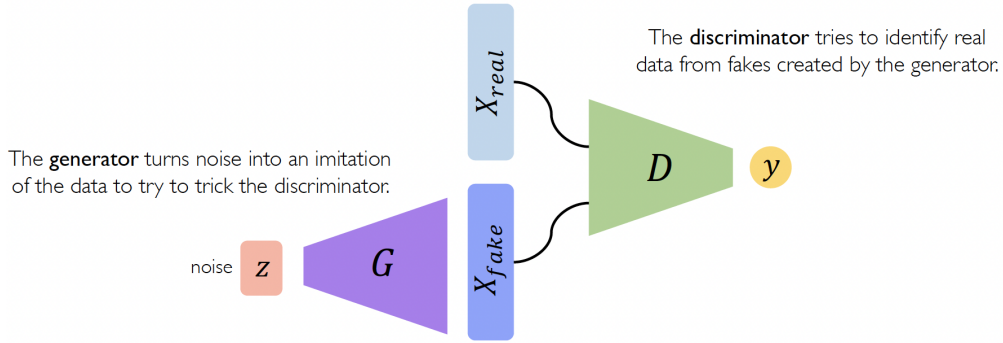


Figure 3.3: The architecture of GANs. Image taken from [69].

3.1.2 Principal Mechanisms in GANs

At the glance, GANs is a generative modeling class that includes two neural networks compete with each other. In other words, GAN tries to learn Generator via *class probability estimation* where the Generator and Discriminator are neural networks. Figure 3.3 illustrates general principles of Generator and Discriminator. Intuitively, the generator tries to create fake samples to trick the discriminator, and the discriminator tries to identify real data from imitated ones generated from the generator.

Training GANs: Loss Function

Denote D, G be the adversarial objectives for discriminator and generator respectively. The global optimum is reached when G reproduces the true distribution of original data. The objective of discriminator D is to maximize the probability of identifying fake data. It is the cross-entropy loss of the true distribution and the distribution generated by the network:

$$\arg \max_D \mathbb{E}_{z,x} [\underbrace{\log D(G(z))}_{\text{Fake}} + \underbrace{\log(1 - D(x))}_{\text{Real}}] \quad (3.1)$$

As generator cannot directly access to the true data distribution, it focuses on minimizing distribution of $D(G(z))$, which is equivalent to minimizing the probability of correctly identifying generated data as “fake”:

$$\arg \min_G \mathbb{E}_{z,x} [\log D(G(z)) + \log(1 - D(x))] \quad (3.2)$$

Combining together, the two networks play a minimax game. This is a bilevel optimization problem:

$$\arg \min_G \max_D \mathbb{E}_{z,x} [\log D(G(z)) + \log(1 - D(x))] \quad (3.3)$$

Generating New Data with GANs

After fully training the generator, it is used to generate new data instances that have never been seen before from the learned distribution. For example, figure 3.4 shows a

sample drawn from Gaussian noise is fed into the trained generator to create a duck image that maps to a data point from the learned target data distribution.

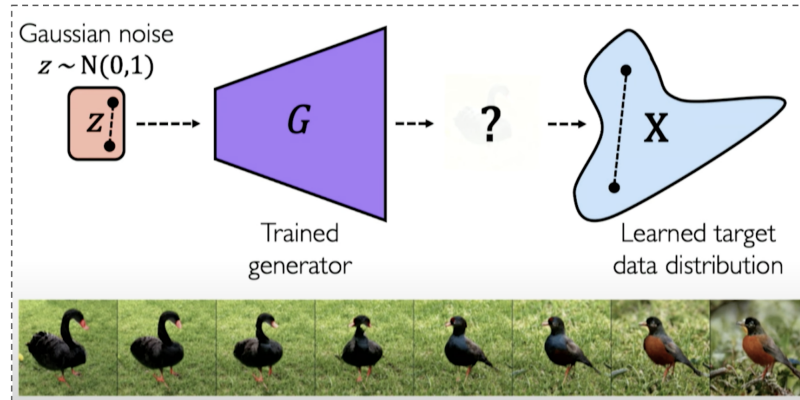


Figure 3.4: GANs as distribution transformers. Image taken from [69]

This mapping is learned over the training itself. In addition, one can perform interpolation from the noise space to create many new instances in the target distribution space. This procedure is illustrated in figure 3.5.

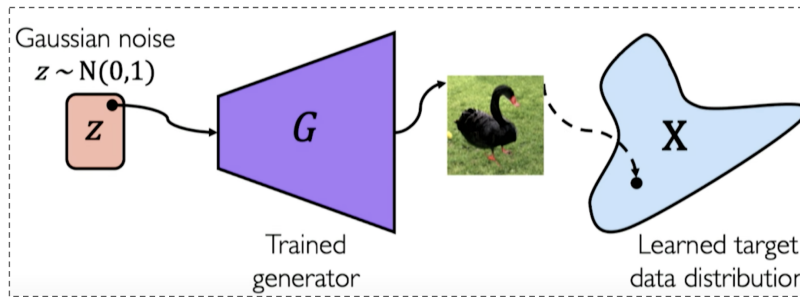


Figure 3.5: Interpolation in the noise space to create new image variations. Image taken from [69]

3.1.3 Recent Advances of GANs

This section covers recent advances of GANs and discusses some GAN-based models that produced the latest results in data synthesis.

Progressive Growing of GANs

One of the current GANs advances is Progressive Growing [70], which enables GANs to add a layer to each generator and discriminator as a training function. This helps to iteratively build up more detailed image generations due to progressive training. This process is described in figure 3.6. The results were improvement of quality and spatial resolutions of generated images, speeding up training and making training process more stable.

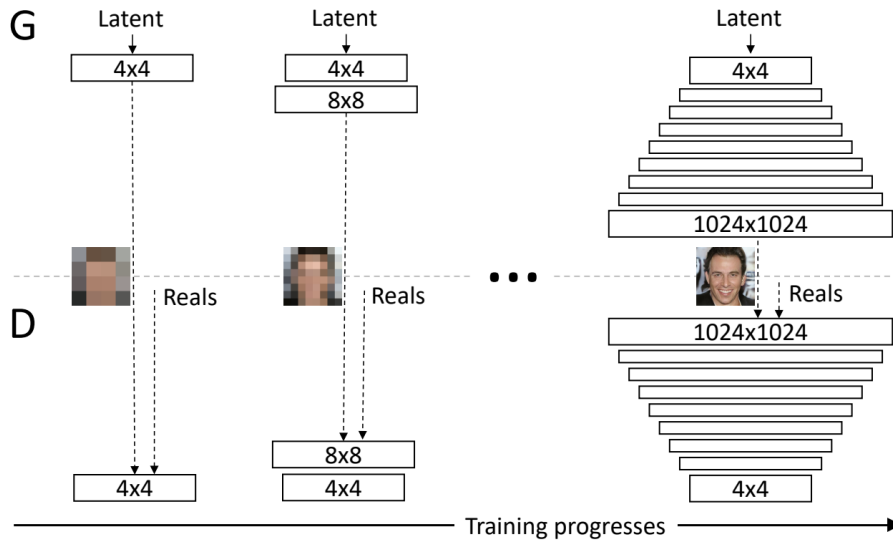


Figure 3.6: Progressive Grow [70].

StyleGAN [71] is another variation that combines progressive growing and style transfer. Features and effects from one series of images can be transferred to a series of target images. This resulted in input target was transformed in style of those source images that features were drawn from. Figure 3.7 compares traditional generator and style-based generator.

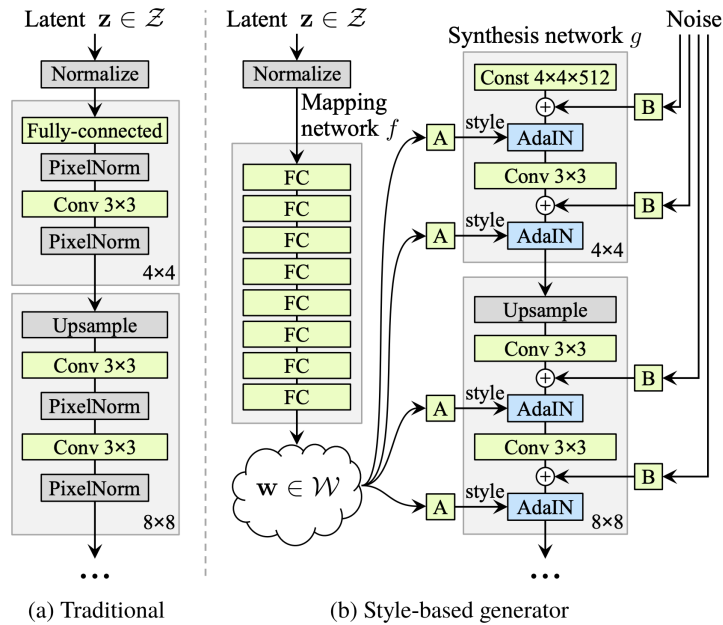


Figure 3.7: Style-based Generator [71].

GANs as data synthesis applications

A conditional version of GANs [70] enables more control on how outputs should look like by conditioning on to both the generator and discriminator. The model can generate MNIST digits conditioned on class labels. In addition, model can be used to learn a multi-model model and provide preliminary instances of an application to image tagging. Figure 3.8 illustrates a simple Conditional GAN.

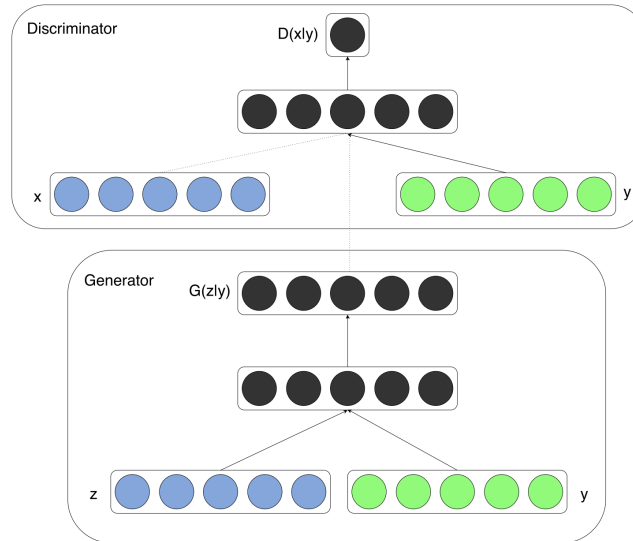


Figure 3.8: Conditional GAN's architecture [70].

One application of Conditional GANs is image-to-image translation using pix2pix software [72]. These networks learn both mapping from input image to output image and a loss function to train this mapping. This allows effective synthesizing photos from map-type inputs, object reconstructions from edges, and coloring from edges of a sketch. Training can also be extended to other instances in the art domain. Figure 3.9 shows several results using Conditional GANs.

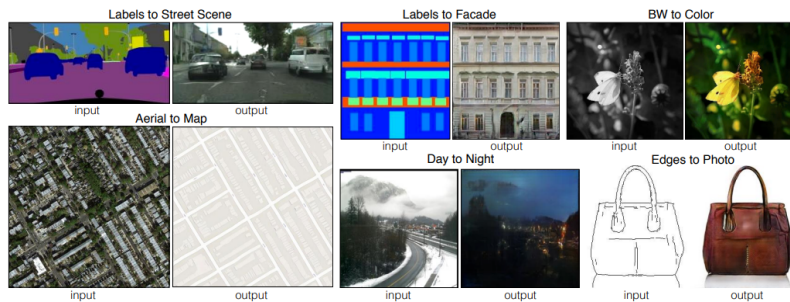


Figure 3.9: Conditional GANs are general-purpose strategy that works well with various problems: labels to street scene, aerial to map, labels to facade, day to night, black-white to color and edges to photo [72].

Finally, CycleGAN [73] is an advanced version of GANs that is used for unpaired image-to-image translation application. While traditional GAN feeds noise into the generator to transform images to the target distribution, CycleGAN learns transformations across domain with unpaired data. The network consists of two generators G and two discriminators D that operate on their own distributions. The objective is to learn a function mapping that translates between two data distributions. Figure 3.10 describes several translation results from this technique.

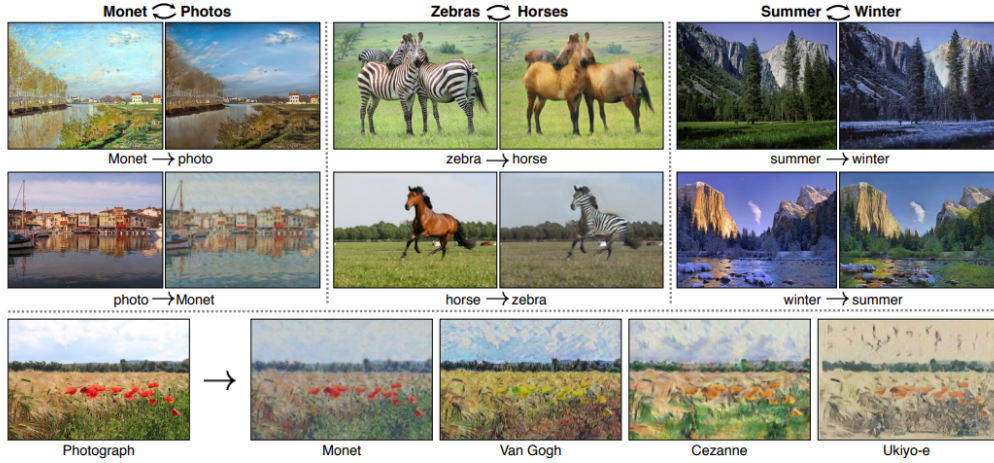


Figure 3.10: Given any two unordered image collections X and Y , image can be transformed from one domain into the other and vice versa [73].

Besides, CycleGAN was applied to speech translating between audio waveforms in Spectrogram image domains. For instance, in [69] authors translated their audios from Spectrogram image domain to that of Former US President Obama in order to synthesize his voice with an impressive performance.

3.2 Self-Attention in Transformer Architecture

3.2.1 Fundamental Concepts of Transformer

Sets and tokenization The proposal of transformer [39] started with a simple idea: exploiting the entire input sequence so that there are no dependencies between hidden states. To this end, a mechanism of encoding sentences into machine-understandable identity numbers is required. This pre-processing step is called tokenization, *i.e.*, creating tokens from the input sentences. Tokenization is an essential process so that model can understand the sentences. For example, the sentence “hello world!” can be represented as “token IDs” as shown in figure 3.11. After tokenization, instead of a sequence of words, we obtain a set of token IDs, where the order of the elements in the set is irrelevant. We denote the input set as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_N\}$ where $\mathbf{x} \in \mathbb{R}^{N \times d_{in}}$, and \mathbf{x}_i is a token. Then, we build word embeddings from the tokenized words, *i.e.*, projecting them into a distributed geometrical space.

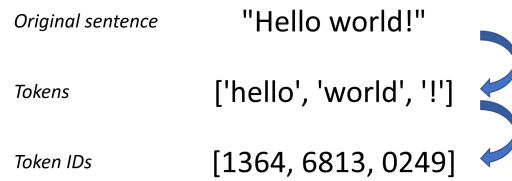


Figure 3.11: Representing a sentence as tokens.

Word embeddings In natural languages, we often encounter similar word meanings or similar grammar structures. Exploiting this property, word embedding represents words in the form of continuous-valued vectors such that vectors that have small distances in the vector space are expected to represent words with similar meanings. As words are not discrete symbols and are strongly correlated with each other, projecting them into a continuous euclidean space can reveal associations between them. Dependent on the task, we can manipulate word embeddings to push them further away or keep them close together. Word embeddings can be projected into 2D or 3D for concise visualization, as shown in figure 3.12. Next, as order is irrelevant in sets, we need a mechanism to produce notion of order in the set so that sentences can be precisely represented.



Figure 3.12: Visualizing word embeddings.

Positional encodings When we convert sentences into sets of tokens, the information of words' order is lost, resulting in broken sentences. To help the neural networks have a sense of order, we slightly modify the word embeddings using their original position. Positional encoding is a set of small constants, which is added to the word embedding vector before the first self-attention layer. When the same word appears in a different position, its actual representation will be slightly different, depending on where it appears in the input sentence.

In the transformer paper, the sinusoidal function is employed for the positional encoding. This function directs the model to pay attention to a specific wavelength λ . Given a signal $y(x) = \sin(kx)$, the wavelength is calculated as $\lambda = \frac{2\pi}{k}$. In positional encoding, the value of λ will be dependent on the position in the sentence. The positional encodings for even and odd positions are defined as follows, given the dimensionality of the embedding vectors is 512.

$$PE_{(\text{pos}, 2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/512}}\right) \quad (3.4)$$

$$PE_{(\text{pos}, 2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/512}}\right) \quad (3.5)$$

3.2.2 Main Components in Transformer Architecture

The Key, Value, and Query

Key-value-query concepts were originally proposed for information retrieval systems. For example, when we search for a particular item, the search engine will map the query (input text) against a set of keys (name, description, etc.) associated with possible stored items. Finally, the search engine will return the best-matched items (values). This is the foundation of content/feature-based lookup. The transformer exploits this idea for constructing its attention mechanism to overcome the bottleneck problem caused by using fixed-length encoding vectors, as the dimension of the representations would be forced to be the same as for both long and short sequences. The mechanism is illustrated in figure 3.13. We can weight the queries by defining a degree of similarity between the representations. We use the keys to define the attention weights to look at the data and the values as the information that we will actually get.

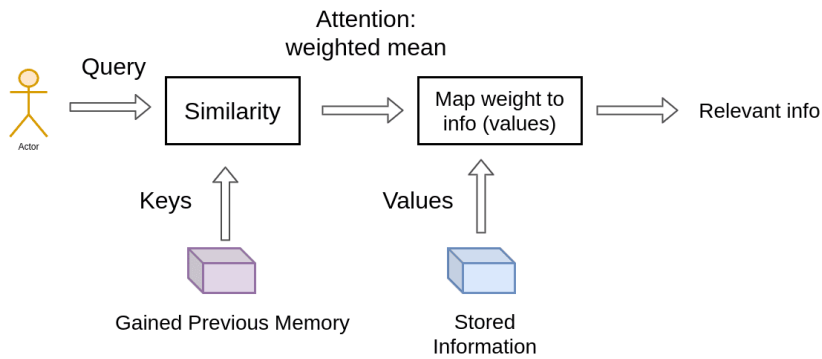


Figure 3.13: The key, value, and query concepts. Image taken from [74].

Self-attention: The Transformer Encoder

Self-attention is a mechanism that connects different positions in a single sequence to produce a representation for that sequence. Self-attention is able to explore correlations between different words in the sentences, providing the grammatical and contextual structures of the sentences. Similarly to the previously mentioned database-query paradigm, this mechanism finds the similarity between the searching query and an entry in a database. Finally, a softmax function is applied to get the final attention weights as a probability distribution. Specifically, the Transformer uses 3 different representations: the Queries, Keys and Values of the embedding matrix. These representations can be computed by multiplying the input $\mathbf{X} \in R^{N \times d_k}$ with 3 different weight matrices \mathbf{W}_Q , \mathbf{W}_K and $\mathbf{W}_V \in R^{d_k \times d_{model}}$. Having the Query, Value and Key matrices, we can now apply the softmax layer to compute the self-attention as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}. \quad (3.6)$$

The process is illustrated in figure 3.14.

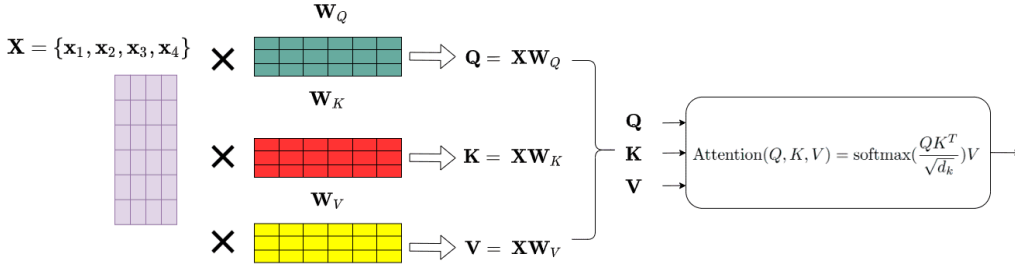


Figure 3.14: The self-attention calculations.

Multi-head Attention

In the original paper, the idea of self-attention is extended to multi-head attention, *i.e.*, the attention mechanism is executed several times. In each run, an independent set of Key, Query, Value matrices is projected into different lower dimensional spaces and the attention is computed in such spaces, producing the outputs called “head”. The idea of multi-head attention is that the model has different and independent ways to understand the input, as it can pay attention to different parts of the sequence in different runs. Consequently, the model can capture better positional and contextual information, producing more robust representations.

The projections of Key, Query, Value matrices are calculated by multiplication with corresponding weight matrices, denoted as \mathbf{W}_i^K , $\mathbf{W}_i^Q \in R^{d_{model} \times d_k}$, and $\mathbf{W}_i^V \in R^{d_{model} \times d_k}$. To reduce the complexity, the output vector size is divided by the number of heads. Specifically, in the vanilla transformer, they use $d_{model} = 512$ and $h = 8$ heads, producing vector sizes of 64. The heads are then concatenated and transformed using a weight matrix $\mathbf{W}^O \in R^{d_{model} \times d_{model}}$, since $d_{model} = h d_k$.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O \quad (3.7)$$

where

$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \quad (3.8)$$

and

$$\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k} \quad (3.9)$$

3.2.3 Vision Transformer Architecture

The Transformer has become the state-of-the-art architecture for natural language processing, however, applications of Transformer to machine vision tasks are still limited even though attention is a crucial information in vision. Seeing the limitation, Dosovitskiy *et al.* [64] proposed Vision Transformer (ViT) to utilize the attention mechanism of Transformer for computer vision tasks. ViT has achieved remarkable results in various vision task. ViT has achieved excellent results compared to state-of-the-art convolutional networks in various image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.) while requiring significantly fewer computational resources.

An overview of ViT is illustrated in figure 3.15. As the original Transformer takes 1D sequences of token embeddings as inputs, we partition the 2D image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where (H, W, C) is the height, width, and number of channels, respectively, of the image, into N patches with sizes of (P, P) , then sequentially concatenating them into a sequence $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ to make compatible inputs for the Transformer architecture. We then flatten the patches and employ a trainable linear projection to map the flattened patches to D -dimension space to create a sequence of “patch embeddings”. Similarly to the original Transformer, we employ learnable 1D position embeddings to preserve positional information, as 2D position embeddings do not provide remarkable advantages. The Transformer encoder take the result patch embedding sequences as inputs.

An additional learnable embedding is added to the patch embeddings sequence ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$). The output of this embedding by the Transformer encoder (\mathbf{z}_L^0) acts as the representation of the input image. In both pre-training and fine-tuning processes, we attach a classification head, which is implemented by a multi-layer perceptron in pre-training and a linear layer in fine-tuning, to \mathbf{z}_L^0 . The architecture of the Transformer is described in the previous subsection. ViT is typically pre-trained on large datasets, and then fine-tuned for downstream tasks. To this end, the pre-trained prediction head is removed and a $D \times K$ feedforward layer is attached, where K is the number classes in the downstream tasks.

3.3 Explainable Deep Learning

There are several surveys on explainable AI [75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85] and explainable deep learning [86, 87, 88]. Even though these surveys provide thorough and comprehensive studies, they cover an enormous of work that readers might find it hard to follow. Instead, we focus on a small number of methods which are foundational. A method is considered as foundational if it is widely used or if it introduce novel concepts that conventional work relies upon. By studying this set of foundational methods, readers might have better insight when they study more modern techniques. We present a simple three-dimensional space encompassing:

- Visualization methods: by employing scientific visualization techniques, visualization methods highlights the characteristics of inputs that strongly affect the outputs

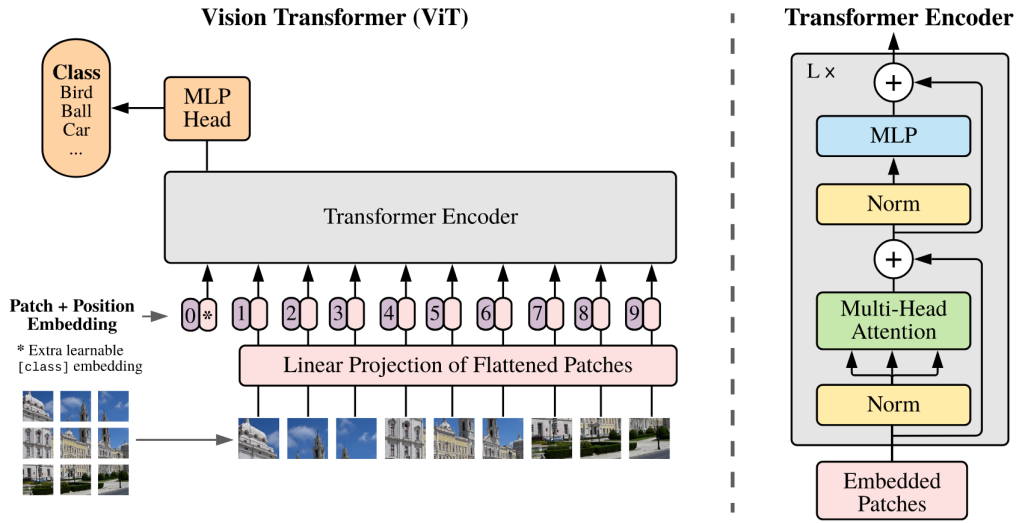


Figure 3.15: The Vision Transformer architecture [64].

of a DNN.

- Model distillation: a separate, “white-box” machine learning algorithm is developed and trained to imitate the behaviors of the DNN. As the white-box algorithm is inherently explainable, we can explore the learned rules or features that affect DNN outputs.
- Intrinsic methods: these methods are actually DNNs developed to produce explanations along with the outputs. Consequently, intrinsic methods can be exploited to simultaneously increase performance and produce high-quality explanation.

In this section, we focus on visualization-based methods, specifically CAM and Grad-CAM.

3.3.1 Methods for Explaining DNNs

Visualization methods associate the degree of importance toward final decisions of the networks to input features. This association is also referred as attribution. Widely-used forms of visualization methods are saliency maps and heatmaps, which are represented by transparent color maps overlaid on the original input images. These maps indicate input features that are most important, *i.e.*, the most influential factors to the model’s output. Visualization methods can be categorized into two types, namely backpropagation and perturbation-based visualization. We will focus on backpropagation-based methods in the scope of this manuscript.

Backpropagation-based methods estimate the importance of input features by evaluating gradient signals during the training process. For example, in a scene recognition problem, a high saliency score for features representing the object “bed”, when a CNN decides that the image belongs to the class “bedroom”, may indicate that the decision of the CNN is highly sensitive to the occurrence of the bed. In [89, 90], the authors visualize the scaled partial derivative of the model’s output with respect to each input feature to identify the corresponding sensitivity. In contrast, other gradient-based methods estimate the

sensitivity with respect to the output by exploiting different feature maps at different CNN network layers [91, 92, 93, 94]. We will focus on CAM and Grad-CAM in the scope of this manuscript.

3.3.2 Visualization-Based Methods: CAM and GradCam

3.3.2.1 CAM Method

Zhou *et al.* [33] proposed a visualization method by creating class activation maps (CAM) using global average pooling (GAP) in CNNs. In [45], Lin *et al.* applied a GAP on the activation maps of the last convolutional layer before those maps are fed to the fully connected (FC) output layer, *i.e.*, the final layers of the CNN are implemented as

$$\text{GAP (Conv)} \rightarrow \text{FC} \rightarrow \text{softmax}. \quad (3.10)$$

The FC layer has C nodes corresponding to C class. The CAM method computes the weighted sum of the activations \mathbf{A}_k produced by Conv , which contains K convolutional filters, using the weights $w_{k,c}$ produced by FC , where the (k, c) pair indicates the specific weighted connection from Conv to FC , to create the saliency map:

$$\text{map}_c = \sum_k^K w_{k,c} \mathbf{A}_k \quad (3.11)$$

The saliency map map_c is then upsampled to match the size of the input images, resulting in the final class activation map. Each class has a unique map, indicating the most influential image regions toward the network prediction for that class. However, CAM can only be employed in CNNs that use the $\text{GAP (Conv)} \rightarrow \text{FC} \rightarrow \text{softmax}$ configuration.

3.3.2.2 GradCam Method

To overcome the limitation of CAM, Gradient-weighted Class Activation Map (Grad-CAM) [46] is proposed. Grad-CAM exploits the gradients of the network output with respect to the last convolutional layer to compute the class activation map. This strategy enables Grad-CAM to be applied to a wider range of CNNs. The only requirement is that the final activation function that produce the network output must be a differentiable function, *e.g.*, softmax .

For each activation map \mathbf{A}_k produced by the final convolutional layer of the network, a gradient of the score y_c (the value before softmax) of class c with respect to every node in \mathbf{A}_k is computed and averaged to get a saliency score $\alpha_{k,c}$ for the activation map \mathbf{A}_k , *i.e.*,

$$\alpha_{k,c} = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n \frac{\partial y_c}{\partial \mathbf{A}_{k,i,j}} \quad (3.12)$$

where $\mathbf{A}_{k,i,j}$ is a neuron at the location (i, j) in the activation map \mathbf{A}_k . Then, the saliency scores of feature maps are combined and passed through a ReLU to produce the saliency map

$$\text{map}_c = \text{ReLU} \left(\sum_k^K \alpha_{k,c} \mathbf{A}_k \right) \quad (3.13)$$

The saliency map map_c is then upsampled to match the size of the input images to produce the final class activation map. Figure 3.16 illustrates the Grad-CAM method and an example of Grad-CAM saliency map for the prediction “cat”.

The CAM-based methods can be exploited to determine, given an input and a class, the most influential information in the input that affect the final decision of the network. From this information, one can interpret the predictions of the network to assess its stability and consistency. For example, given two network that have the same prediction accuracy, the network with saliency maps which are more consistent with human experience is considered more robust and trustworthy than the other. The CAM-based methods can also be utilized to detect class bias.

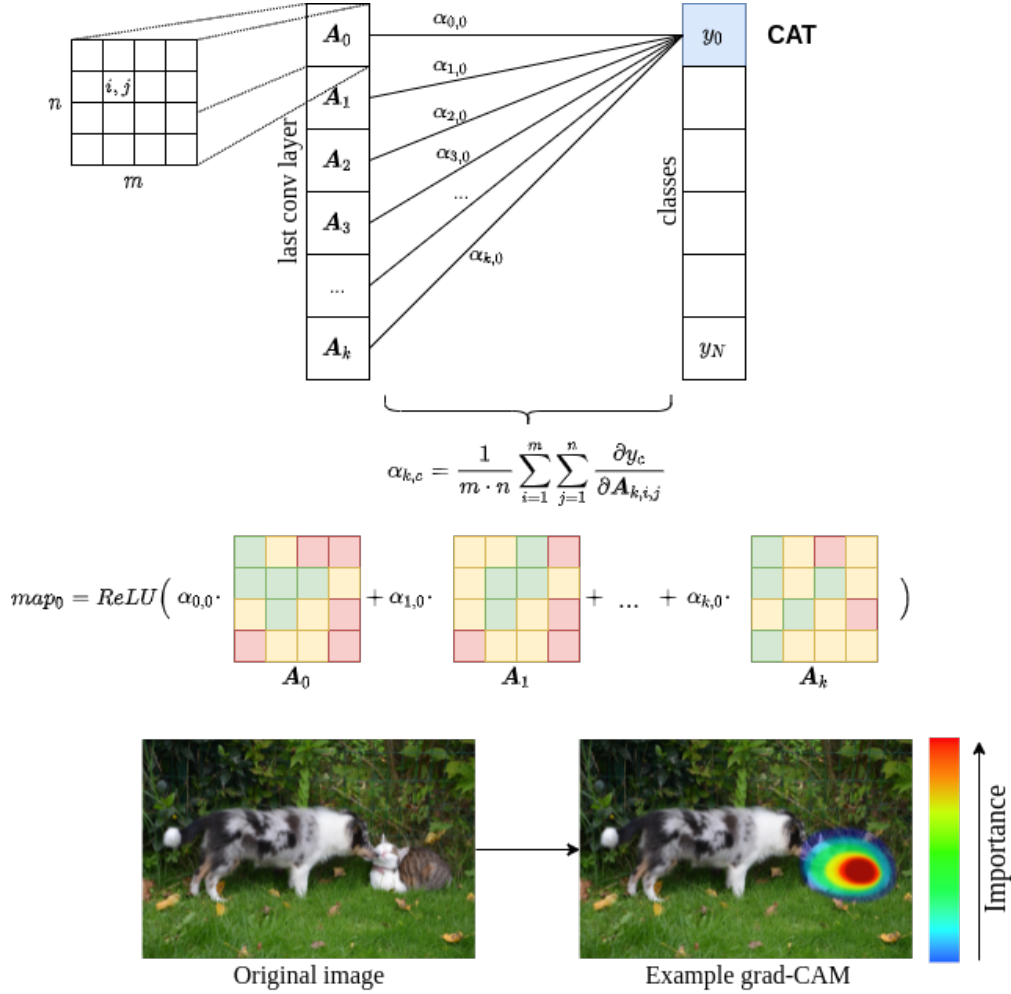


Figure 3.16: Illustration of Grad-CAM method. **Top:** Illustration of (3.12) for calculating the saliency scores $\alpha_{i,j}$ for each activation map A_k . **Middle:** The computation of saliency map for a specific class. **Bottom:** Saliency map for the prediction “cat”. Image taken from [95].

Chapter 4

Methodology

4.1 Methodology Overview

4.1.1 Interactive Machine Learning Flow

This thesis aims to build a deep learning-based prediction engine for an intelligent user-interactive (IUI) diagnosis system focusing on diabetic retinopathy (DR) disease. Our method is inspired by the clinical diagnosis behavior of ophthalmologists in diabetic retinopathy grading. In particular, to determine the severity of the disease, ophthalmologists, in the first step, usually locate and identify different lesion regions by closely observing the retinal image [7]. These lesion attributes are essential clinical features for the disease diagnosis and grading progression on the severity scale. In practice, recent studies [29, 96] have shown that incorporating this additional information in the learning process can play as an expert’s feedback to improve the model performance and robustness for DR grading tasks.

However, as we discussed previously, training a disease grading classification model with additional lesion information requires solving several challenges. First, the dataset containing image-level and pixel-level supervision is expensive to annotate. For example, though more than five retinal image datasets are available containing disease severity grading annotations, there are only two public datasets for retinal lesion annotation. One of them comprises less than 100 images. Second, integrating lesion information into a deep network for classification tasks is still an open problem because each DR grading task depends on a group of different features with an extended level and properties. Finally, it is crucial to consider the users’ roles in this framework; thereby, (i) the system’s predictions could be explainable to the users by providing relevant evidence to medical priors; (ii) the system allows users to inspect predicted results and be able to enhance accuracy given new samples annotated by users which usually contain to a certain noise level.

In this work, we aim to address the above challenges by introducing an interpretable

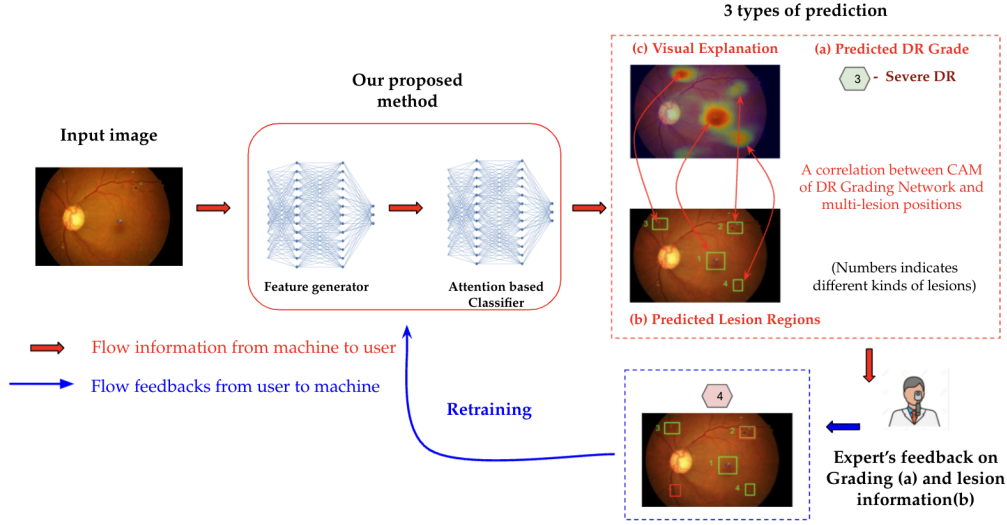


Figure 4.1: A high-level overview of our proposed method in terms of IML workflow. Given a retinal image, our deep learning models will simultaneously generate 3-types of prediction (DR grade, lesion region, visual explanation). By using an Intelligent User Interface (IUI), ophthalmologists can observe the system’s predictions and provide feedback to fine-tune the model.

diabetic retinopathy diagnosis system. The method is depicted in figure 4.1 showing the high level of our framework in terms of the IML workflow. Given a retina image input, our method will simultaneously predict three outputs: DR grade severity on a scale of 0 to 4 (by *Attention-based Classifier*), different lesion regions available in the input image (by *Feature generator*), and a visual correlation between class activation maps of the model with lesion regions has influenced the grading prediction.

In terms of expert users, by observing the network’s prediction with their explanations, an ophthalmologist can validate the output and re-annotate the prediction labels (both grading labels and segmentation masks) as a form of user feedback if required. This feedback can be used to fine-tune the model in the future iteration to obtain better performance. Furthermore, as our proposed network can already highlight the tiny lesion regions, the data annotation effort can be minimal. Also, by equipping attention mechanisms, we enable the user to provide lesion annotations in weak supervision, i.e., sketch a boundary around lesion regions rather than detail segmentation masks. This saves time for annotation effort and is convenient to deploy in practice.

4.1.2 Deep Network Architectures for Lesion Attributes Segmentation and DR Grading Prediction

We now detail two main neural networks in our pipeline: *Feature Generator* and *Attention-based Classifier*, which are illustrated in figure 4.2. In this, the *Feature Generator* is learned using Multi-lesion Segmentation (*S-Net*) and the *Attention-based Classifier* is formulated by an Attention Network (*Att-Net*) and a Grading Network (*G-Net*). Besides, we propose other networks such as Patch Discriminator *PD-Net*, Wasserstein discriminator network

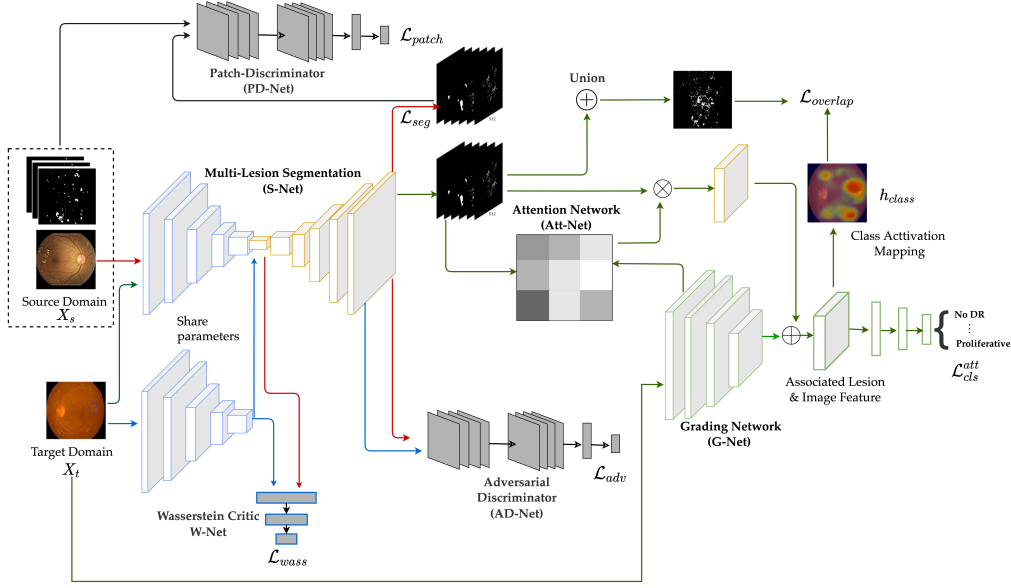


Figure 4.2: Detailed architectures and pipeline of the proposed method. It consists of five inter-dependent neural networks: *S-Net*, *PD-Net*, *AD-Net*, *W-Net*, *Att-Net*, and *G-Net*. These networks are learned in two phases. The first stage aims to train the multi-lesion segmentation networks *S-Net*. To overcome the lack of training data in a source domain and domain adaptation issues in a target domain, *PD-Net*, *W-Net*, and *AD-Net* are jointly integrated. The second stage optimizes the grading network *G-Net* using lesion information obtained from *S-Net*, where the attention network *Att-Net* automatically decides the most influential parts.

(*W-Net*) and Adversarial domain discriminator (*AD-Net*). These components are employed for domain adaptation purposes, which permit *S-Net* to generate predictions in a new target domain without using labeled data in the training step.

4.1.2.1 Notations and Settings

Before describing training procedures for mentioned neural networks, we introduce the notions and settings being used throughout this work. In particular, we require two types of image annotations:

- *Image-level annotations*: An image is labeled with a particular disease grade. There are five [0-4] different disease grades for diabetic retinopathy. '0' is a healthy eye with no sign of diabetic retinopathy, and 4 is the most severe stage (figure 1.1, top row). For a set of images $X \subset \mathbb{R}^{h \times w \times 3}$, we assume to have associated D -class disease grades $y \subset \mathbb{R}^D$, where $D = 5$ and h and w are the height and width of the image.
- *Pixel-level annotations*: Each of the pixels in an image is labeled with a pre-defined lesion class. In this work, we consider four distinct DR-related lesion attributes: Microaneurysms, Hemorrhages, Soft Exudates, and Hard Exudates, denoted as $L = \{\text{MA}, \text{HE}, \text{SE}, \text{EX}\}$ (figure 1.1, bottom row). Formally, each image $x \in X \subset$

$\mathbb{R}^{h \times w \times 3}$, we have associated L -class lesion segmentation maps, $z_l \in \mathbb{R}^{h \times w \times 1}$ where $l \in L$, i.e., four binary segmentation maps, and h and w are the height and width of the image.

Our setting distinguishes data points in two domains:

- *Source domain/dataset* (X_s): The dataset has ground-truth labels used in the training of the model. For this, we assume to have both image-level and pixel-level annotations.
- *Target domain/dataset* (X_t): The dataset which we desire to generate final predictions. In our setting, we suppose only having image-level annotations for the DR Grading task and without (or a few samples) pixel-level annotations for lesion segmentation tasks.

4.1.2.2 Overview Training Procedures

We construct different segmentation models for learning distinct lesion features. This strategy is followed by prior works [22, 38] to avoid the imbalance of data problems among classes. For simplification, when describing the whole system in general, we will use *S-Net* to refer segmentation models with parameters denoted as θ_S . For a specific purpose, we will use S_l with parameters θ_{S_l} , denoting a dedicated model segmenting a feature $l \in L$.

All segmentation models *S-Net* in this work are defined based on the *encoder - decoder* U-Net architecture [15] parameterized by $\theta_S = \{\theta_e, \theta_d\}$. The encoder blocks use *feature extraction layers* of the ResNet50 [97] architecture as the network backbone. We firstly pre-train *S-Net* with data from the source domain X_s using task agnostic transfer learning techniques [22] (section 4.2.1). This network is then fine-tuned using available pixel-level annotations in X_s . To enhance the segmentation quality of *S-Net* for tiny regions, we further apply adversarial learning techniques on the predicted segmentation masks using the *PD-Net* parameterized by θ_{pd} , implying *S-Net* is constrained by a segmentation and another discriminate objective loss functions (section 4.2.2).

While the *S-Net* is required to generate predictions in the target domain X_t , which has a distribution shift compared with the source domain X_s and only has unsupervised training data, we propose a novel unsupervised domain adaptation approach to alleviate this challenge. This includes the adversarial entropy minimization network (*AD-Net*) and the Wasserstein critic discriminator network (*W-Net*) parameterized by θ_{ad} and θ_w respectively. To learn a domain-invariant feature representations for *S-Net*, both three networks *S-Net*, *AD-Net*, *W-Net* are jointly optimized using both pixel-level annotations in X_s and only input images in X_t (section 4.2.3).

To learn a diabetic retinopathy disease grading model, we use available image-level annotations in the target domain X_t and train *G-Net* parameterized by θ_g . Besides, the lesion attention model *Att-Net* parameterized θ_{att} is integrated with *G-Net* to highlight the disease-related lesion attributes obtained from *S-Net*. In this work, we formulate two variations of *Att-Net* for two types of architectures: CNN-based methods (section 4.3.1) and Transformer-based methods (section 4.3.2). Furthermore, we present a new loss function based on overlapping heatmap constraints estimated from class activation mapping of trained networks (*G-Net*) and extracted lesion regions from *S-Net* (section 4.3.1.2). This constraint advances performance overall and generates explainable proper-

ties to the users by observing the correlation of lesion attributes (medical priors) to the class activation map of trained networks (section 5.5.2).

Main Contributions In a nutshell, we make the following contributions:

1. First, we construct novel lesion attribute segmentation models *S-Net* (or *Feature Generator*) that use a new transfer learning scheme based on task agnostic to pre-train models using limited labeled images in the source domain. Domain adaptation concepts are also formulated to guide neural networks in learning domain-invariant feature representations in the target domain given unsupervised data. Furthermore, the adversarial learning-based constraints are also integrated during training processes to enhance the model’s robustness and accuracy.
2. Second, relevant lesion regions extracted by *Feature Generator* are automatically combined for DR grading tasks using attention mechanisms. For this, we propose different attention methods for both CNN and Transformer-based methods. Moreover, a new constraint that explicitly models the relation of lesion attributes with heatmap regions of the DR Grading network is proposed.
3. Third, our framework provides diverse information to the end-users with explainable predictions. Given this, the user can inspect predicted results and give feedback to gradually improve systems’ performance with less effort for annotation tasks.
4. Finally, the empirical experiments in different datasets confirmed the effectiveness of our framework as it improved the performance of several baseline methods by a large margin and demonstrated an increased performance when more data feedback from users was provided.

In the following sections, we describe mathematical formulations for the aforementioned factors in detail.

4.2 Learning Domain-Invariant Lesion Attributes Segmentation

The aim of this section is to learn the *Feature Generator* through *S-Net*. This task relates to semantic segmentation problems [98], where the optimization goal is to solve the pixel-level classification with classes pre-defined. In the context of diabetic retinopathy (DR) images, each pixel can be annotated as one of the four related lesion classes L or healthy. Conventionally, using pixel-level annotation from the source domain X_s , a segmentation model S_l for a lesion attribute $l \in L = \{\text{MA}, \text{HE}, \text{SE}, \text{EX}\}$ is trained in a fully supervised manner:

$$\mathcal{L}_{seg} = \min_{\theta_{S_l}} \frac{1}{|X_s|} \sum_{(x, z_l) \in X_s} \mathcal{L}_{wbce}(S_l(x), z_l) \quad (4.1)$$

where x and z_l are the input image and corresponding segmentation map for the attribute l , \mathcal{L}_{wbce} is the weighted binary cross-entropy loss defined as:

$$\mathcal{L}_{wbce}(z, \hat{z}) = -(\beta \cdot z \log(\hat{z}) + (1 - z) \log(1 - \hat{z})), \quad (4.2)$$

with z is the ground-truth binary lesion mask and \hat{z} is the model prediction, β is the class balancing weight. The loss \mathcal{L}_{wbce} in general can be replaced with other segmentation losses [99].

However, training Eq. (4.1) is (i) hard to converge to a good optimal solution due to a scarcity of training data in the source domain X_s . Besides, (ii) for lesion attributes whose regions are small, neural networks tend to learn for other dominant classes with more extensive areas. Lastly, (iii) optimizing with fully supervised data in one source domain X_s is not yet guaranteed a good generalization to a test case in the target domain X_t due to the domain shift problem.

To overcome those obstacles, we address (i) by using the task agnostic transfer learning framework [22] (section 4.2.1), which enables discovering a shared feature representation space for similar lesion attributes and has been shown to be effective in the skin attribute detection case. We hypothesis that our setting for diabetic retinopathy-related lesions is relevant to the skin attribute detection in the sense that the targeted objects are also small, disconnected regions and non-uniform distributions (figure 1.1, bottom row). In addition, the training data among classes are also insufficient and imbalanced. The second challenge (ii) is handled by equipping the adversarial learning [32], i.e., a two-player game, to force the segmentation outputs of *S-Net* to look "real" as ground-truth data and tiny regions have to be taken into account in training step. We present in section 4.2.2 formulations for this idea. Finally, the domain adaptation problems in (iii) are handled by jointly learning *S-Net* with Wasserstein distance *W-Net* [31] and Adversarial Domain Discriminator *AD-Net* [12] using both labeled data in the source domain and unlabeled data in the target domain. At a glance, these additional networks control learned feature representations of *S-Net* to be invariant across data distributions. The section 4.2.3 describes in detail our approach for this step.

4.2.1 Task Agnostic Transfer Learning for Lesion Segmentation in Source Domain

Class imbalance and lack of training data are common problems in medical domain. The publicly available IDRID [21], which we have used as the source domain X_s for the lesion segmentation task, contains only 53 annotated images with lesion information. For the lesion class *Soft Exudates*, the annotated images are even fewer, only 26. To lessen this issue, a popular transfer learning technique is to use a pre-trained ImageNet model [11, 100] for the weight initialization. However, recent studies [22, 60, 61] have argued that this transfer learning method is sub-optimal in several scenarios and is not consistently better than random initialization for medical image analysis. Nguyen et al. [22] thus has proposed a novel *task agnostic transfer learning* (TATL) approach for skin cancer attribute detection. Motivated by the ophthalmologist's behavior in DR related lesion detection, we have adopted the self-supervised TATL technique containing an *attribute-agnostic segmenter* and a *task-specific classifier* for detecting each of the lesion regions.

Figure 4.3 describes our adopted TATL framework. In the first step, we train the *attribute-agnostic segmenter* with encoder-decoder layers as the U-shape network, denoted as $S_U = \{S_U^e, S_U^d\}$ where S_U^e is the encoder and S_U^d is the decoder part of the model. To train S_U , we define an intermediate dataset of attribute agnostics $D_U = \{X_s, M_U\}$ where M_U is the corresponding binary mask to an image whose value is 1 whenever a pixel is an attribute from L . In our setting, given an image $x \in X_s$ and a set of attributes masks z_l , ($l \in L$), M_U is the union of all the masks (figure 4.3) and can be easily constructed by

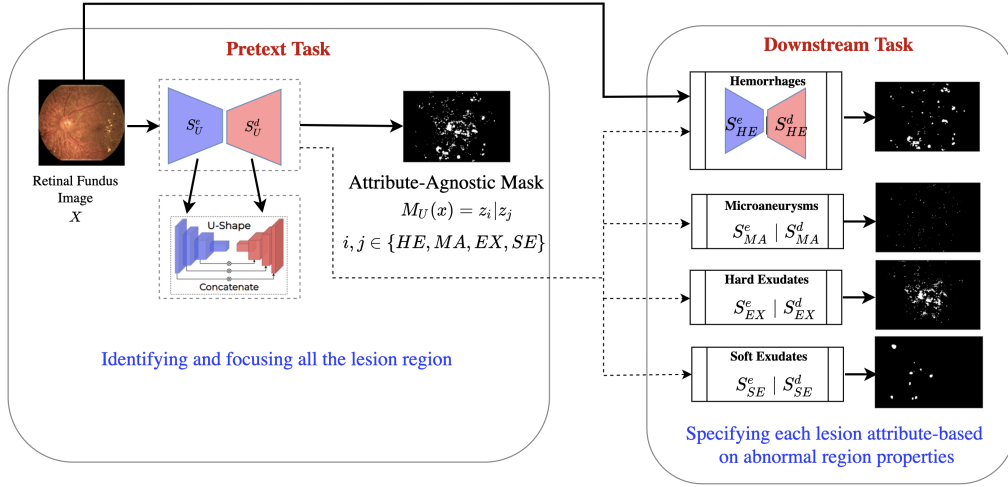


Figure 4.3: Task agnostic transfer learning (TATL) procedure used in the pre-training step in lesion attribute generation. *Pretext Task*: an U-shaped model is trained to recognise all the regions containing any lesion attributes through an *attribute agnostic mask*. *Downstream Task*: Trained parameters from the *Pretext Task* are transferred to the downstream tasks of locating and identifying each of the lesion attributes independently. The image is adopted from [22] and modified as per fundus image lesion feature generation task.

performing bitwise OR operator as:

$$M_U(x) \triangleq z_{MA} | z_{HE} | z_{SE} | z_{EX}. \quad (4.3)$$

where $|$ denotes the *bitwise OR* operator.

Using this intermediately constructed dataset D_U and equation (4.1), we train S_U so that it can detect lesion attribute regions belonging to any of the lesions in L . We define this part as a *Pretext Task* (left part in figure 4.3).

Next, we define four separate segmentation networks ($S_l, l \in L$) to segment four different lesion attributes. The architecture of S_l is similar to the S_U network, i.e., $S_l = \{S_l^e, S_l^d\}$, $l \in L$, and we initialize the encoder and decoder of S_l as:

$$S_l^e \leftarrow S_U^e; S_l^d \leftarrow S_U^d, \forall l \in L \quad (4.4)$$

These networks are then fine-tuned independently using pixel-level annotations z_l of each image x in the source domain X_s to generate multi-lesion information (downstream task in figure 4.3). The loss function in equation (4.1) seeking parameters θ_{S_l} for each network S_l given these supervised data.

Note that, instead of choosing four different models for each of the lesion attributes, a single model can also be used to minimize a semantic segmentation task, to predict multi-class outputs. However, in our case, a pixel could represent more than one lesion class [101]; therefore, a single semantic model with multi-class settings breaks down this property. In the experiment (table 5.1, 5.2), we found that using TATL transfer learning helps improve overall accuracy compared with the conventional approach initialized from the pre-trained ImageNet. We argue that this gain arises from learning shared feature representation of TATL in the Pretext-Task step, thereby allowing knowledge sharing among attribute models.

4.2.2 Adversarial Learning on Predicted Segmentation Maps in Source Domain

In order to improve semantic segmentation accuracy, especially with lesion attributes whose regions are tiny, we follow prior studies in semantic segmentation of natural images [11, 102] and adapt for each lesion segmentation model S_l a corresponding discriminate network. Precisely, the Generative Adversarial Networks (GANs) [32] consist of a generative network and a discriminate network, which play in a competitive min-max. In our setting, we choose the generative network as the lesion attribute generator S_l , $l \in L$ and employ the discriminate network as the conditional GAN [103], denoted as Patch Discriminator *PD-Net*. The *PD-Net* aims to distinguish *real* samples from the generated ones. The architecture of *PD-Net* is similar to [101], which is formed based on the ideas of PatchGAN [104].

Figure 4.4 describes our workflow for training a specific lesion feature network S_l using the adversarial learning with *PD-Net*. In particular, an input image is split into 16×16 smaller patches, and each of these patches is applied with the cross-entropy loss to decide whether that patch is fake or real. The input for the *PD-Net* is the concatenation of the original image patch with its corresponding lesion map predicted from S_l and actual ground truths. Thus, the discriminator *PD-Net* learns the joint distribution of both images and the lesion map, conditioned on the input data. In other words, the *PD-Net* will force the output of S_l to look ‘real’ as the ground-truth data as much as possible given the input image $x \in X_s$. The objective loss function for this formulation can be defined as:

$$\mathcal{L}_{\text{patch}} = \min_{\theta_{S_l}} \max_{\theta_{\text{pd}}} \frac{1}{|X_s|} \sum_{(x, z_l) \in X_s} [\log(\text{PD}(x \oplus z_l)) + \log(1 - (\text{PD}(x \oplus \hat{z}_l)))] \quad (4.5)$$

where \oplus being the concatenation operator and $\hat{z}_l = S_l(x)$. Combining this adversarial objective with Eq. (4.1), we derive an optimization problem for the segmentation model S_l using the source domain X_s as:

$$\mathcal{L}_{\text{source}} = \mathcal{L}_{\text{seg}} + \lambda_p \mathcal{L}_{\text{patch}}. \quad (4.6)$$

where λ_p is a parameter controls the contribution of $\mathcal{L}_{\text{patch}}$.

4.2.3 Incorporating Domain Adaptation with Unlabeled Data in Target Domain

So far, we have trained the segmentation model S_l in the source domain X_s . However, in the target domain, we can not use Eq. (4.1) to train S_l because the pixel-level annotations z_l for image samples $x \in X_t$ are not available. Prior approaches [38, 105, 106] proposed self-training approaches in which the predictions of models trained in the source domain are used to infer new images in the target domain. Then images with high confidence scores are utilized to fine-tune the model for an adaptation step.

Unlike prior works, we propose constraining the objective function of the lesion attribute generation model S_l so that its embedding feature representations estimated from images in the source and target domain are close to each other. We realize this idea by minimizing Wasserstein distance [31, 107] in the encoder layers of S_l^e and further applying an adversarial learning mechanism on entropy segmentation maps [12, 108, 109] computed by S_l for images in different domains (figure 4.5). Compared with prior methods on DR-Grading or lesion attribute segmentation [2, 7, 8, 29, 37, 38], we are the first to consider

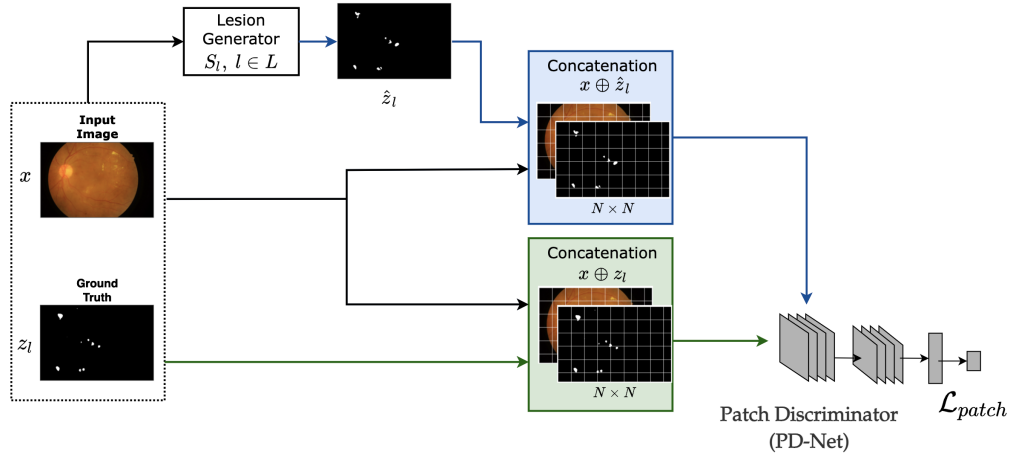


Figure 4.4: Our workflow for the supervised lesion segmentation using adversarial learning. Output from the segmentation network S_l is concatenated with the input image and split into patches. Then, the patch discriminator $PD-Net$ computes the adversarial loss by predicting whether each patch is real or fake.

domain adaptation aspects through aligning feature representations across domains using tools from Wasserstein distance estimation and adversarial entropy minimization.

4.2.3.1 Wasserstein Distance Minimization on Feature Encoder

For each lesion segmentation network S_l , we aim to minimize discrepancies between different domain's feature embedding estimated at the encoder layers S_l^e of S_l (section 4.2.1). To this end, we introduce the domain critic $W-Net$ parameterized by θ_w , whose goal is to estimate the Wasserstein distance [31] between the source and target distribution in the feature representation space (figure 4.5). Given an encoder feature representation $h = S_l^e(x)$ for an image x from any domain, we define:

$$h_s = S_l^e(x_s), x_s \in X_s, \quad (4.7)$$

$$h_t = S_l^e(x_t), x_t \in X_t. \quad (4.8)$$

and the domain critic function $W(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}$ which maps a feature representation to a real number. As proposed by [28], if the parameterized family of domain critic function $W(\cdot)$ are all 1-Lipschitz, for the source and target distributions X_s and X_t , an empirical Wasserstein distance can be approximated by *maximising domain critic loss* \mathcal{L}_{wd} with respect to θ_w :

$$\mathcal{L}_{wd}(X_s, X_t) = \frac{1}{|X_s|} \sum_{x_s \in X_s} W(h_s) - \frac{1}{|X_t|} \sum_{x_t \in X_t} W(h_t). \quad (4.9)$$

To make the training progress to be stable, we further enforce parameters θ_w to *minimize the Lipschitz constraint* using gradient penalty \mathcal{L}_{grad} proposed by [110] of the domain critic as:

$$\mathcal{L}_{grad}(\hat{h}) = \left(\left\| \nabla_{\hat{h}} W(\hat{h}) \right\|_2 - 1 \right)^2, \quad (4.10)$$

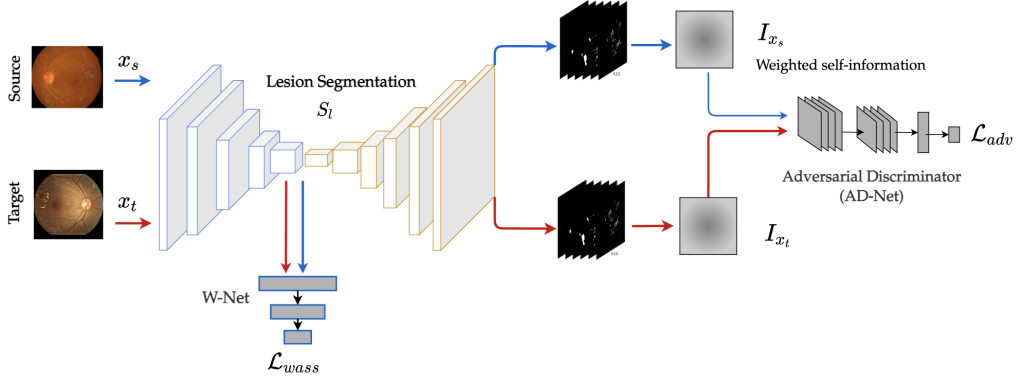


Figure 4.5: Domain Adaptation Components in our method with *W-Net* and *AD-Net*. The *W-Net* is designed to minimize the Wasserstein Distance Minimization between two feature embedding spaces in the source and target domain. The *AD-Net* is employed as the discriminate network to distinguish entropy distributions from two domains.

where $\hat{h} = \{h_s, h_t, h\}$ is the feature representations at which to penalize the gradients and it is defined by the source and target representations h_s, h_t as well as at the random points h along the straight line between source and target representation pairs [28, 110]. Combining Eq. (4.9) and Eq. (4.10), the optimization problem for the *W-Net* and the lesion attribute segmenter S_l ($l \in L$), which minimizes discrepancy between two domains X_s and X_t in terms of Wasserstein distance, is defined as:

$$\mathcal{L}_{wass} = \min_{\theta_{S_l}} \max_{\theta_w} \{\mathcal{L}_{wd} - \eta \mathcal{L}_{grad}\}, \quad (4.11)$$

where η is the regularization coefficient.

4.2.3.2 Adversarial Entropy Minimization on Target Domain

In order to exploit the structural consistency between the source and target domain, we employ an entropy loss \mathcal{L}_{ent} to directly minimize an uncertainty prediction [12, 109, 111] in the target domain of S_l . Given an input image in the target domain $x_t \in X_t$, we compose the Shannon Entropy map [108] $E_{x_t} \in [0, 1]^{h \times w}$:

$$E_{x_t}^{(h,w)} = \frac{-1}{\log(C)} \sum_{i=1}^C P_{x_t}^{(h,w,i)} \log P_{x_t}^{(h,w,i)}, \quad (4.12)$$

at each pixel (h, w) . Here, C being the number of output class, i.e., $C = 2$ in our setting with binary segmentation for each lesion attribute $l \in L$, and $P_{x_t}^{(h,w,i)}$ is the pixel-wise predicted class score estimated from $S_l(x_t)$. We now define an entropy loss \mathcal{L}_{ent} which is the sum of all pixel-wise normalized entropy:

$$\mathcal{L}_{ent}(x_t) = \sum_{h,w} E_{x_t}^{(h,w)} \quad (4.13)$$

By combining Eq.(4.1) and Eq. (4.13), we can jointly optimize the supervised segmentation loss with the samples in the source domain X_s and an unsupervised entropy loss on

the target domain X_t by:

$$\min_{\theta_{S_l}} \frac{1}{|X_s|} \sum_{(x, z_l) \in X_s} \mathcal{L}_{wbce}(S_l(x), z_l) + \frac{\lambda_{ent}}{|X_t|} \sum_{x_t \in X_t} \mathcal{L}_{ent}(x_t), \quad (4.14)$$

with λ_{ent} is the weight factor for the entropy part \mathcal{L}_{ent} .

Training the lesion segmentation model S_l with the joint loss function in Eq. (4.14) has shown improvement in experiments. However, merely minimization this does not entirely capture the structural dependencies between the local semantics. In other directions, authors in [112] have argued that adaptation to the structural output space is favorable for unsupervised domain adaptation in semantic segmentation tasks. Similarly, Shen et al. [28] have also shown that adaptation in the latent space can enhance the model generalization ability. These premises are based on the fact that the source and target domain usually share strong similarities in the semantic layout. To exploit such observations, we incorporate an adversarial training framework to implicitly guide the target domain's entropy distributions to be similar to the source ones [12, 109]. Our motivation is based on the fact that a trained model naturally produces a high confidence score for one target class and low for the rest on source-like images. Therefore, entropy for the source-like images will be low, and that of target images will be higher.

Using the equation (4.12), we define a *weighted self-information map* I_x for any input image x composed of pixel-level vectors at each pixel (h, w) as:

$$I_x^{(h,w)} = -P_x^{(h,w)} \circ \log P_x(h, w) \quad (4.15)$$

where \circ stands for hadamard product, $P_x^{(h,w)} = \left[P_x^{(h,w,i)} \right]_{i=1}^C$ is the probability score for C classes at (h, w) estimated by $S_l(x)$.

Given this, we define I_{x_s} , I_{x_t} as weighted self-information maps for the source and target domain X_s, X_t respectively computed by:

$$I_{x_s}^{(h,w)} = -P_{x_s}^{(h,w)} \circ \log P_{x_s}(h, w) \quad (4.16)$$

$$I_{x_t}^{(h,w)} = -P_{x_t}^{(h,w)} \circ \log P_{x_t}(h, w) \quad (4.17)$$

We now formulate the adversarial network *AD-Net* with parameters θ_{ad} as convolutional networks. This network takes I_x as an input and produces domain classification as an output with class label 1/0 for the source/target domain respectively (right side in figure 4.5). Note that, *AD-Net* architecture and its parameters are different from the discriminator network *PD-Net* that we have used to optimised S_l in Eq. (4.6). Similar to the learning procedure of the original GAN method [32], the discriminator *AD-Net* is trained to discriminate outputs coming from source and target images and simultaneously, the lesion network S_l is trained to fool the discriminator *AD-Net*. The optimization objective of the discriminator is:

$$\mathcal{L}_{adv} = \min_{\theta_{S_l}} \max_{\theta_{ad}} \frac{1}{|X_s|} \sum_{x_s \in X_s} \log(AD(I_{x_s})) + \frac{1}{|X_t|} \sum_{x_t \in X_t} \log(1 - AD(I_{x_t})) \quad (4.18)$$

In summary, by jointly optimizing equations (4.6), (4.11), and (4.18), we end up with a total loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_p \mathcal{L}_{patch} + \lambda_w \mathcal{L}_{wass} + \lambda_{adv} \mathcal{L}_{adv}. \quad (4.19)$$

where, λ_p , λ_w and λ_{adv} are the weighting factors for the patch-based adversarial learning, Wassertein term, and entropy-based adversarial term respectively.

4.3 Integrating Lesion Features into DR Grading Networks

In DR grading task, conventionally, human experts diagnose by observing lesions and attributing signs of illness in detail to grade a disease’s severity. While deep neural classification models [29, 113, 114] based on image-level supervision can achieve promising performance, we follow lesion integration-based strategies [4, 7, 38]. We hypothesize that such an approach can improve the model’s generalizations and make the leaned network predictions transparent to the end-users by taking into account medical priors. To this end, using lesion segmentation models S_l from the previous step, we generate L distinctive lesion maps $\hat{z}_{l=1}^L$ for each input image x . The attention module *Attn-Net* then uses this preliminary information to select the most relevant parts contributing toward the better performance of the disease classification network *G-Net*.

This work investigates two different jointly learning frameworks between *Att-Net* and *G-Net* for network architectures based on CNN (section 4.3.1) and Transformer (section 4.3.2). Furthermore, besides integrating lesion information at a feature level through attention gates, denoted as *low-level concepts* (section 4.3.1.1), we also formulate a novel overlapping constraint among heatmap regions of *G-Net* and segmented multi-lesion regions obtained from *S-Net*, denoted as *high-level concepts* (section 4.3.1.2). Figure 4.6 gives an illustration for these attention concepts. In our setting, we develop these constraints for both CNN- and Transformer-based architectures and discover that they contribute in improving accuracy for several baselines by a large margin (tables 5.6, 5.5). Last but not least, these strategies enable an explainable visualization of DR grading predictions when ophthalmologists can observe the correlation of high-responding areas in trained networks with medical priors represented as lesion regions.

4.3.1 Integrating Lesion Features for CNN-based Methods

4.3.1.1 Attention Lesion Regions at Low-level Concepts

Inspired by [38], we consider attention mechanisms at feature maps of the classification network *G-Net* given multi-lesion maps predicted by S_l . For this, feature maps at the first and last layers of *G-Net* are jointly combined with lesion maps $\hat{z}_{l=1}^L$ to define attention maps that return high responses to different lesion regions characterizing the disease. It is worthy to note that we do not integrate the lesion masks initially predicted by S_l as a direct input for the classification model *G-Net* because the initially predicted masks are usually very tiny and sparse. Their contributions rather than being controlled by the attention network *Att-Net*. In our setting, we choose the ResNet-50 [97], including five CNN blocks for the architecture of *G-Net*, and therefore have to modify formulations in [38] as follows.

We indicate five CNN blocks of *G-Net* as $G = \{\mathbf{g}^1, \mathbf{g}^2, \mathbf{g}^3, \mathbf{g}^4, \mathbf{g}^5\}$. The feature maps of an image x at the first block are computed by:

$$\mathbf{f}^{\text{first}} = \mathbf{g}^1(x) \quad (4.20)$$

Similarly, the feature maps at the last layer estimated by:

$$\mathbf{f}^{\text{last}} = \mathbf{g}^5(h^5) \quad (4.21)$$

where

$$h^i = \mathbf{g}^{i-1}(h^{i-1}), i \in [2, 5] \quad (4.22)$$

$$h^1 = x \quad (4.23)$$

In the first step, we train *G-Net* with all blocks in a fully supervised manner using the image-level annotations in the target domain $(x, y) \in X_t$. The optimization problem is:

$$\mathcal{L}_{cls} = \min_{\theta_g} \frac{1}{|X_t|} \sum_{(x,y) \in X_t} \mathcal{L}_{mce}(G(x), y), \quad (4.24)$$

where \mathcal{L}_{mce} is the multi-class cross entropy classification loss defined as:

$$\mathcal{L}_{mce}(\hat{y}, y) = - \sum_{y=k}^K y^{(k)} \log \hat{y}^{(k)} \quad (4.25)$$

with $y^{(k)}$ is 0 or 1, indicating whether class label k is the correct classification.

Once the model is pre-trained, the feature maps $\mathbf{f}^{\text{first}}$ and \mathbf{f}^{last} are computed using equations (4.20), (4.21). Then we define an attentive feature for the l -th lesion \hat{z}_l obtained from S_l by:

$$\mathbf{f}_l^{\text{first-att}} = \text{ReLU}(\mathbf{W}_l^{\text{first}} \text{concat}(\hat{z}_l, \mathbf{f}^{\text{first}}) + b_l^{\text{first}}), \quad (4.26)$$

where $\text{concat}(\cdot)$ is the channel-wise concatenation; $\mathbf{W}_l^{\text{first}}$ and b_l^{first} are additional learnable parameters and bias terms for the l -th lesion.

In a next step, the last feature maps \mathbf{f}^{last} acting as a global feature embedding is correlated with the first-level attentive features to generate attention weights for the l -th lesion:

$$\alpha_l = \text{Sigmoid}(\mathbf{W}_l^{\text{last}} [\mathbf{f}_l^{\text{first-att}} \odot \mathbf{f}^{\text{last}}] + b_l^{\text{last}}), \quad (4.27)$$

where \odot is the element-wise multiplication; $\mathbf{W}_l^{\text{last}}$ and b_l^{last} are other parameters and bias terms to learn attention features at the global level. To make \mathbf{f}^{last} and $\mathbf{f}_l^{\text{first-att}}$ be compatible in channel dimensions, we also use a 1×1 convolution over the \mathbf{f}^{last} .

By applying Eq. (4.27) for each lesion $l \in L$, we aggregate all attention lesion maps and use them to separately conduct element-wise multiplication with the first-level features $\mathbf{f}^{\text{first}}$ of *G-Net*. The output feature vectors then are concatenated and utilized as final attention features to fine-tune the DR Grading network *G-Net*. In terms of optimization, parameters of *Att-Net* include $\theta_{Att} = \{(\mathbf{W}_l^{\text{first}}, \mathbf{W}_l^{\text{last}}, b_l^{\text{first}}, b_l^{\text{last}})_{l=1}^L\}$.

Finally we compose an optimization, that jointly learns the grading network *G-Net* and the attention-based lesion *Att-Net* at the low-level concept as:

$$\mathcal{L}_{cls}^{att} = \min_{\theta_g, \theta_{att}} \frac{1}{|X_t|} \sum_{(x,y) \in X_t} \mathcal{L}_{mec}(G(x) \cdot \text{Att}(S_l(x)_{l=1}^L), y). \quad (4.28)$$

where \mathcal{L}_{mec} is the multi-class cross entropy defined in Eq. (4.25). We characterize the loss \mathcal{L}_{cls}^{att} as the low-level concept because it purely combines *G-Net* and *Att-Net* at the feature level through attention weights. While this approach boosts performance, it is still a black box to end-users. In the next section, we introduce a new constraint between two types of networks which are able to enhance accuracy and provide explainable properties for the learned system.

4.3.1.2 Attention Lesion Regions at High-level Concepts

Based on the Grad-CAM [115] discussed in the section 3.3.2.2, we can get the class activation map of the last layer in the grading model *G-Net*. For a input image x , let $f_{l,k}$ be the activation of unit k in l -th layer. For each of the ground-truth class c , we

compute the corresponding gradient score s^c , with respect to activation maps of $f_{l,k}$. These gradient scores then pass through 1×1 convolution of global average pooling layer to obtain the neuron importance weights $w_{l,k}^c$:

$$w_{l,k}^c = \text{GAP} \left(\frac{\partial s^c}{\partial f_{l,k}} \right), \quad (4.29)$$

where $\text{GAP}(\cdot)$ is the global average pooling. Because $w_{l,k}^c$ indicates the important of activation map $f_{l,k}$ contributing the prediction of class c , we thus apply the weight matrix w_c as a kernel and apply 2D convolution over the feature map f_l to aggregate all activation maps, followed by a ReLU function to get the activation map AM^c for the c -th class:

$$AM^c = \text{ReLU}(\text{conv}(f_l, w^c)) \quad (4.30)$$

where l represents the last convolution layer.

In the next step, we normalize the AM^c so that its class channel values are normalized to $[0, 1]$, denoted as \widetilde{AM}^c , using a thresholding operation $T(\cdot)$ [116] as follows:

$$\widetilde{AM}^c = T(AM^c) \quad (4.31)$$

$$T(AM^c) = \frac{1}{1 + \exp(-\omega(AM^c - \sigma))} \quad (4.32)$$

where σ is the threshold matrix whose elements are equal to σ . ω is the scale parameter forcing $T(AM^c)_{i,j}$ approximately equals to 1 if AM_{ij}^c is greater than σ , or to 0 otherwise.

Finally, we propose an overlapping loss function $\mathcal{L}_{overlap}$ for images whose lesion areas are not empty as:

$$\mathcal{L}_{overlap} = \min_{\theta_g, \theta_{att}} \frac{1}{hw} \|\widetilde{AM}^c - L_U\|_2 \quad (4.33)$$

where h, w are the height and width of image x , L_U is the union region of all of lesion types computed by:

$$L_U = \bigcup_{l \in L} S_l(x) \quad (4.34)$$

Note that we only compute the $\mathcal{L}_{overlap}$ for an image x if its grading label is different 0, i.e., input image has some stages of DR disease. By optimizing $\mathcal{L}_{overlap}$, we jointly learn parameters for both grading network *G-Net* and the attentive network *Att-Net*.

Combining equations (4.33), (4.28), we end up with a new total object loss function that incorporates attention mechanisms for DR grading task at both low-level and high-level constraints:

$$\mathcal{L}_{grading} = \mathcal{L}_{cls}^{att} + \mathcal{L}_{overlap} \quad (4.35)$$

At a glance, our proposed loss $\mathcal{L}_{overlap}$ is comparable to existing semi-supervised learning [116] or Covid-19 detection [48]; however, we extend it for the multi-lesion scenario in the context of the DR grading task. Furthermore, our system is superior to these works and current approaches to the DR problem [29, 38] in that we develop lesion information-based attention mechanisms for classification tasks at both feature-level and high-level concepts, thereby improving performance and providing explainable properties to the entire system (table 5.5).

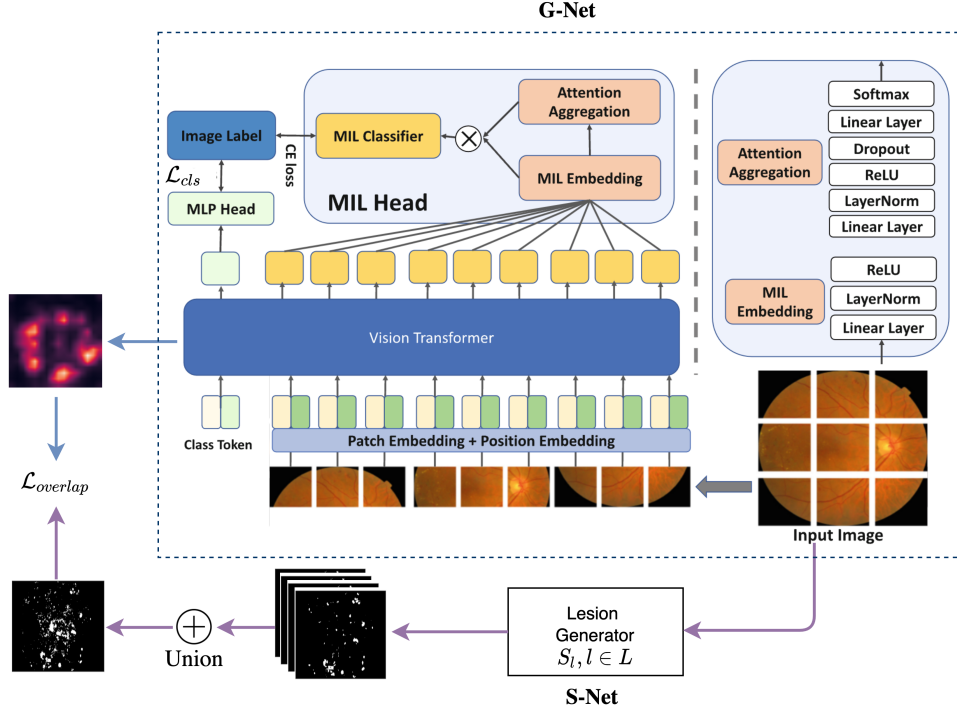


Figure 4.7: Our proposed DR-grading transformer architecture. We integrate predicted lesion feature maps using the **lesion generation** S_l for an image with the multi-head lesion generation G-Net based on MIT-VT [117]. We compute $\mathcal{L}_{overlap}$ and the grading model is trained using both \mathcal{L}_{cls} and $\mathcal{L}_{overlap}$.

4.3.2 Integrating Lesion Features for Transformer-based Methods

Transformer models have been widely used for natural language processing (NLP) and achieved superior performance [118]. These models use inherent self-attention techniques automatically learn to focus on the task-specific important regions during the training process. Recently, transformer-based models such as ViT [64] have been exploited in vision-related tasks, and they have achieved competitive performance against the CNN. The main difference between CNN and transformer models is that CNN uses pixel arrays whereas vision transformers split the images into visual tokens. In Transformer, an input image $x \in \mathbb{R}^{h \times w \times 3}$ is divided into individual patches with size $p \times p$. Therefore, we can obtain $N = hw/p^2$ patches from one single image. The patches are further flattened into a 1D format and then get embedded together using a linear layer of D dimensions to make a compatible data format for the transformer networks.

In this section, we investigate the performance of Transformer architecture for the DR grading task. For this, we choose the MIL-VT method proposed by Yu et al. [117] (figure 4.7). Compared with ViT, MIL-VT further adds multiple-instance learning heads to leverage the features extracted from individual patches. However, MIL-VT solely employs image-level data in the training stage and ignores the responsibilities of lesion areas. This motivates us to incorporate the MIL-VT attention’s mechanism with our overlapping heatmap concepts presented in section 4.3.1.2. To establish such a constraint, we apply the Attention Rollout technique mentioned in recent Transformer papers

[64, 119] and use it to compute the heatmap regions for MIL-VT.

In particular, Attention rollout is a concept used to track the information propagated from the input layer to the embeddings in the higher layers. Given a Transformer with N layers, in each layer $n \in N$, this technique takes the average of all attention weights across all heads to form an attention matrix A_n where $(A_{ij})_l$ defines how much attention is going to flow from token j in the layer $n - 1$ to token i in the layer n . Then the attention rollout matrix at layer n , denoted as \tilde{A}_n , is computed in a recursive way:

$$\tilde{A}_n = (A_n + I) \tilde{A}_{n-1} \quad (4.36)$$

where I is the identity matrix.

By computing equation (4.36) at the last layer N , we can account for the combination of attention across tokens through all layers. In our method, we choose the attention map AM^c as:

$$AM^c = \tilde{A}_N. \quad (4.37)$$

Then the loss $\mathcal{L}_{overlap}$ in equation (4.33) is defined in analogous way (figure 4.7). In experiment, we discover that this extended version significantly enhances MIL-VT's performance in a variety of situations (table 5.6).

4.4 Human Interaction with Trained Systems

Our proposed architecture is essentially a joint disease diagnosis system. Given a retina fundus image, the system can automatically detect the associated lesion attributes and predict its disease grade. In addition, the detected lesion masks from the segmentation module, as-they-are, can work as a support for the disease grading prediction by the grading module. Unlike most of the medical decision support systems [53, 55?] where the lesion maps and disease grading predictions modules are independent of each other, our architecture is trained to learn these tasks collaboratively. In the learning phase of our grading model, we exploit predicted lesion information. Besides, the system also facilitates the utilization of expert feedback throughout the training process. Generally, our diagnosis system provides an interactive way with the expert user, as shown in 4.8. For this, the method can deliver a visual explanation to the user about its predictions. Also, the method is robust enough to incorporate user feedback in weakly-supervised annotations and can use them to fine-tune the model. We will discuss in the following sections detailing the interaction process.

4.4.1 DR Grading Predictions with Explainable Properties

Explainability is an imperative feature required for intelligent decision support systems, especially in the healthcare domain. When a healthcare model predicts a disease, the medical practitioner needs to know which factors the model is taking into account. In our work, we focus on the local interpretability [46, 120] of our prediction model computed based on the class activation map AM^c . This gives us the discriminative regions used by the model to predict the disease grade for a certain class. By drawing the AM^c over input image x (figure 4.8 - *Visual Explanation*), we can observe the highlighted region on the input x , which the grading network G -Net thought to be the most essential region for its decision (figure 4.8 - *Predicted DR Grade*). Moreover, detected lesion maps can be used

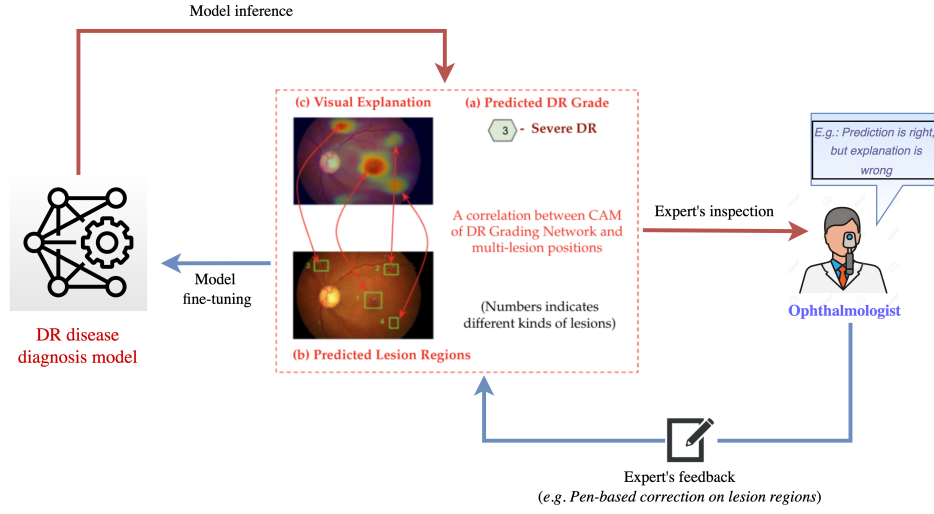


Figure 4.8: Illustration for the interaction between experts and the trained model. The model can infer explanatory predictions via presenting disease grade, related lesions, and class activation mapping. Upon inspection, the expert validates the results and, if necessary, provides input (e.g., drawing bounding box on missing lesion regions). This input is used to improve model performance in subsequent model retraining.

as visual explanations. In particular, considering the overlapping region between lesion maps $\hat{z}_{l=1}^L$ (figure 4.8 - *Predicted Lesion Regions*) with activation maps AM^c of x gives a visual interpretation of how much these lesions influenced the grading model prediction (figure 4.8).

4.4.2 Improving System's Performance through User Feedback

In general, user feedback is the information that a user sends to a learning agent in order to update the agent's knowledge. In interactive machine learning, user feedback can guide the intelligent system to achieve the desired behavior [49]. Our proposed method is inherently designed to be able to integrate user feedback in its learning procedure. However, getting user feedback is a costly and time-consuming task, especially in the medical domain. For instance, lesion regions in diabetic disease can be very tiny to cover only a small group of pixels in the image. Fortunately, our framework alleviates this issue when we do not require pixel-level training data in the target domain to train segmentation models. It instead is handled by the domain adaptive networks as the Wasserstein network *W-Net* (Section 4.2.3.1) and the Adversarial Discriminator-based on Entropy *AD-Net* (Section 4.2.3.2). Given this, we can generate different lesion regions while only using pixel-level labels from another source domain.

When the system is deployed in practice, the expert can provide two feedback forms. Firstly, given new annotations on lesion masks, we can fine-tune our lesion generation models *S-Net*. Specifically, for each lesion generator model S_l , we update the segmenta-

tion model using *new labeled data in the target domain* by training:

$$\mathcal{L}_{seg} = \min_{\theta_{S_l}} \frac{1}{|X_t|} \sum_{(x, z_l) \in X_t} \mathcal{L}_{wbce}(S_l(x), z_l) \quad (4.38)$$

where \mathcal{L}_{wbce} is the weighted binary cross-entropy loss defined in equation (4.2). Note that in this case, we can ignore the domain adaptation parts, which are already optimized during training with data in the source domain X_s .

Secondly, when users provide both new image grading and lesion segmentation labels, we can update either:

- Segmentation models *S-Net* using \mathcal{L}_{seg} in equation (4.38).
- Attention-based disease grading model *Att-Net* and *G-Net* by learning the objective function $\mathcal{L}_{grading}$ defined in equation (4.35).

It is worth noting that our technique automatically learns to identify critical lesion locations influencing the DR Grading task by utilizing attention mechanisms; hence, it can be resilient to a certain degree of noise in segmentation annotations provided by experts. In other words, rather than a precise segmentation mask, the expert can mark the region of interest in the form of bounding boxes or circles around the lesion locations using any pen-based input device. As the attention model inherently learns to filter out the noise and focus on the area of interest, our experiments confirm that only these soft annotations are sufficient to boost the performance of the model (table 5.8).

Chapter 5

Experiments and Results

5.1 Data Description

Retina images are generally captured in two forms, which are Optical Coherence Tomography (OCT) capturing a cross-sectional images of retina and Color Fundus retinal photography capturing 3D retina images using fundus cameras. We employed datasets with color fundus retina images for all our experiments in this work. For domain-invariant lesion attributes segmentation, we used two publicly available datasets. To the best of our knowledge, these two are the only datasets that provide pixel-level annotations for diabetic retinopathy. The datasets are:

- *IDRID segmentation* [21]: This dataset contains 81 high resolution (4288×2800 pixels) images with four different types of lesion annotations and is split into 54 training images and 27 testing images. The lesions are: microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE). Each of these lesions are annotated in the forms of binary masks.
- *FGADR segmentation* [8]: Similarly as IDRID, this dataset contains 1843 images with 4 kinds of lesion masks (MA, HE, EX, and SE). The training set consists of 1500 images and the remaining 343 images are for test set. This dataset has imbalance in the lesion classes. The class HE is available in 1471 images which is 78% of the total images where the class MA is available for only 610 images which consists of 33% of the total images.

For the classification task, we used three publicly available datasets.

- *IDRID classification* [21]: This dataset consists of 413 training and 103 test images with 5 severity grading labels. The five labels include: *normal*, *mild*, *moderate*, *severe* and *proliferative*, which are annotated as 0, 1, 2, 3, 4, respectively. These images only have image-level labels and do not have any pixel-level lesion masks annotations.

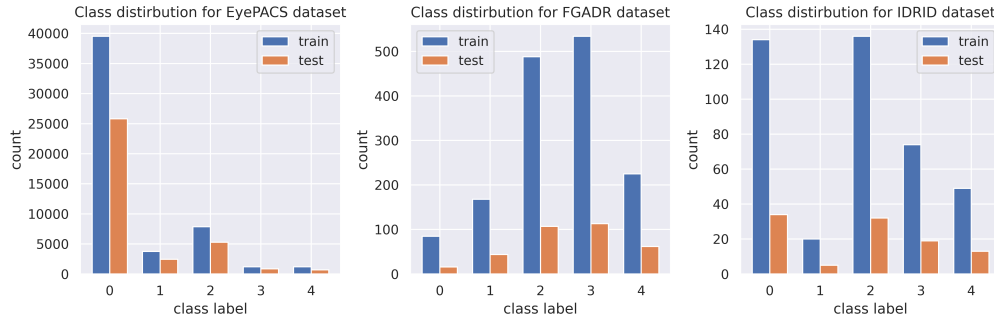


Figure 5.1: Class distributions in the three classification datasets used in our work.

- *EyePACS* [121]: This dataset consists of 35,126 training images and 53,576 testing images with similar grading protocol of classes with 5 categories. Images in this dataset is captured using different types of cameras settings under various light conditions.
- *FGADR classification* [8]: this dataset consists of 1500 training images and 343 testing image for diabetic retinopathy grading. This is the only one that contains both image-level pixel disease grading and pixel-level lesion annotation maps.

Figure 5.1 shows the class distribution for the three classification datasets we used in our work. We can observe that there are great numbers of class imbalance in all three datasets.

5.2 Evaluation Metric

To evaluate the segmentation performance in section 4.2, we used two common metrics. Area Under the Curve of Receiver Operating Characteristics (AUC-ROC) and Area Under the Curve of Precision-Recall (AUC-PR). The positive class means the existence of lesion and negative is the otherwise.

- *AUC-ROC* measures the class separability at various threshold settings. ROC is the probability curve and AUC represents the degree of measures of separability. It compares true positive rate (sensitivity/recall) versus the false positive rate (1 - specificity). The higher the AUC-ROC, the bigger the distinction between the true positive and false negative.
- *AUC-PR*: It combines the precision and recall, for various threshold values, it compares the positively predicted value (precision) vs the true positive rate (recall). Both precision and recall focus on the positive class (the lesion) and unconcerned about the true negative (not a lesion, which is the majority class). Thus, for class imbalance, PR is more suitable than ROC. The higher the AUC-PR, the better the model performance.

For evaluating multi-class disease grading classification task in section 4.3, we used two evaluation metrics.

- *Accuracy*: The normal classification accuracy which simply measures how many observations are correctly classified.
- *Quadratic Weighted Kappa (Kappa)*: It is similar to the Cohen’s kappa metric [122] by the weights are set to ‘Quadratic’. Cohen’s kappa measures the agreement between two raters who each classify N number of items into C mutually exclusive categories. The formulation is:

$$\text{kappa} = \frac{p_o - p_e}{1 - p_e}, \quad (5.1)$$

where p_o denotes the relative observed agreement between the raters and p_e is the hypothetical probability of chance agreement.

5.3 Implementation Details

We implemented all experiments using PyTorch framework [123] and executed on a computer 2 NVIDIA TITAN RTX 24GB GPUs. For all experiments, the sizes of the input images are 512×512 .

Domain Invariant Lesion Attribute Learning

For the experiments in lesion attribute segmentation discussed in section 4.2, we used pixel values in the range $[0, 1]$ for retina images and ground-truth binary lesion segmentation masks. In this step, the input image are labeled images from source domain and unlabeled images for the target domain. We applied following preprocessing steps for input images as proposed in [101]:

- *Contrast limited adaptive histogram equalization (CLAHE)*: Instead of processing the entire image, CLAHE processes smaller regions. We applied the CLAHE technique with to 8×8 patches.
- *Denoising*: Assuming that the images contain Gaussian white noise, we applied Non-local means denoising algorithm [124] with a filter strength of 10.

Moreover, we applied the built-in data augmentation procedure of PyTorch framework. Each image and its corresponding lesion masks jointly are randomly crop to 512×512 pixel and randomly rotated with a maximum angle of 30° . Finally, we applied normalization to each channel of the input lesion image with mean = $[0.485, 0.456, 0.406]$ and std = $[0.229, 0.224, 0.225]$.

Our *S-Net* model described in section 4.2 is a U-shaped [15] encoder-decoder segmentation network. The encoder architecture is identical to the convolutional block of the ResNet50 [97] model and the decoder architecture is up-scaled accordingly. The architecture of the *PD-Net* described in section 4.2.2 is similar to the InfoGAN [125]. This discriminator model provides stable performance for our case. It consists of two convolution layers with 64 and 128 kernels, respectively, followed by two fully-connected layer of 1024 dimensions and then the sigmoid layer.

For training the *SD-Net* and the *PD-net* described in 4.2.1 and 4.2.2, we used SGD with momentum [126] as optimizer with initial learning rate of 10^{-4} and momentum value of 0.9. We adopted the cyclic learning scheduler [127] with max value of 10^{-2} . The

class balancing hyper-parameter weight β for the segmentation loss \mathcal{L}_{wbce} described in equation 4.2 is set to 10 following [101] to address class imbalance for lesion and non-lesion class. The value of λ_p for the source domain adversarial joint optimization in equation (4.6) is set to 10^{-2} . For domain adaptation fine-tuning training discussed in section. 4.2.3, pretrained *S-Net* from 4.2.2 is trained with Wasserstein critic model *W-Net* (section 4.2.3.1) and entropy-based discriminatory network *AD-Net*. *W-Net* is a fully connected neural network which takes 1D encoder block outputs from *S-Net* as inputs. *W-Net* has two fully connected layers with number of neurons of 128 and 64, respectively, and final output is linear layer which gives a single scalar value, the critic score. *AD-Net* architecture is adopted from [12] which have five convolutional layers. The input for *AD-Net* is the weighted self-information maps computed on the output of *S-Net* using equation (4.17) and output is the binary class probability score on input being source or target domain.

For training both *W-Net* and *AD-Net*, we used Adam optimizer [128] with a learning rate 10^{-4} . The optimizer hyper-parameters for *S-Net* in this step are kept unchanged and they are similar to the one discuss in the previous step. Following [12], the weight factor λ_{ent} in Eq. (4.14) and λ_{adv} in Eq. (4.19) are 10^{-3} . We used batch size of 8 for our experiments.

Lesion Attentive Grading Model

To train the lesion attentive grading models *G-Net* and *Att-Net* discussed in section 4.3, the input images are 512×512 . We followed the similar pre-processing steps for the input images as with the grading model. We also used data augmentation functions in PyTorch as random-crop, horizontal-flip, vertical-flip, color-distortion, rotation, and translation.

The backbone of our lesion attentive grading model is a ResNet50 [97] architecture with the output fully connected layer modified to predict five classes. The attention module *Att-Net* described in section 4.3.1.1 consists of multiple convolutional layers. This network is optimized using Eq. (4.35) using the cross-entropy loss. As we can see from figure 5.1, there is a significant class imbalance among different types of lesions across all datasets, we thus compute class-weight for each lesion attribute during the optimization step. We used the SGD optimizer [129] with initial learning rate 10^{-3} and applied the cyclic learning strategy as in lesion segmentation steps. For all the experiments, we varied the batch size between 16 and 32. For all of our experiments we have used 10% of training data as validation set and final results are computed on test set.

5.4 Performance of Multi-Lesion Segmentation Task

5.4.1 Influence of Pre-training Steps on Lesion Generator Models

As discussed in section 4.2, our lesion segmentation model *S-Net* is firstly pre-trained using the lesion segmentation labels from the source domain data in a fully supervised manner. To evaluate the effectiveness of task agnostic transfer learning (TATL) ([22]) and the adversarial learning (*PD-Net*) during the pre-training step, we test with three settings:

- *S-Net*: Lesion feature generator segmentation model (*S-Net*) is trained only using

the segmentation loss \mathcal{L}_{seg} in equation 4.1. We do not adapt either TATL strategy or adversarial trainings.

- **S-Net + PD-Net**: We incorporate adversarial learning strategy on source domain as discussed in section 4.2.2. We jointly the *PD-Net* with *S-Net* segmentation model by optimization the formula in equation 4.6.
- **S-Net + PD-Net + TATL**: We apply both *task agnostic transfer learning* (TATL) and adversarial learning with *PD-Net* to learn *S-Net* using data from the source domain.

Table 5.1 and 5.2 compare the results of different settings for *IDRID* and *FGADR* segmentation datasets respectively. The segmentation performance is evaluated by AUC-ROC and AUC-PR values on four different lesions, including microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE). For *IDRID-segmentation* dataset in table 5.1, we can observe slight improvement in performance for settings S-Net + PD-Net compared with original S-Net settings. With the combinations of TATL in settings S-Net + PD-Net + TATL, we observe significant improvements in results for all of the lesion classes. For instance, comparing with S-Net settings, we gained 2.5% and 5.5% on AUC-ROC and AUC-PR scores respectively, averaged over all four lesions. For the lesion class *HE*, our method achieved highest improvement in AUC-PR scores of 9.4%.

We can observe the similar trend in results on the test set of *FGADR-segmentation* dataset in table 5.2. There is an improvement of 8.5% in the AUC-PR scores averaged over all four lesion classes. To summarize from the results, we conclude that combining adversarial training using the *PD-Net* and the TATL strategy provided a better pre-training performance for *S-Net* in the source domain supervised learning.

Lesions	MA		HE		EX		SE	
Methods	ROC	PR	ROC	PR	ROC	PR	ROC	PR
S-Net	0.937	0.44	0.896	0.459	0.931	0.722	0.953	0.583
S-Net + PD-Net	0.932	0.439	0.917	0.481	0.946	0.735	0.965	0.611
S-Net+PD-Net+TATL	0.953	0.456	0.931	0.553	0.961	0.772	0.970	0.643

Table 5.1: Contributions of task agnostic transfer learning TATL (section 4.2.1) and Adversarial training PD-Net (section 4.2.2) in learning lesion segmentation model S-Net in the source domain. Results are evaluated using the training and testing set of *IDRID segmentation*.

Lesions	MA		HE		EX		SE	
Methods	ROC	PR	ROC	PR	ROC	PR	ROC	PR
S-Net	0.901	0.373	0.941	0.611	0.947	0.602	0.927	0.410
S-Net + PD-Net	0.926	0.394	0.955	0.638	0.959	0.667	0.941	0.492
S-Net+PD-Net+TATL	0.937	0.417	0.963	0.652	0.970	0.714	0.954	0.553

Table 5.2: Contributions of task agnostic transfer learning TATL (section 4.2.1) and Adversarial training PD-Net (section 4.2.2) in learning lesion segmentation model S-Net in the source domain. Results are evaluated using the training and testing set of *FGADR segmentation*.

5.4.2 Influence of Domain Adaption on Lesion Generator Models

In this section, we present the results of our domain adaption approaches in different settings and compare with different baselines. As discussed in section 4.2.3, the aim of domain adaptation is to train a neural network using available labeled data from source domain and secure a good accuracy on target domain. We evaluated our approach on IDIRD and FGADR datasets in the forms of *source* \rightarrow *target*:

- *IDRID* \rightarrow *FGADR*: evaluation on *FGADR-segmentation* test set when domain adaptive segmentation model is trained using the *labeled* data from *IDRID-segmentation* dataset and *unlabeled* images from *FGADR-segmentation* dataset.
- *FGADR* \rightarrow *IDRID*: evaluation on *IDRID-segmentation* test set when domain adaptive segmentation model is trained using the *labeled* data from *FGADR-segmentation* dataset and *unlabeled* images from *IDRID-segmentation* dataset.

Target Domain	Lesions	MA		HE		EX		SE	
	Methods	ROC	PR	ROC	PR	ROC	PR	ROC	PR
0%	S-Net	0.752	0.243	0.796	0.280	0.794	0.311	0.728	0.264
	S-Net + Entropy	0.807	0.313	0.843	0.357	0.855	0.407	0.819	0.341
	S-Net + AD-Net	0.841	0.348	0.903	0.448	0.917	0.473	0.902	0.443
	S-Net+AD-Net+W-Net	0.894	0.357	0.911	0.502	0.939	0.538	0.915	0.522
40%	S-Net+AD-Net+W-Net	0.938	0.411	0.953	0.613	0.966	0.682	0.965	0.634
60%	S-Net+AD-Net+W-Net	0.946	0.438	0.969	0.648	0.979	0.728	0.971	0.655
80%	S-Net+AD-Net+W-Net	0.954	0.458	0.973	0.671	0.981	0.732	0.977	0.684
100%	S-Net+AD-Net+W-Net	0.958	0.462	0.979	0.676	0.990	0.739	0.984	0.693
100%	FCN-8s [8]	0.925	0.363	0.962	0.606	0.981	0.686	0.963	0.642
	U-Net [8]	0.927	0.382	0.967	0.643	0.982	0.726	0.977	0.683
	DL-V3+ [8]	0.934	0.364	0.973	0.619	0.981	0.708	0.967	0.659
	Attention U-Net [8]	0.942	0.435	0.974	0.678	0.984	0.731	0.980	0.685

Table 5.3: *IDRID* \rightarrow *FGADR*: Semantic segmentation performance on *FGADR* (*target domain*). Models are trained with labeled data on *IDRID* and increasingly labeled data in the target domain from 0 \rightarrow 100%. **Red** indicates the best results for settings using 0% labeled data from target domain in the training step. **Green** stands for settings using 40% – 60% labeled data in target domain but outperform at least one of baselines trained with 100% data. **Bold** are the best values in all methods.

We report the results for domain adaptation on tables 5.3 and 5.4. Our domain adaptation approach discussed in section 4.2.3 consists of modules for Adversarial Entropy Minimization and Wasserstein based distance minimization between source and target domain. To evaluate the effectiveness of different constraints for our domain adaptive segmentation model, we considered four experiments settings:

1. **S-Net**: The baseline approach for our domain adaptation approaches. *S-Net* is trained on source domain by adopting the pre-training method including PD-Net and TATL. This model is trained using only the equation 4.6. Neither domain adaptation constraints are considered.
2. **S-Net + Entropy**: We apply direct entropy minimization constraint for target domain in the learning process of *S-Net* as discussed in section 4.2.3.2. The entropy loss \mathcal{L}_{ent} in equation (4.13) is used for minimization on the unlabeled data from target domain and model is optimized with the segmentation loss \mathcal{L}_{seg} using equation (4.14).

3. **S-Net + AD-Net** : Instead of using direct entropy minimization \mathcal{L}_{ent} , we use the proposed domain adaptation with minimizing entropy-based adversarial learning *AD-Net*. *S-Net* and *AD-Net* are optimized together by combining adversarial loss \mathcal{L}_{adv} in equation (4.18) with \mathcal{L}_{seg} in equation (4.14).
4. **S-Net + AD-Net + W-Net** : Additional domain critic *W-Net* is added along with *S-Net* and *AD-Net* to minimize the domain gap in feature representation using Wasserstein loss \mathcal{L}_{wass} . This is our proposed setting where all modules are jointly trained using our optimization problem in equation (4.19).

Target Domain	Lesions	MA		HE		EX		SE	
	Methods	ROC	PR	ROC	PR	ROC	PR	ROC	PR
0%	S-Net	0.813	0.232	0.803	0.251	0.837	0.363	0.783	0.241
	S-Net + Entropy	0.884	0.331	0.874	0.375	0.902	0.531	0.865	0.404
	S-Net + AD-Net	0.925	0.408	0.902	0.447	0.911	0.697	0.879	0.447
	S-Net+AD-Net+W-Net	0.947	0.436	0.911	0.462	0.927	0.772	0.890	0.476
40%	S-Net+AD-Net+W-Net	0.971	0.483	0.957	0.631	0.946	0.829	0.937	0.634
60%	S-Net+AD-Net+W-Net	0.983	0.502	0.968	0.658	0.969	0.841	0.943	0.669
80%	S-Net+AD-Net+W-Net	0.985	0.510	0.979	0.682	0.977	0.847	0.959	0.710
100%	S-Net+AD-Net+W-Net	0.988	0.511	0.982	0.700	0.978	0.851	0.961	0.714
100%	Adv. HEDNet [101]	-	0.439	-	0.483	-	0.840	-	0.481
	AdvSeg [11]	0.961	0.470	0.924	0.592	0.945	0.79	0.939	0.675
	ASDNet [38]	0.969	0.478	0.932	0.628	0.950	0.809	0.948	0.692
	CoLL [38]	0.965	0.473	0.954	0.657	0.967	0.845	0.953	0.716

Table 5.4: *FGADR* \rightarrow *IDRID*: Semantic segmentation performance on *IDRID* (*target domain*). Models are trained with labeled data on *FGADR* and increasingly labeled data in the target domain from 0 \rightarrow 100%. **Red** indicates the best results for settings using 0% labeled data from target domain in the training step. **Green** stands for settings using 40% – 60% labeled data in target domain but outperform at least one of baselines trained with 100% data. **Bold** are the best values in all methods.

Results on *IDRID* \rightarrow *FGADR*: In table 5.3, we report the results for different settings on *FGADR* segmentation test set. These models are trained using the labeled data from *IDRID* training dataset and unlabeled data from *FGADR* dataset. We can observe that the settings without any domain adaptation *S-Net* performed poorly during the inference on *FGADR* domain. The AUC-PR scores for all of the lesion classes are below 0.3. Direct entropy minimization method (*S-Net* + Entropy) has significant improvement in both the metrics with respect to *S-Net*. Introducing Adversarial network for entropy minimization (*S-Net* + *AD-Net*) has improved the AUC-ROC and AUC-PR scores by 12% and 15.3%, respectively, averaged over all four lesion classes. The best results among our domain adaptive settings with 0% labels data in target domain are highlighted in red colour. In short, we achieved the best results for all of the lesion classes when both adversarial entropy minimization and Wasserstein domain critic model are considered together.

Plots in figure 5.2 compare the improvement in AUC-PR scores on the lesion classes of *FGADR* segmentation test set for different domain adaptation settings. In summary, we observe that all of our domain adaptive settings contribute for significantly improving performance over non-adaptive settings *S-Net*. Besides, it can be seen in figure 5.3 that the training and validation loss curves of all lesion classes for the setting *S-Net*

+ AD-Net + W-Net using 0% labeled data from the FGADR monotonously decline, confirming the stability in the learning process.

We also compare our method with several baselines trained directly on the target domain *FGADR* using 100% training labels. In table 5.3, the baselines we considered are Fully convolutional segmentation (*FCN-8s*) [130], Deep Lap model (*DLV3+*) [131], U-shaped models (*U-Net*) [15], and *Attention U-Net* [132]). The results for those baselines are taken from [8]. For a detailed comparison, we trained our domain adaptive model along with different percentages of labeled data from target domain. Results in green indicate the settings where we surpassed one or more baselines while using 40% \rightarrow 60% labeled data. Results in bold indicate the best scores in all settings. In general, we observe that, for all of the lesion classes, by using only 40% – 60% of labeled data in the target domain, we are able to outperform more than one baselines and with 80% – 100% labeled data, we derive better AUC-ROC and AUC-PR scores compared with other methods in most cases. For instance, we gained 1.6%, 0.5%, 0.6% and 0.4% AUC-ROC score improvement on *MA*, *HE*, *EX*, and *SE* lesion classes respectively. The AUC-PR scores also improved for *MA*, *EX* and *SE* by 2.7%, 0.8% and 1.2% respectively. Figure 5.4 illustrates qualitative comparisons of lesion map predictions between S-Net settings and our proposed multi-lesion segmentation model trained with domain adaption constraints S-Net + AD-Net + W-Net.

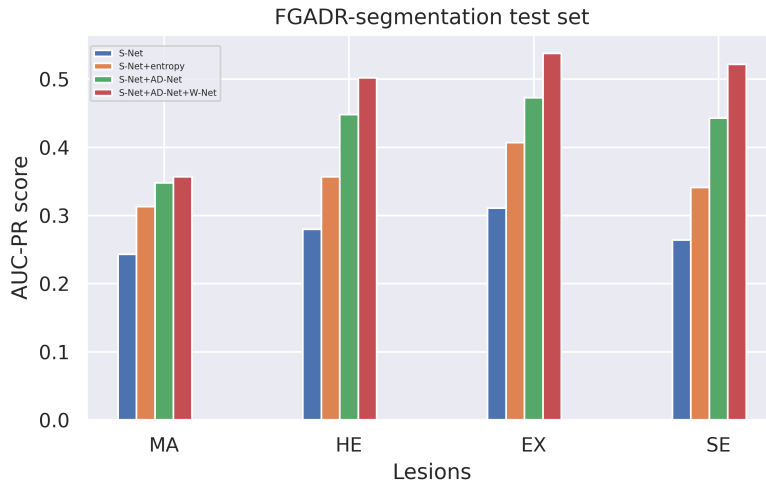


Figure 5.2: Comparison of our different domain adaption approaches for lesion segmentation on four types of lesions. Results are evaluated on the setting *IDRID* \rightarrow *FGADR* with 0% in target domain.

Results on *FGADR* \rightarrow *IDRID*: Table 5.4 shows the results on four lesion classes on *IDRID-segmentation* test set where the segmentation model is trained using the labeled data from *FGADR-segmentation* set and unlabeled images from *IDRID-segmentation* training dataset. For different domain adaptation settings, we discover similar trends as the *IDRID* \rightarrow *FGADR* case. In particular, we compare our method with various baselines trained with 100% labeled instances in the target domain (*IDRID*). These baselines include adversarial learning based segmentation networks *Adv. HEDNet* [101], *AdvSeg* [11] and semi-supervised collaborative learning networks *ASDNet* and *CoLL* [38]. With 0% of label data, the configuration S-Net + AD-Net + W-Net already has comparable

performance with the first baseline *Adv. HEDNet*. When using 40% labeled data, we are able to outperform most of competitive baselines in both AUC-ROC and AUC-PR metrics. This shows that, the proposed lesion generator model can be fine-tuned with minimal annotation data while still attains good performance in the target domain.

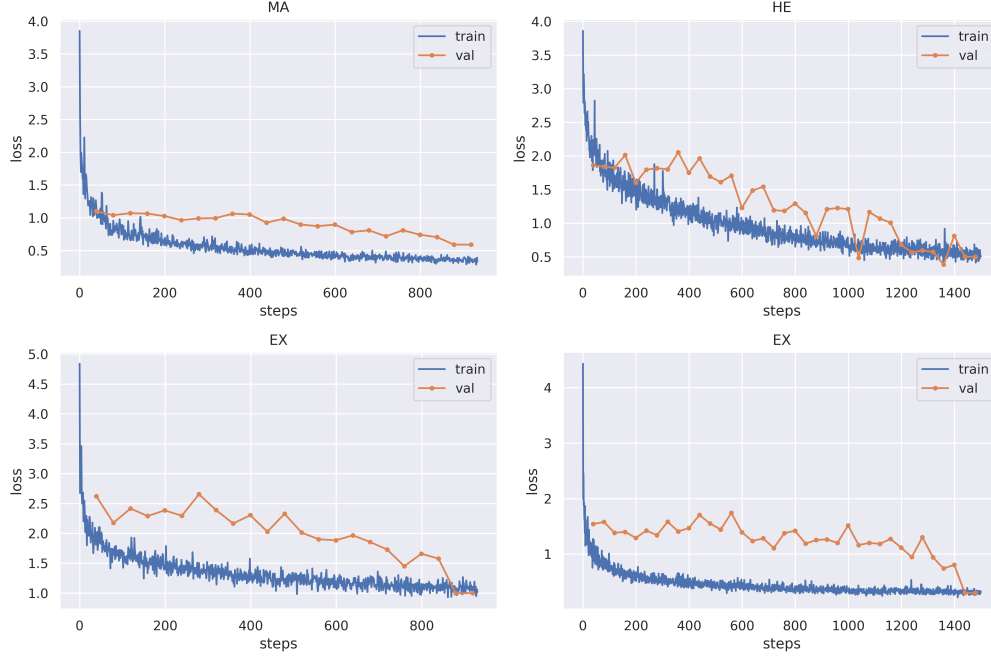


Figure 5.3: Training and validation loss curves for our domain adaptive lesion segmentation models. Results are on the setting $IDRID \rightarrow FGADR$ with 0% labeled data in the target domain. It can be seen that all training and validation curves tend to converge to stable points given more training steps.

5.5 Performance of Diabetic Grading Tasks with Attentive Lesion Information

5.5.1 Influence of Attention Mechanism on Grading Networks

We evaluated the effectiveness of predicted lesion segmentation maps in disease grading classification tasks. Our evaluation is on two types of vision related foundation models: CNN-based architecture and Vision Transformer-based architecture. We conducted experiments with the following settings.

1. *G-Net*: In this baseline, we directly train the disease grading classification network using retina fundus images and their associated disease grading labels using the classification loss \mathcal{L}_{cls} in equation (4.24). In this step only *G-Net* is trained and no attention functionality is used, therefore, we do not use any lesion information in the training process.

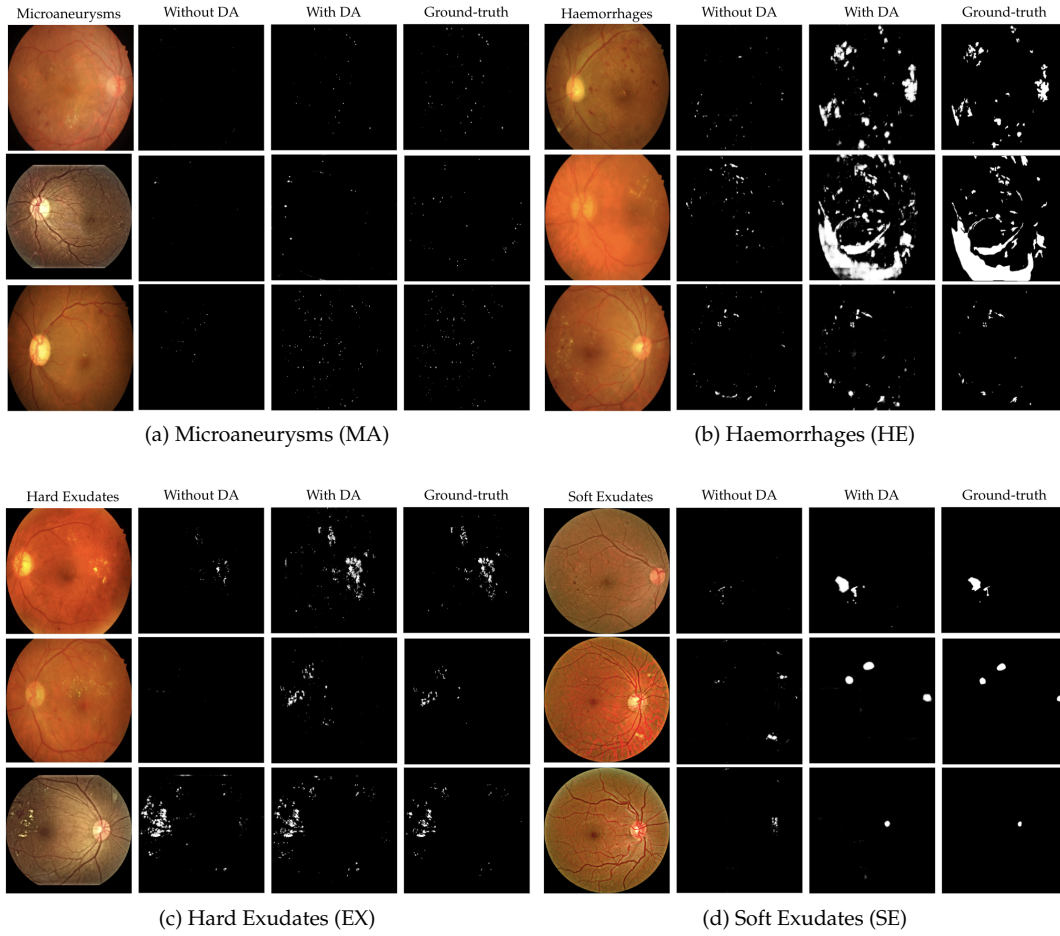


Figure 5.4: Qualitative multi-lesion segmentation results for four different lesion classes. Comparison of lesion maps predicted by the segmentation models trained without (with) Domain Adaption (DA) S-Net (S-Net + AD-Net + W-Net).

2. G-Net + Attention (low-level): We integrate an attention model *Att-Net* to highlight generated lesion features in the training process of the grading network *G-Net*. The attention mechanism we integrated here is on structural feature level of the grading network described in section 4.3.1. Here, the *G-Net* and *Att-Net* are jointly trained using equation (4.28).
3. G-Net + Attention (high-level): In this setting, we integrate the attention mechanism for lesion features following our high-level concept described in section 4.3.1.2. Unlike the previous low-level concept, this attention constrain is more intuitive where the normalized class activation maps are directly compared with the lesion feature maps to compute the attention overlapping loss $\mathcal{L}_{overlap}$ in equation (4.33). The *G-Net* and *Att-Net* then are jointly trained for classification.
4. G-Net + Attention (two-level): We combine both of the proposed *low-level* and *high-level* attention concept in the training process of the disease grading network *G-Net*. This is our final proposed method for the lesion attentive DR

grading model. The *G-Net* and *Att-Net* are jointly optimized using the equation (4.35).

Datasets	IDRID		EyePACS		FGADR	
Methods	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
G-Net (ResNet-50)	0.823	0.844	0.836	0.819	0.807	0.751
G-Net + Attention (low-level)	0.897	0.904	0.881	0.867	0.860	0.852
G-Net + Attention (high-level)	0.859	0.873	0.862	0.849	0.839	0.827
G-Net + Attention (two-level)	0.904	0.911	0.895	0.883	0.872	0.861

Table 5.5: Performance of the proposed attention mechanisms on the CNN-based architecture (ResNet-50). Results are evaluated on three datasets *IDRID*, *EyePACS*, and *FGADR*.

Datasets	IDRID		EyePACS		FGADR	
Methods	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
G-Net (MIL-ViT)	0.780	0.824	0.755	0.791	0.751	0.781
G-Net + Attention (high-level)	0.822	0.841	0.819	0.849	0.797	0.826

Table 5.6: Performance of the proposed attention mechanisms on the Vision Transformer architecture (MIL-ViT). Results are evaluated on three datasets *IDRID*, *EyePACS*, and *FGADR*. Because MIL-ViT already integrated attentions at the feature-level, we only extend it with the high-level case.

Datasets	IDRID		EyePACS		FGADR	
Methods	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
JCS [48]	-	-	0.886	0.877	0.856	0.842
AFN-Net [8]	-	-	0.861	0.856	0.836	0.784
CoLL [38]	0.913	0.904	0.891	0.872	0.86	0.848
G-Net (MIL-ViT) + Attention	0.822	0.841	0.819	0.849	0.797	0.826
G-Net (ResNet-50) + Attention	0.904	0.911	0.895	0.883	0.872	0.861

Table 5.7: Comparison of our lesion attentive disease grading models with other *state-of-the-art* baselines. These baseline methods are also used lesion information in solving the DR grading task.

For the CNN-based architecture, we evaluated all four of our proposed settings using *ResNet-50* for the *G-Net* model. For transformer-based architecture, we considered multi-head vision transformer (*MIL-ViT*) [117] as *G-Net*. Transformer models are inherently built on attention modules at the structural level (low-level); therefore, we only test *G-Net* and *G-Net + Attention (high-level)* methods for the transformer architecture.

We tested our approach on three publicly available datasets including *IDRID*, *EyePACS* and *FGADR* DR grading datasets. Table 5.5 and 5.6 present the evaluation results for different settings on these datasets for CNN-based and Transformer-based models respectively. For CNN-based models in table 5.5, we observe that both low-level and high-level attention mechanisms improved the performance for the original classification network *G-Net*. Integrating low-level attention with grading network in settings *G-Net + Attention (low-level)*, we could increase in Kappa scores by 6%, 5% and 10% for *IDRID*, *EyePACS* and *FGADR* datasets respectively. For the *G-Net + Attention (high-level)*, the kappa scores are improved by 3%, 3% and 7% respectively. We also gained best performance (in bold) for each of the datasets when combining both

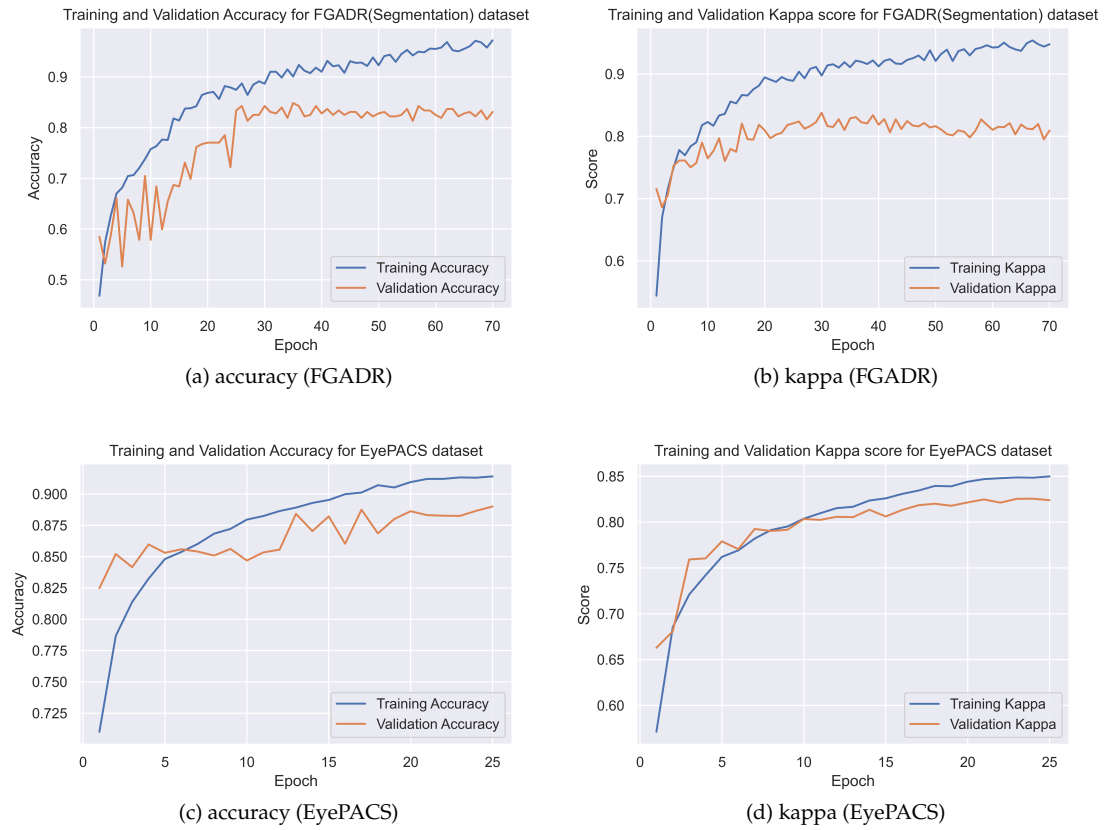


Figure 5.5: Illustration training and validation classification accuracy and kappa measurement for our proposed lesion attentive grading network G-Net + Attention (two levels). Plots in the top row (a) and (b) are the accuracy and kappa score curve in FGADR dataset. Plots in the bottom row (c) and (d) are results in EyePACS dataset.

low-level and high-level attention mechanisms for the setting `G-Net + Attention (two-level)`. Figure 5.5 describes the comparison between training and validation accuracy in terms of kappa scores for the *FGADR* and *EyePACS* data. We can see that the networks tend to achieve higher performance given more epochs.

Table 5.6 presents the results for transformer-based model. We can see that, integrating high level attention mechanism with *G-Net* has improved both classification accuracy and kappa scores for all three grading datasets. In table 5.7, we compared our best lesion attentive disease grading models (one each for CNN-based and Transformer-based methods) with some competitive baselines. Those methods includes, *JCS* ([48]), Denoising Attention Fusion Network (*AFN-Net*) [8]) and Collaborative DR grading (*CoLL*) [38]. Results of our CNN-based attention model *G-Net (CNN) + Attention* on *IDRID* is comparable with the other baselines. For *EyePACS* and *FGADR* dataset, we outperform other methods by around 2% in accuracy and kappa scores.

5.5.2 DR Grading Prediction with Explainable Property

Our aim is to construct an explainable intelligent decision support system for diabetic retinopathy diagnosis. In the inference step, along with the predicted class output, we compute the class activation map (CAM) for the input image. CAM highlights the discriminative regions for the predicted class based on which the deep neural model has taken its decision. For DR grading tasks, different lesion regions observed in retinal fundus image are the key clinical factors to diagnose the patient disease progression to a particular grading class. The explanation decision of a model for its predictions is required to be consistent with the human expert reasoning. In terms of diabetic retinopathy diagnosis, class activation map (CAM) for the predicted grading class should be able to highlight relevant lesion regions. Figure 5.6 illustrates an example of the explainable prediction in our system. For a given fundus image, our domain invariant lesion generator model *S-Net* predicts lesion maps for four different lesion classes (figure 5.6 (j) to 5.6 (l)). Then the attention-based grading model *G-Net* provides the disease grade and its decision explanation in the form of class activation map (figure 5.6 (b)). By comparing the CAM region with the predicted lesion positions, we can compare the quality of the model explanation. Intuitively, the overlapping rate of CAM on lesion maps and input images can provide explanation properties to the experts about the network’s decisions (figure 5.6 (c) and 5.6 (d)).

In figure 5.7 we qualitatively compare the explanation results between the grading model that is trained only using classification loss in *G-Net* settings and our proposed attention based grading model that incorporate both low-level and high-level attention concepts in settings *G-Net (CNN) + Attention (two-level)*. For some typical images and their lesion maps, we discover that our attention-based grading (figure 5.7 (d)) was able to capture almost all the important lesion regions whereas *G-Net* (figure 5.7 (c)) failed to capture these lesion regions and considered some irrelevant positions as the class discriminating mapping. In opposite, even some tiny lesion regions are also considered in the prediction decision of our grading model.

5.6 Performance of Trained System using User Feedback

To assess the consistency and robustness of our proposed architecture, we considered evaluating the performance of our models given user feedback. In real world deploy-

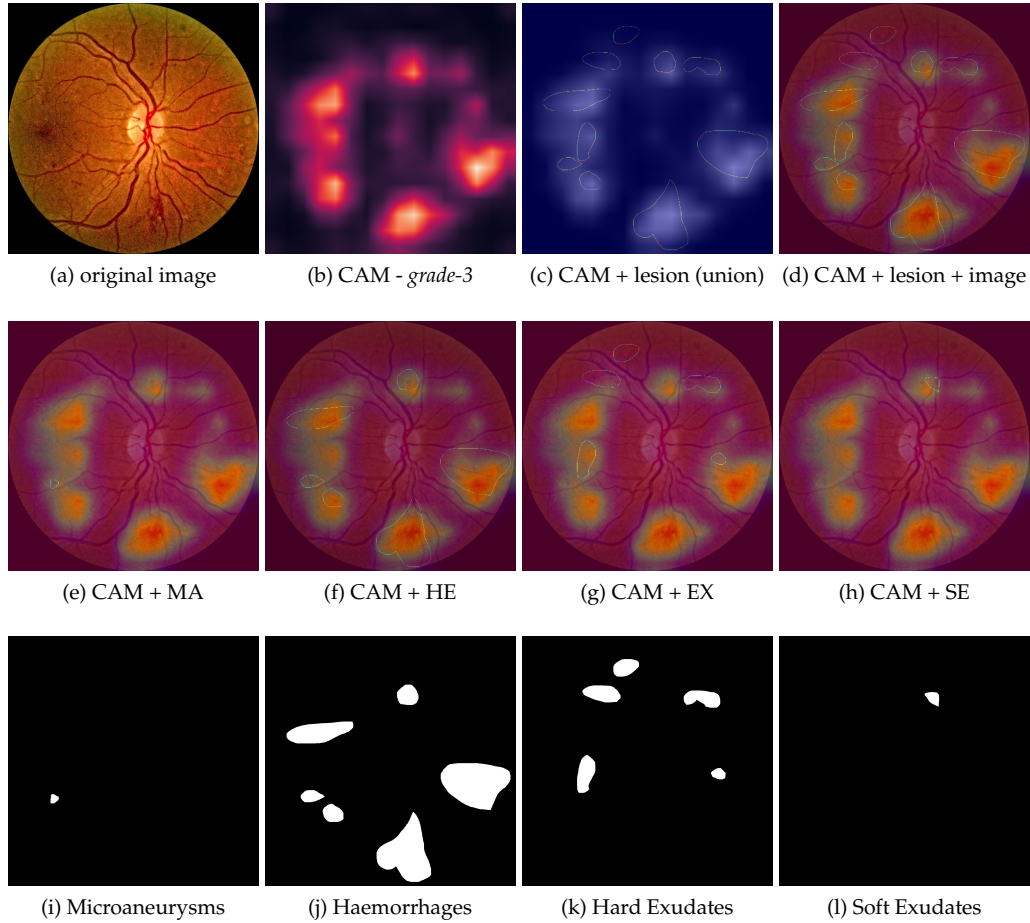


Figure 5.6: Input and outputs of our proposed intelligent diagnosis system for diabetic retinopathy (DR) disease. For an input retina fundus image (a), our domain invariant lesion generator generates the lesion maps for four lesion classes (the bottom row). In the top row, our network provides a (b) class activation map (CAM) for the predicted class (*grade-3*) (c) the overlap of CAM with the union of the predicted lesion maps, and (d) overlapping of CAM and union of predicted lesion maps on the original input image. In the second row, the CAM overlap with each of the lesion maps.

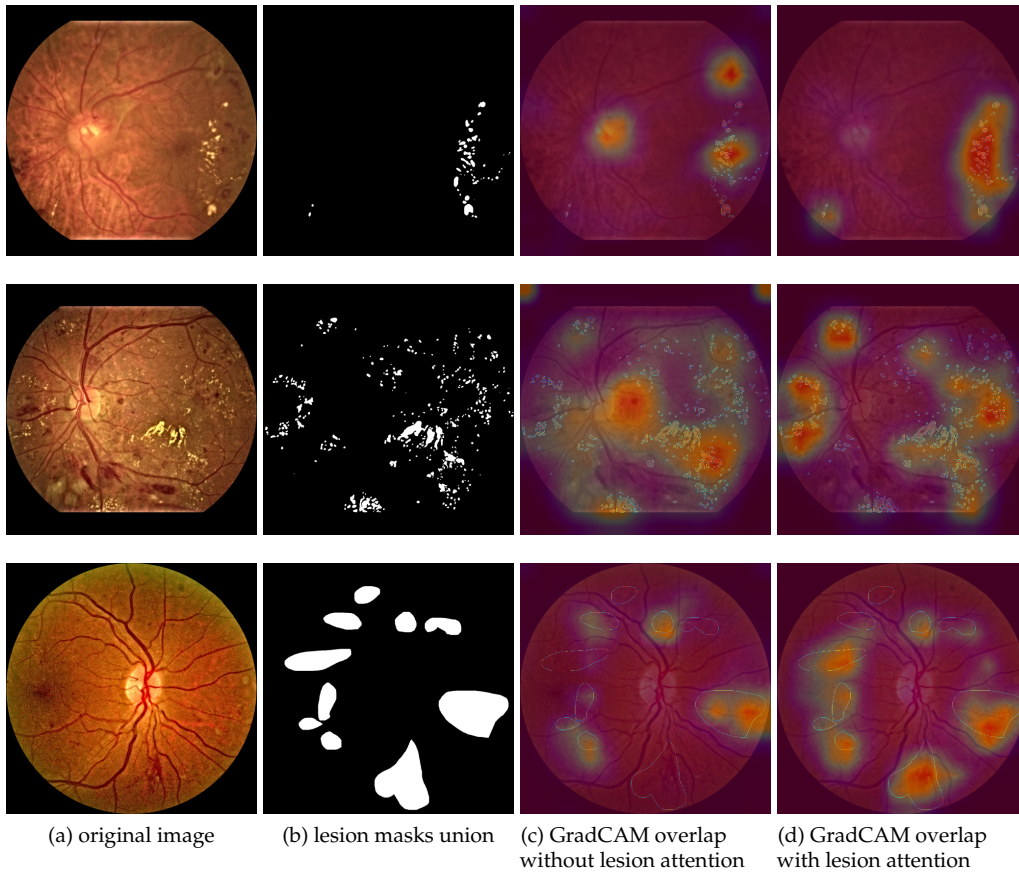


Figure 5.7: Comparison on explainability of the disease classification model trained with and without lesion attention. (a) original inputs, (b) union of ground-truth lesion maps, (c) overlapping of class activation map (CAM) of the predicted grading class with the (a) and (b) for *G-Net* trained without lesion attention model *Att-Net*, (d) overlapping of class activation map (CAM) of the predicted grading class with the (a) and (b) for *G-Net* trained with lesion attention model *Att-Net*. We can observe that CAMs in column (d) overlaps with most of the lesion regions.

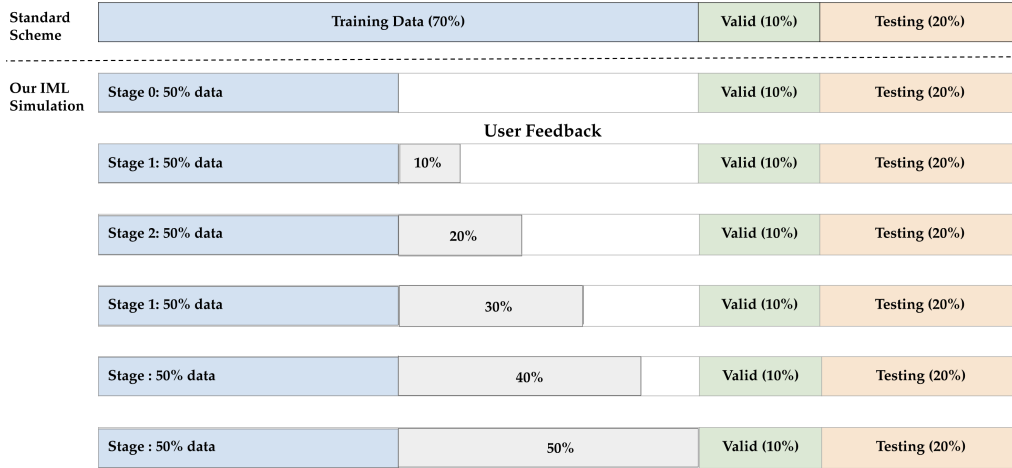


Figure 5.8: Illustration of user-feedback simulation on slices of data. In a standard scheme, traditional data split strategy is followed. In our interactive simulation case, we initially split the training data into two equal parts, and used one part of the split (50% of training data) to train our proposed attention-based grading model. For the remaining part of 50% training data, it is further split into five equal slices, serving for increasingly data feedback. After each time getting data feedback, the model is fine-tuned using both initial data and new ones. The performance of a new model then is evaluated on the hold-out testing set.

ment, inference for given input by the intelligent system can be validated and verified by domain experts. Within an interactive framework, given an input image, ophthalmologists can validate predicted lesion feature maps and grading class outputs. Using interactive input devices, they can provide feedback on these predictions. For example, if the predicted lesion maps have false positive or false negative regions, they can highlight the region and upon verification, these lesion maps can be used to further fine-tune our intelligent diagnosis system. As discussed in section 4.4.2, we can use these expert annotations directly to improve the performance of our lesion attentive grading model. In order to evaluate the effect expert’s validation and feedback process on our learning system, we conducted experiments by simulating user-feedback action. For this, we used *FGADR* dataset to simulate user feedback because this dataset has both annotations for lesion maps and grading tasks.

In figure 5.8, we illustrate our data split approach for user-feedback simulation. The original training data is divided into two equal splits. The disease grading model *G-Net* in beginning is trained using a half of training data and their lesion masks predicted by *S-Net*. The rest of training data is divided into five parts serving as increasingly data collected from users. After each update, a new model will be tested again on the fixed hold-out test set to justify performance. In our setting, instead of directly utilizing ground-truth samples as user feedback, we assumed that these data contain a certain noise level. This assumption makes sense in practice as we cannot guarantee accurate annotations from the user in all cases. Therefore, we simulated this scenario by applying some morphological operations on the ground-truth data and used them in the fine-tuning step. Specifically we randomly applied erosion and dilation operations for the lesion maps with a kernel size 15. Figure 5.9 demonstrates the effect of these operations on a ground-truth lesion mask.

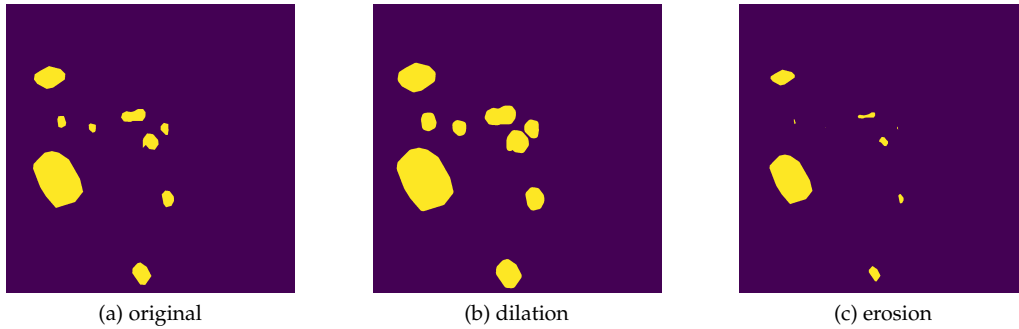


Figure 5.9: Illustration of noisy user-feedback on lesion features maps. We simulate the user behaviour to evaluate the robustness of our method by randomly performing dilation and erosion on original lesion masks and training models using these perturbed data. The examples are (a) original segmentation map, (b) lesion map after performing dilation, and (c) lesion map after erosion morphological operations using kernel size 15.

Table 5.8 presents our obtained results on *FGADR* data for the DR classifier network with rising user feedback data. The visualizations for these results are also illustrated in figure 5.10. At each stage, 10% of additional simulated user-feedback data is used to fine-tune a current stage of the grading network. We report both classification and explanation scores in our experiments, where the explanation score is computed based on the overlapping between the network’s heatmap regions and detected lesion positions using the Jaccard similarity [133]. Table 5.8 demonstrates to us a trend of improvement in both classification and explanation scores given enlarged user-feedback data. For example, with 30% or more response data, we achieved comparable scores against the other competitive baselines discussed previously. In summary, these tests validate the effectiveness of our approach in terms of Interactive Machine Learning (IML) and exhibit further improvement is possible if more data from users are provided with minimal annotations.

Method	% direct user feedback	Accuracy	Kappa	Explanation
G-Net + Attention	0%	0.773	0.781	0.272
	10%	0.786	0.801	0.295
	20%	0.822	0.838	0.331
	30%	0.841	0.853	0.363
	40%	0.850	0.863	0.366
	50%	0.855	0.866	0.387

Table 5.8: Performance of our framework utilizing increasingly user-feedback. We iteratively fine-tune the model using simulated user-feedback. Experiments are conducted on *FGADR* dataset.

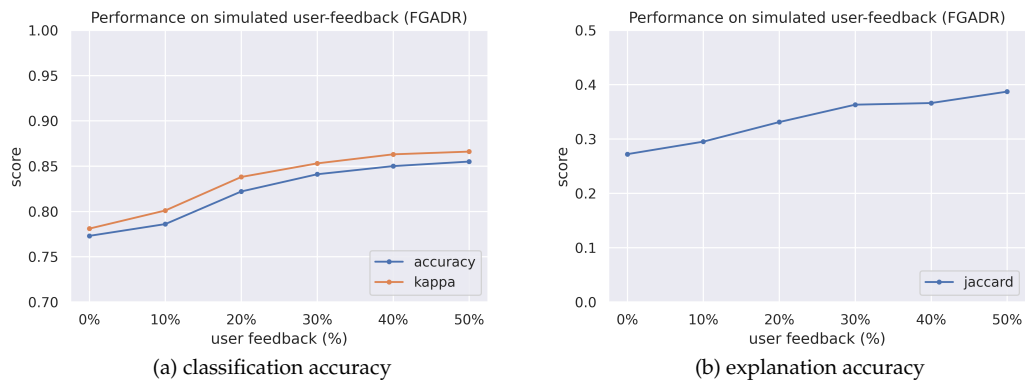


Figure 5.10: Visualizing performance of our framework utilizing increasingly user-feedback computed by (a) accuracy and kappa metrics and (b) explanation scores. Experiments are conducted on *FGADR* dataset.

Chapter 6

Discussion and Future Works

6.1 Discussion

This thesis provides a unified system for diabetic retinopathy (DR) grading task, which simultaneously learns to predict the disease grading and important lesion features. Motivated by ophthalmologists' clinical behavior in identifying DR disease, we incorporate lesion characteristics into the learning process of our disease grade prediction neural model. Experiments revealed that this framework significantly improves baseline performance and outperforms other competitive benchmarks. Generally, our method has the following strengths:

- First, given each retina fundus image as an input, the feature generator network automatically detects different types of lesions across domains while using only annotations in the source domain and unlabeled data in the target domain. We highlight that this feature is precious in practice since it allows for the rapid deployment of applications in new datasets without the need for extensive pixel-level annotation, which is often expensive and time-consuming to prepare. Technically, we built the feature generator network using innovative transfer learning algorithms and then restricted the feature representations to be domain-invariant using strategies from adversarial learning and the Wasserstein distance.
- Second, our attention network, which identifies and exploits the most significant lesion locations during grading network learning, performs at both low-level and high-level concepts. While the low-level concept considers merging latent embedding features of lesion and grading tasks and is trained to improve performance, the high-level idea in the other direction is based on explainable principles. The beyond hypothesis is that by observing the association between class activation mapping of grading network with detected lesion positions, the expert can inspect and uncover insight reasons to validate the network's predictions. In the experiment, we discovered that both attention components lead to improved accuracy and provide end-users with explainable attributes. Furthermore, we demonstrated

that these attention methods could be generalized to both CNN and Transformer neural architectures.

- Finally, the user role is taken into account during the method development process. We equip attention methods that make annotation progress more comfortable for users and make neural networks more robust in the presence of noise in new data. The empirical experiments validate this property; thereby, trained systems progressively improve their prediction given weakly supervised annotations created by users. Furthermore, the model also offers a variety of outputs to the user, such as DR Grading prediction, lesion positions, heatmap activation of the network, and especially showing the overlapping between these regions, serving as interpretative property in our system.

6.2 Future Works

In future works, we consider the following problems for further investigations:

- Even though our approach has integrated lesion information for the DR Grading task, there is still another medical criterion that can be utilized to improve performance. For instance, extensive characteristics of the lesion regions such as geometry, area, radius, or the degree of occurrence of different types of lesions, are other essential factors that need to formulate during the learning strategies. However, expressing such constraints is not straightforward since most of them are not differentiable, making end-to-end learning unfeasible. Fortunately, recent advances in machine learning subjects like discrete optimization [134], geometric deep learning [135, 136], and physic-informed machine learning [137] may offer us viable methods for incorporating these restrictions. For that reason, we believe that expanding our suggested method in those directions is worth investigating.
- The lack of training data is a primary obstacle that hinders the robustness and generalizability of a trained deep network. While we proposed techniques based on transfer learning and domain adaptation to alleviate these challenges, having a powerful pre-trained model is still in high demand. Currently, self-supervised learning methods trained on large-scale unlabeled data, namely the foundation model, have succeeded in various downstream tasks in natural language processing with well-known models such as BERT [64], DALL-E [138], and GPT-3 [139]. This raises the question of whether foundation models trained on large-scale medical datasets can bring similar performance for downstream medical tasks. In our setting, given such a foundation model, we expect to advance accuracy for neural networks in both lesion generator and DR Grading tasks, yielding increasing the performance of the whole system and reducing user efforts in preparing data annotations.
- This study organized experiments to confirm that the proposed method grows accuracy over time when provided user feedback in weakly-supervised forms. However, because these results are simulated in the computer system, they may not cover all real scenarios in practice. This encourages us to build a real intelligent user interface and deploy it for real-world applications. Such a system when operating in practice requires overcoming various barriers. For example, developing a user-friendly and intuitive UI/UX system so that experts easily engage with and provide

feedback to systems. When experts are in the annotation step, equipping computer-aided detection (CAD) [140] to discover and recommend similar marked locations to users is also necessary to save time and accelerate the progress. Finally, when there is a large amount of data feedback accessible, it poses concerns about learning for new instances while not forgetting past samples. Such questions are active topic research in continual learning and active learning, which is also a future direction for investigation.

Bibliography

- [1] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*. 2019;157:107843.
- [2] Yang Y, Shang F, Wu B, Yang D, Wang L, Xu Y, et al. Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image. *IEEE Transactions on Cybernetics*. 2021;.
- [3] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*. 2016;316(22):2402–2410.
- [4] Antal B, Hajdu A. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE transactions on biomedical engineering*. 2012;59(6):1720–1726.
- [5] Wang Z, Yin Y, Shi J, Fang W, Li H, Wang X. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2017. p. 267–275.
- [6] Buhrmester V, Münch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*. 2021;3(4):966–989.
- [7] Lin Z, Guo R, Wang Y, Wu B, Chen T, Wang W, et al. 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. p. 74–82.
- [8] Zhou Y, Wang B, Huang L, Cui S, Shao L. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*. 2020;40(3):818–828.
- [9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer; 2015. p. 234–241.
- [10] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer; 2018. p. 3–11.
- [11] Hung WC, Tsai YH, Liou YT, Lin YY, Yang MH. Adversarial learning for semi-supervised semantic segmentation. *Proceedings of the British Machine Vision Conference (BMVC)*. 2018;.

- [12] Vu TH, Jain H, Bucher M, Cord M, Pérez P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 2517–2526.
- [13] Sonntag D, Huber M, Möller M, Ndiaye A, Zillner S, Cavallaro A. Design and implementation of a semantic dialogue system for radiologists. arXiv preprint arXiv:170107381. 2017;.
- [14] El-Baz A, Jiang X, Suri JS. Biomedical image segmentation: advances and trends. CRC Press; 2016.
- [15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.
- [16] Bozorgtabar B, Ge Z, Chakravorty R, Abedini M, Demyanov S, Garnavi R. Investigating deep side layers for skin lesion segmentation. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE; 2017. p. 256–260.
- [17] Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE transactions on medical imaging. 2018;37(12):2663–2674.
- [18] Mehta S, Mercan E, Bartlett J, Weaver D, Elmore JG, Shapiro L. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 893–901.
- [19] Al-Masni MA, Kim DH, Kim TS. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Computer methods and programs in biomedicine. 2020;190:105351.
- [20] Li L, Verma M, Nakashima Y, Kawasaki R, Nagahara H. Joint learning of vessel segmentation and artery/vein classification with post-processing. In: Medical Imaging with Deep Learning. PMLR; 2020. p. 440–453.
- [21] Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data. 2018;3(3):25.
- [22] Nguyen DM, Nguyen TT, Vu H, Pham Q, Nguyen MD, Nguyen BT, et al. TATL: task agnostic transfer learning for skin attributes detection. Medical Image Analysis. 2022;78:102359.
- [23] Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics. 2016;3(2):119–131.
- [24] Wang J, Chen Y, Li W, Kong W, He Y, Jiang C, et al. Domain adaptation model for retinopathy detection from cross-domain OCT images. In: Medical Imaging with Deep Learning. PMLR; 2020. p. 795–810.
- [25] Sun S, Shi H, Wu Y. A survey of multi-source domain adaptation. Information Fusion. 2015;24:84–92.

- [26] Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:14123474. 2014;.
- [27] Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7167–7176.
- [28] Shen J, Qu Y, Zhang W, Yu Y. Wasserstein distance guided representation learning for domain adaptation. In: Thirty-second AAAI conference on artificial intelligence; 2018. .
- [29] Sun R, Li Y, Zhang T, Mao Z, Wu F, Zhang Y. Lesion-Aware Transformers for Diabetic Retinopathy Grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 10938–10947.
- [30] Zacharias J, Barz M, Sonntag D. A survey on deep learning toolkits and libraries for intelligent user interfaces. arXiv preprint arXiv:180304818. 2018;.
- [31] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: International conference on machine learning. PMLR; 2017. p. 214–223.
- [32] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.
- [33] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2921–2929.
- [34] Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia computer science. 2016;90:200–205.
- [35] Sankar M, Batri K, Parvathi R. Earliest diabetic retinopathy classification using deep convolution neural networks. pdf. Int J Adv Eng Technol. 2016;10:M9.
- [36] Alban M, Gilligan T. Automated detection of diabetic retinopathy using fluorescein angiography photographs. Report of standford education. 2016;.
- [37] Yang Y, Li T, Li W, Wu H, Fan W, Zhang W. 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. p. 533–540.
- [38] Zhou Y, He X, Huang L, Liu L, Zhu F, Cui S, et al. Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 2079–2088.
- [39] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need; 2017.
- [40] Sarafianos N, Xu X, Kakadiaris IA. Deep imbalanced attribute classification using visual attention aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 680–697.

- [41] Yu C, Wang J, Peng C, Gao C, Yu G, Sang N. Learning a discriminative feature network for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1857–1866.
- [42] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135–1144.
- [43] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:180607421. 2018;.
- [44] Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 3429–3437.
- [45] Lin M, Chen Q, Yan S. Network in network. arXiv preprint arXiv:13124400. 2013;.
- [46] Selvaraju RR, 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; 2017. p. 618–626.
- [47] Nguyen DM, Nguyen DM, Vu H, Nguyen BT, Nunnari F, Sonntag D. An attention mechanism using multiple knowledge sources for COVID-19 detection from CT images. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Workshop: Trustworthy AI for Healthcare; 2021. .
- [48] Wu YH, Gao SH, Mei J, Xu J, Fan DP, Zhang RG, et al. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. IEEE Transactions on Image Processing. 2021;30:3113–3126.
- [49] Amershi S, Cakmak M, Knox WB, Kulesza T. Power to the people: The role of humans in interactive machine learning. Ai Magazine. 2014;35(4):105–120.
- [50] Sonntag D, Wennerberg P, Buitelaar P, Zillner S. Pillars of ontology treatment in the medical domain. IGI Global; 2010.
- [51] Sonntag D, Schulz C, Reuschling C, Galarraga L. Radspeech’s mobile dialogue system for radiologists. In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces; 2012. p. 317–318.
- [52] Prange A, Chikobava M, Poller P, Barz M, Sonntag D. A Multimodal Dialogue System for Medical Decision Support inside Virtual Reality. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue; 2017. p. 23–26.
- [53] Sonntag D, Nunnari F, Profitlich HJ. The Skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions. Technical Report. arXiv preprint arXiv:200509448. 2020;.
- [54] 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; 2016. p. 770–778.
- [55] Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature communications. 2021;12(1):1–11.

- [56] Ventura D, Warnick S. A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences. 2007;19.
- [57] Yan W, Wang Y, Gu S, Huang L, Yan F, Xia L, et al. The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2019. p. 623–631.
- [58] Zhao S, Li B, Xu P, Keutzer K. Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv preprint arXiv:200212169. 2020;.
- [59] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.
- [60] Cheplygina V. Cats or CAT scans: Transfer learning from natural or medical image source data sets? Current Opinion in Biomedical Engineering. 2019;9:21–27.
- [61] He X, Yang X, Zhang S, Zhao J, Zhang Y, Xing E, et al. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. medrxiv. 2020;.
- [62] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. The Journal of Machine Learning Research. 2012;13(1):723–773.
- [63] Kpanou R, Osseni MA, Tossou P, Laviolette F, Corbeil J. On the robustness of generalization of drug–drug interaction models. BMC bioinformatics. 2021;22(1):1–21.
- [64] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). 2021;.
- [65] Kingma DP, Welling M. Auto-encoding variational bayes. Proceedings of the International Conference on Learning Representations (ICLR). 2014;.
- [66] Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. PMLR; 2014. p. 1278–1286.
- [67] Bank D, Koenigstein N, Giryas R. Autoencoders. arXiv preprint arXiv:200305991. 2020;.
- [68] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
- [69] Amini A, Soleimany A. MIT 6.S191: Introduction to Deep Learning;. Available from: http://introtodeeplearning.com/slides/6S191_MIT_DeepLearning_L4.pdf.
- [70] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. International Conference on Learning Representations (ICLR). 2018;.
- [71] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 4401–4410.

- [72] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1125–1134.
- [73] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2223–2232.
- [74] Adaloglou N. Transformers in Computer Vision. <https://theaisummer.com/>. 2021;.
- [75] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82–115.
- [76] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. *Electronics*. 2019;8(8):832.
- [77] Mueller ST, Hoffman RR, Clancey W, Emrey A, Klein G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv*. 2019;abs/1902.01876.
- [78] Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*. 2020;32(11):4793–4813.
- [79] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE; 2018. p. 80–89.
- [80] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160.
- [81] Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*. 2019;267:1–38.
- [82] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*. 2018;51(5):93.
- [83] Lipton ZC. The mythos of model interpretability. *Communications of the ACM*. 2018;61(10):36–43.
- [84] Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*. 2017;1(1):48–56.
- [85] Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE; 2018. p. 0210–0215.
- [86] Ras G, van Gerven M, Haselager P. Explanation methods in deep learning: Users, values, concerns and challenges. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer; 2018. p. 19–36.
- [87] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018;73:1–15.

- [88] Zhang Qs, Zhu SC. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*. 2018;19(1):27–39.
- [89] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: *In Workshop at International Conference on Learning Representations*. Citeseer; 2014. .
- [90] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. *ArXiv*. 2014;abs/1412.6806.
- [91] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer; 2014. p. 818–833.
- [92] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*. 2015;10(7):e0130140.
- [93] Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*. 2017;65:211–222.
- [94] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org; 2017. p. 3145–3153.
- [95] Ras G, Xie N, van Gerven M, Doran D. Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*. 2022;73:329–397.
- [96] Deng C, Ji X, Rainey C, Zhang J, Lu W. Integrating machine learning with human knowledge. *Iscience*. 2020;23(11):101656.
- [97] 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*; 2016. p. 770–778.
- [98] Athanasiadis T, Mylonas P, Avrithis Y, Kollias S. Semantic image segmentation and object labeling. *IEEE transactions on circuits and systems for video technology*. 2007;17(3):298–312.
- [99] Jadon S. A survey of loss functions for semantic segmentation. In: *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE; 2020. p. 1–7.
- [100] Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*. 2016;57(13):5200–5206.
- [101] Xiao Q, Zou J, Yang M, Gaudio A, Kitani K, Smailagic A, et al. Improving Lesion Segmentation for Diabetic Retinopathy using Adversarial Learning. In: *International Conference on Image Analysis and Recognition*. Springer; 2019. p. 333–344.
- [102] Luc P, Couprie C, Chintala S, Verbeek J. Semantic Segmentation using Adversarial Networks; 2016.
- [103] Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint arXiv:14111784*. 2014;.

- [104] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1125–1134.
- [105] Laine S, Aila T. Temporal ensembling for semi-supervised learning. International Conference on Learning Representations. 2017;.
- [106] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems. 2017;30.
- [107] Liu H, Gu X, Samaras D. Wasserstein gan with quadratic transport cost. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 4832–4841.
- [108] Shannon CE. A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review. 2001;5(1):3–55.
- [109] Ma A, Li J, Lu K, Zhu L, Shen HT. Adversarial entropy optimization for unsupervised domain adaptation. IEEE Transactions on Neural Networks and Learning Systems. 2021;.
- [110] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. Advances in neural information processing systems. 2017;30.
- [111] Wu L, Hong R, Wang Y, Wang M. Cross-entropy adversarial view adaptation for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology. 2019;30(7):2081–2092.
- [112] Tsai YH, Hung WC, Schuler S, Sohn K, Yang MH, Chandraker M. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7472–7481.
- [113] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124(7):962–969.
- [114] Li X, Hu X, Yu L, Zhu L, Fu CW, Heng PA. CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE transactions on medical imaging. 2019;39(5):1483–1493.
- [115] Selvaraju RR, 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; 2017. p. 618–626.
- [116] Li K, Wu Z, Peng KC, Ernst J, Fu Y. Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 9215–9223.
- [117] Yu S, Ma K, Bi Q, Bian C, Ning M, He N, et al. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 45–54.

- [118] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. Available from: <https://aclanthology.org/N19-1423>.
- [119] Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4190–4197. Available from: <https://aclanthology.org/2020.acl-main.385>.
- [120] Saini A, Prasad R. Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability. arXiv; 2021. Available from: <https://arxiv.org/abs/2108.06907>.
- [121] EyePACS Challenge Kaggle Diabetic Retinopathy Dataset;. Accessed: 2021-10-12. <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>.
- [122] McHugh M. interrater reliability: the kappa statistic. *Biochemica Medica*, 22 (3), 276–282; 2012.
- [123] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc.; 2019. p. 8024–8035. Available from: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [124] Buades A, Coll B, Morel JM. Non-local means denoising. *Image Processing On Line*. 2011;1:208–212.
- [125] Chen X, Duan Y, Houthoofd R, Schulman J, Sutskever I, Abbeel P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*. 2016;29.
- [126] Liu C, Belkin M. Accelerating sgd with momentum for over-parameterized learning. arXiv preprint arXiv:181013395. 2018;.
- [127] Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE; 2017. p. 464–472.
- [128] Zhang Z. Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE; 2018. p. 1–2.
- [129] Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:160904747. 2016;.
- [130] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–3440.
- [131] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*. 2017;40(4):834–848.

- [132] Islam M, Vibashan V, Jose V, Wijethilake N, Utkarsh U, Ren H. Brain tumor segmentation and survival prediction using 3D attention UNet. In: International MICCAI Brainlesion Workshop. Springer; 2019. p. 262–272.
- [133] Murphy AH. The Finley affair: A signal event in the history of forecast verification. *Weather and forecasting*. 1996;11(1):3–20.
- [134] Pogančič MV, Paulus A, Musil V, Martius G, Rolinek M. Differentiation of blackbox combinatorial solvers. In: International Conference on Learning Representations; 2019. .
- [135] Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 5115–5124.
- [136] Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:210413478*. 2021;.
- [137] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nature Reviews Physics*. 2021;3(6):422–440.
- [138] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR; 2021. p. 8821–8831.
- [139] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Advances in neural information processing systems*. 2020;33:1877–1901.
- [140] Yanase J, Triantaphyllou E. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*. 2019;138:112821.