CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading

Xiaomeng Li, *Student Member, IEEE*, Xiaowei Hu, Lequan Yu, *Student Member, IEEE*, Lei Zhu, *Member, IEEE*, Chi-Wing Fu, *Member, IEEE*, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*— Diabetic retinopathy (DR) and diabetic macular edema (DME) are the leading causes of permanent blindness in the working-age population. Automatic grading of DR and DME helps ophthalmologists design tailored treatments to patients, thus is of vital importance in the clinical practice. However, prior works either grade DR or DME, and ignore the correlation between DR and its complication, *i.e.*, DME. Moreover, the location information, *e.g.*, macula and soft hard exhaust annotations, are widely used as a prior for grading. Such annotations are costly to obtain, hence it is desirable to develop automatic grading methods with only image-level supervision. In this article, we present a novel cross-disease attention network (CANet) to jointly grade DR and DME by *exploring the internal relationship between the diseases* with only image-level supervision. Our key contributions include the disease-specific attention module to selectively learn useful features for individual diseases, and the disease-dependent attention module to further capture the internal relationship between the two diseases. We integrate these two attention modules in a deep network to produce disease-specific and disease-dependent features, and to maximize the overall performance jointly for grading DR and DME. We evaluate our network on two public benchmark datasets, *i.e.*, ISBI 2018 IDRiD challenge dataset and Messidor dataset. Our method achieves the best result on the ISBI 2018 IDRiD challenge dataset and outperforms other methods on the Messidor dataset. Our code is publicly available at https://github.com/xmengli999/CANet.**

***Index Terms*— Diabetic retinopathy, diabetic macular edema, joint grading, attention mechanism.**

I. Introduction

DIABETIC Retinopathy (DR) is a consequence of microvascular retinal changes triggered by diabetes. It is the most common leading cause of blindness and visual disability in the working-age population worldwide [1]. Structures such as microaneurysms, hemorrhages, hard exudates, and soft exudates are closely associated with DR and the presence of each of the aforementioned anomaly determines the grade of DR in the patient, as shown in Figure 1. Diabetic Macular Edema (DME) is a complication associated with DR, which is normally due to the accumulation of fluid leaks from blood vessels in the macula region or retinal thickening that occurs at any stage of DR [2]. The grading of the severity of DME is based on the shortest distances of the hard exudates to the macula. The closer the exudate is to the macular, the more the risk increases; see examples in Figure 2. The most effective treatment for DR and DME is at their early stage, for example, by laser photocoagulation. Therefore, in clinical practice, it is important to classify and stage the severity of DR and DME, so that DR/DME patients can receive tailored treatment at the early stage, which typically depends on the grading.

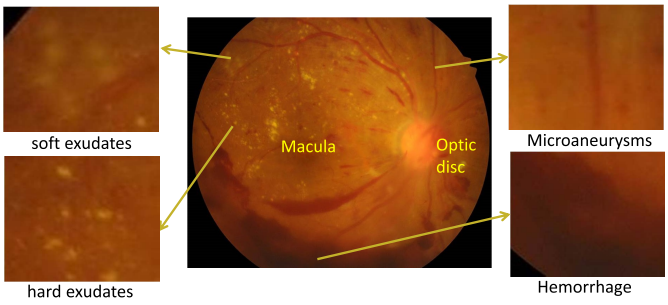


Fig. 1. Early pathological signs of DR, *e.g.*, soft exudates, hard exudates, microaneurysms, and hemorrhage, in a diabetic retinopathy image. Early pathological signs of DME is determined by the shortest distance of macula and hard exudates.

Convolutional neural networks (CNNs) have been proven to be a powerful tool to learn features for DR [3]–[5] and DME [6], [7] grading. For example, Islam *et al*. [3] developed a network to detect early-stage and severity grades of DR with heavy data augmentation. Zhou *et al*. [4] presented a multi-cell multi-task learning framework for DR grading by adopting the classification and regression losses. Regarding the DME grading, Ren *et al*. [6] presented a semi-supervised learning method with vector quantization. These methods, however, adopted different deep networks independently for grading each disease, ignoring the internal relationship between DR and DME, for example, the DME is the complication of DR.

Manuscript received September 30, 2019; revised October 26, 2019; accepted November 2, 2019. Date of publication November 6, 2019; date of current version April 30, 2020. This work was supported in part by the Research Grants Council of HKSAR under Project 14225616, in part by the Hong Kong Innovation and Technology Fund under Project ITS/311/18FP, and in part by the Shenzhen Science and Technology Program under Project JCYJ20170413162617606. *(Corresponding authors: Xiaomeng Li; Xiaowei Hu.)*

The authors are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: xmli@cse.cuhk.edu.hk; xwhu@cse.cuhk.edu.hk; lqyu@cse.cuhk.edu.hk; lzhu@cse.cuhk.edu.hk; cwfu@cse.cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2951844

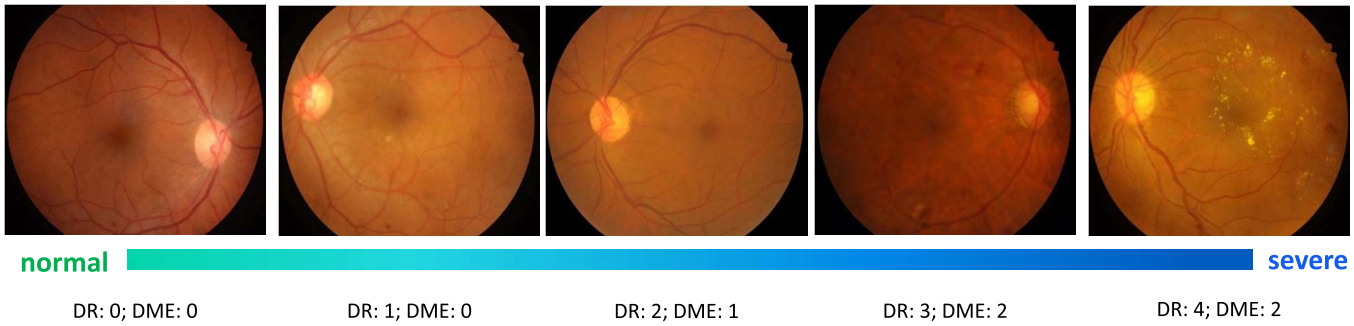


Fig. 2. Examples of fundus images with different pathological severity of DR and DME.

Recently, some works began to explore joint grading of DR and DME [5], [8]. Gulshan *et al*. [8] employed the Inception-v3 architecture for DR and DME grading, while Krause *et al*. [5] further improved the performance by utilizing the Inception-v4 architecture. However, these works focused on the network design and simply regarded the joint grading task as a multi-label problem, without considering the implicit relationship between these two diseases. In the medical imaging community, some work [9]–[11] employed multi-task learning to explore the relationship between different diseases (tasks). A key factor for the success in multi-task learning is that *the information among different tasks is shared, thereby promoting the performance of each individual task*.

To explore the feature relationship of DR and DME diseases and improve the grading performance for both diseases, it requires *an understanding of each disease, and also the internal relationship between two diseases*. To this end, we present a novel deep network architecture, called cross-disease attention network (CANet), to selectively leverage the features learned by the deep convolutional neural network, and produce disease-specific (within each disease) and disease-dependent features (between diseases) for joint DR and DME grading. In particular, we first develop a disease-specific attention module to select features from the extracted feature maps for individual disease (*i.e*., DR & DME). We then present a disease-dependent attention module to explore the internal relationship between two diseases by learning a set of attention weights, such that a larger weight indicates a higher risk of complication (*e.g*., DME may lead to worsening DR), and vice versa. Through the attention mechanism, our network models the implicit relationship between these two diseases, and improves the joint grading performance.

In summary, our contributions are three folds:

- We present a novel and effective method, named as cross-disease attention network (CANet), to jointly model the relationship between DR and its complication, *i.e*., DME. To the best of our knowledge, this is the first work for joint modeling the disease and its complication for fundus images.
- We propose the disease-specific attention module to selectively learn useful features for individual diseases, and also design an effective disease-dependent attention module to capture the internal relationship between two diseases.
- Experiments on the public IDRiD [12] challenge dataset and the Messidor [13] dataset show that our CANet method outperforms other methods on grading for both diseases, and achieves the best performance on the IDRiD dataset.

II. Related Work

A. Diabetic Retinopathy Grading

Early works on automatic diabetic retinopathy grading were based on the hand-crafted features to measure the blood vessels and the optic disc, and on counting the presence of abnormalities such as microaneurysms, soft exudates, hemorrhages, and hard exudates, *etc*. Then the grading was conducted using these extracted features by different machine learning methods [14]–[21], *e.g*., support vector machines (SVM) and k-nearest neighbor (kNN) and Gaussian mixture model.

In the last few years, deep learning algorithms have become popular for DR grading [22]–[28]. There are mainly two categories of deep learning methods for identifying DR severity. The first category is to use location information of tiny lesions, *e.g*., microaneurysms, hemorrhage, to determine DR grading performance. Van Grinsven *et al*. [29] sped up model training by dynamically selecting misclassified negative samples for hemorrhage detection. Dai *et al*. [30] proposed a multi-modal framework by utilizing both expert knowledges from text reports and color fundus images for microaneurysms detection. Yang *et al*. [31] designed a two-stage framework for both lesion detection and DR grading by using the annotations of locations including microaneurysms, hemorrhage, and exudates. Lin *et al*. [32] developed a new framework, where it first extracted lesion information and then fused it with the original image for DR grading. Zhou *et al*. [33] proposed a collaborative learning method for both lesion segmentation and DR grading using pixel-level and image-level supervisions simultaneously.

The second category uses image-level supervision to train a classification model to distinguish DR grades directly [8], [34], [35]. Gulshan *et al*. [8] proposed an inception-V3 network for DR grading. Gargeya and Leng [34] designed a CNN-based model for DR severity measurements. Wang *et al*. [35] used attention maps to highlight the suspicious regions, and predicted the disease level accurately based on the whole image as well as the high-resolution suspicious patches. It is expensive to annotate the labels on the medical images in a pixel-wise

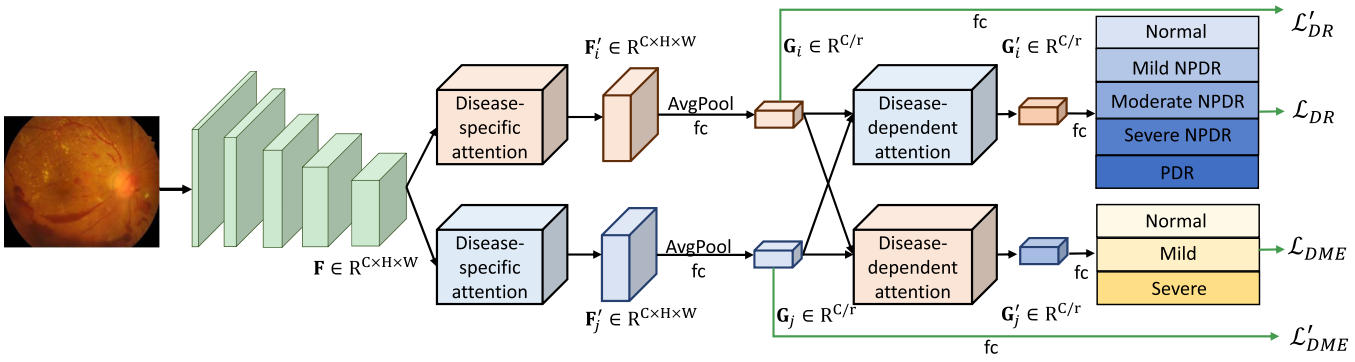


Fig. 3. The schematic illustration of the overall cross-disease attention network (CANet). G_i and G_j denote the disease-specific features for DR and DME, respectively; G_i' and G_j' denote the refined features with disease-dependent information for DR and DME, respectively; r is the ratio to reduce the number of feature channels for saving parameters; fc denotes the fully connected layer; The final loss function is the weighted combination of $\mathcal{L}_{DR}$, $\mathcal{L}_{DR}'$, $\mathcal{L}_{DME}$, and $\mathcal{L}_{DME}'$.

manner, hence, we follow the second category to conduct disease grading with only image-level supervision.

B. Diabetic Macular Edema Grading

Like the DR grading task, grading DME also attracts much attention in the community [36]. The assessment of the severity of DME is based on the distances of the exudate to the macula. The closer the exudate is to the macular, the more the risk increases. Early works used hand-crafted features to represent the fundus images [37], [38]. For example, Akram *et al*. [37] presented a screening system for DME that encompassed exudate detection with respect to their position inside the macular region. The system first extract features for exudate candidate regions, followed by making a representation of those candidate regions. The exact boundaries were determined using a hybrid of GMM model. However, the capacity of the hand-crafted features is limited. CNN based methods [6], [7] have dramatically improved the performance of DME grading. For example, Ren *et al*. [6] proposed a semi-supervised graph-based learning method to grade the severity of DME. Syed *et al*. [7] used knowledge of location information of exudates and maculae to measure the severity of DME. However, all of these work utilize the location information of exudate regions for disease grading. Such annotations (both lesions masks and grading labels) are difficult to obtain, in this work, we grade DME with only image-level supervision. Under the image-level supervision,Al-Bander *et al*. [39] proposed a CNN-based method based on foveae and exudates location for DME screening. However, their method classifies the DME into two classes, which is simpler than ours.

C. Multi-Task Learning in Medical Imaging Domain

Since jointly grading DR and DME diseases is related to the multi-task learning, we also review related works in medical imaging domain [9]–[11], [40], [41] and most of them are designed for image classification or regression tasks. For example, Chen *et al*. [9] trained a classification network for four tasks on the Age-related Macular Degeneration disease grading by using the CNN layers to capture common features then fully connected layers to learn the features for individual tasks. Liu *et al*. [11] employed a margin ranking loss to jointly train the deep network for both lung nodule classification and attribute score regression tasks. Similarly, Tan *et al*. [10] used the multi-level shared features and designed individual decoders to jointly learn the organ probability map and regressing boundary distance map. In contrast to these works that jointly do classification and regression tasks, we design a novel deep network architecture to *explore the relationship between two diseases*, and improve the overall grading performance for both diseases.

III. Methodology

Figure 3 illustrates the overview of our cross-disease attention network (CANet) for joint DR and DME grading, consisting of two disease-specific attention modules [Figure 4 (a)] to learn disease-specific features and two disease-dependent attention modules [Figure 4 (b)] to explore correlative features between these two diseases.

A. Cross-Disease Attention Network

As shown in Figure 3, our cross-disease attention network takes a fundus image as the input and outputs the grading scores for both DR and DME diseases in an end-to-end manner. First, we adopt a convolutional neural network, *i.e*., ResNet50 [42] to produce a set of feature maps with different resolutions. Then, we take the feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ with the smallest resolution and highly-semantic information (the deepest convolutional layer in ResNet50) as the inputs for the following two disease-specific attention modules, which learn the disease-specific features $\mathbf{F}_i' \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_j' \in \mathbb{R}^{C\times H\times W}$ to understand each individual disease. Note that the feature is the one before the AvgPool and fully connected layer of original ResNet. It contains high-level semantic information for DR and DME. Afterwards, we propose disease-dependent attention modules to explore the internal relationship between the two correlative diseases and produce the disease-dependent features for DR and DME, respectively. Finally, we predict the

grading scores for DR and DME based on the learned disease-dependent features.

In the following subsections, we will first elaborate the disease-specific attention module and disease-dependent attention module in details, and then present the training and testing strategies of our network for DR and DME grading.

B. Disease-Specific Attention Module

Each disease has its specific characteristics, *i.e.*, DR is graded by the presence of soft exudates, hard exudates, hemorrhage, and microaneurysms while DME is determined by the shortest distance between the macula and hard exudates [25]. However, the feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ extracted by the convolutional neural network only contain the high-level representations of the input image and it is difficult to capture the specific characteristics for each disease. In order to learn the representation of each individual disease, we present a novel disease-specific attention module to learn the specific semantic features of DR and DME, receptively.

Figure 4 (a) illustrates the detailed structure of the proposed disease-specific attention module, which takes the feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as the input and adopts the channel-wise attention as well as the spatial-wise attention to highlight the inter-channel and inter-spatial relationship of the features related to each disease. Specifically, we first squeeze the spatial information from the shared feature maps $\mathbf{F}$ via spatial-wise average- and max-pooling operations, and obtain two kinds of global spatial features $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{max}$. Then, we feed them into a shared *MLP* (multi-layer perception) to produce the channel-wise attention maps $\mathbf{A_c}$. The channel-wise attention maps $\mathbf{A_c}$ are described in the following:

$$\mathbf{A_c} = \sigma[\mathbf{W}_1 ReLU(\mathbf{W}_0\mathbf{F}^c_{avg}) + \mathbf{W}_1 ReLU(\mathbf{W}_0\mathbf{F}^c_{max})] \quad (1)$$

where σ is a sigmoid function to normalize the attention weights into $[0, 1]$, $\mathbf{W}_0 \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C\times C/r}$ are the weights of the shared *MLP*, and r is the ratio to reduce the number of feature channels for saving the network parameters and we empirically set it as 0.5. After obtaining the learned attention weights $\mathbf{A_c}$, we multiply it with the original feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ to produce the disease-specific feature maps $\mathbf{F}_i$:

$$\mathbf{F}_i = \mathbf{A_c} \otimes \mathbf{F} \quad (2)$$

where $\otimes$ denotes the element-wise multiplication, and the attention weights $\mathbf{A_c}$ are broadcasted along the spatial dimension. Hence, we can select the disease-specific features and suppress the features that are irrelevant to the disease along the feature channels.

To further highlight the disease-specific features across the spatial domain, we follow [43], [44] and adopt another attention model, which aggregates the channel-wise information by applying the max-pooling and avg-pooling operations along the channel dimension and produces the feature maps $\mathbf{F}^s_{i,avg}$ and $\mathbf{F}^s_{i,max}$. Then, we concatenate these two feature maps together and use another convolutional operation to learn the 2D spatial-wise attention map $\mathbf{A_s}$:

$$\mathbf{A_s} = \sigma(Conv([\mathbf{F}^s_{i,avg}; \mathbf{F}^s_{i,max}]) \quad (3)$$

TABLE I

THE DETAILED STRUCTURE OF CROSS-DISEASE ATTENTION MODULES. "FC" REPRESENTS THE FULLY CONNECTED LAYER; "CONV" REPRESENTS THE CONVOLUTION OPERATION; "RELU" AND "SIGMOID" ARE THE RELU AND SIGMOID NON-LINEAR OPERATIONS, RESPECTIVELY; "CONCAT" REPRESENTS THE CONCATENATION OPERATION. FOR "CONV", WE USE PADDING TO KEEP THE SIZE OF THE FEATURE MAPS. THE SYMBOLS ARE DEFINED IN FIGURE 4

	Input feature	Type	Output feature
disease-specific	$\mathbf{F}$	MaxPool, Flatten	$\mathbf{F}^c_{max}$
	$\mathbf{F}$	AvgPool, Flatten	$\mathbf{F}^c_{avg}$
	$\mathbf{F}^c_{avg}$, $\mathbf{F}^c_{max}$	FC 1 (2048×128), ReLU	-
	-	FC 2 (128×2048)	-
	-	Sum, Sigmoid	$\mathbf{A}_c$
	$\mathbf{A}_c$, $\mathbf{F}$	Multiplication	$\mathbf{F}_i$
	$\mathbf{F}_i$	AvgPool	$\mathbf{F}^s_{i,avg}$
	$\mathbf{F}_i$	MaxPool	$\mathbf{F}^s_{i,max}$
	$\mathbf{F}^s_{i,avg}$, $\mathbf{F}^s_{i,max}$	Concat, Conv, Sigmoid	$\mathbf{A}_s$
	$\mathbf{A}_s$, $\mathbf{F}_i$	Multiplication	$\mathbf{F}'_i$
disease-dependent	$\mathbf{G}_i$	FC 1 (1024×64), ReLU	-
	-	FC 2 (64×1024)	-
	-	Sigmoid	$\mathbf{A}_c$
	$\mathbf{A}_c$, $\mathbf{G}_j$	Multiplication	$\mathbf{G}'_j$

where *Conv* is a convolution layer and σ denotes the sigmoid function. Finally, we obtain the disease-specific features $\mathbf{F}'_i$ ($\mathbf{F}'_j$ for another disease; see Figure 3) by multiplying the learned attention weights $\mathbf{A_s}$ with the feature maps $\mathbf{F}_i$ to select the disease-specific features across the spatial dimension:

$$\mathbf{F}'_i = \mathbf{A_s} \otimes \mathbf{F}_i \ . \quad (4)$$

Note that the attention weights $\mathbf{A_s}$ are broadcasted along the channel dimension during the multiplication. In this way, we can further selectively use the disease-specific features by enhancing the disease-relevant features and suppressing the disease-irrelevant features across the spatial domain.

We show the detailed structure of the disease-specific attention module in Table I. The input and output channel number of FC 1 and FC 2 in disease-specific module are 2048×128 and 128×2048, respectively. We use ReLu activation after the first fully connected layer in each attention module.

C. Disease-Dependent Attention Module

As the statistics of the grading labels shown in Table II and Table III, DR and DME have the internal relationship. On the one hand, the more exudates are, the greater risk of the macula may have, *i.e.*, severer of DR may lead to severer DME. On the other hand, the closer of exudates to the macula, the more risk of presences of pathological DR signs, *i.e.*, worser of DME may lead to worser DR. Motivated by this observation, we present the disease-dependent attention module [see Figure 4 (b)] to capture the internal relationship between these two diseases.

As shown in Figure 3, this model takes the disease-specific features of both DR and DME diseases as the inputs, *i.e.*, $\mathbf{G}_i$ and $\mathbf{G}_j$, which are obtained by adopting the average pooling and fully connection operations on $\mathbf{F}'_i$ and $\mathbf{F}'_j$, and then it learns to produce the disease-dependent features for DR or DME, respectively. Figure 4 (b) illustrates the detailed structures of the proposed disease-dependent attention module

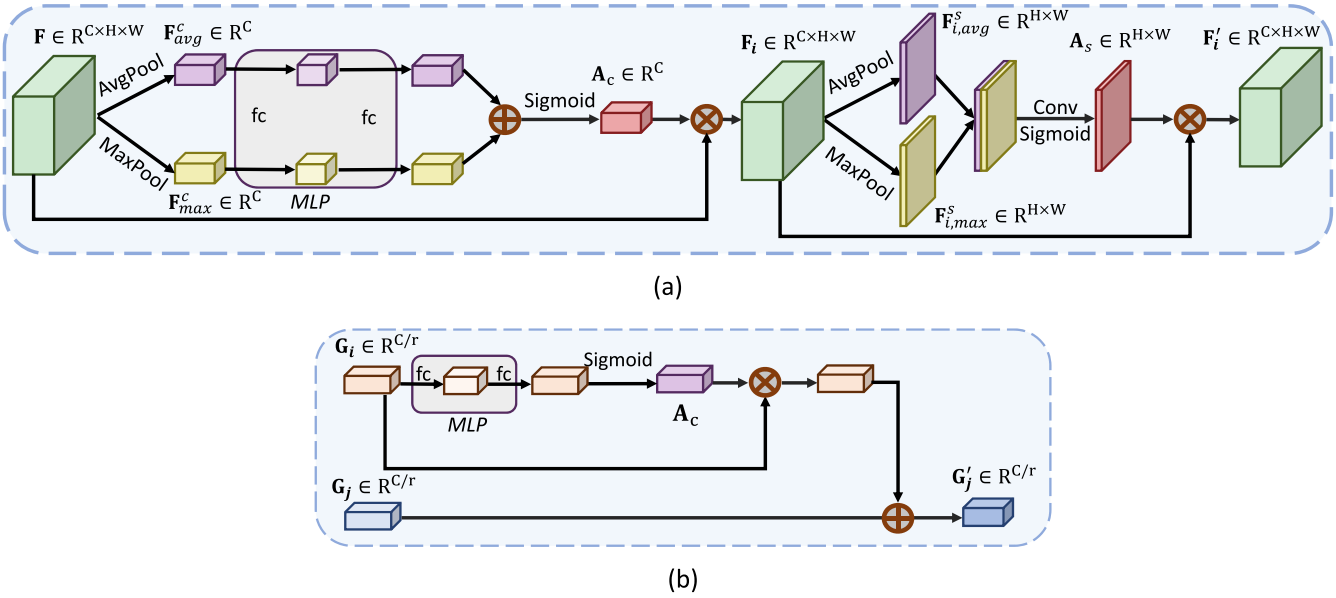


Fig. 4. The architectures of different attention modules. The disease-specific attention module (a) exploits both the inter-channel and inter-spatial relationship of features, while the disease-dependent module (b) explores and aggregates the informative inter-channel features from the other branch; (b) shows an example of disease-dependent attention module used for DME grading; A_c denotes the spatial-wise attention map and A_s denotes the channel-wise attention map; r is the ratio to reduce the number of feature channels for saving parameters. The actual details of the CANet topology can be found in Table I.

TABLE II
THE STATISTICS OF THE LABELS IN THE MESSIDOR DATASET. THE FIRST NUMBER IS THE COUNTS OF LABELS AND THE SECOND ONE IS THE RELATIVE VALUE

DME \ DR	0	1	2	3
0	546, 45.5%	142, 11.8%	182, 15.2%	104, 8.7%
1	0, 0.0%	5, 0.4%	28, 2.3%	42, 3.5%
2	0, 0.0%	6, 0.5%	37, 3.1%	108, 9.0%

TABLE III
THE STATISTICS OF THE LABELS IN THE IDRiD DATASET. THE FIRST NUMBER IS THE COUNTS OF LABELS AND THE SECOND ONE IS THE RELATIVE VALUE

DME \ DR	0	1	2	3	4
0	134, 26.0%	18, 3.5%	36, 7.0%	5, 1.0%	4, 0.8%
1	0, 0.0%	0, 0.0%	24, 4.6%	4, 0.8%	2, 0.4%
2	0, 0.0%	0, 0.0%	140, 27.1%	116, 22.4%	33, 6.4%

used for DME grading, which has the similar structures to the attention model used for DR grading.

Specifically, given the feature maps $\mathbf{G}_i$ of DR disease, we first employ a *MLP* and a sigmoid function to learn a set of attention weights $\mathbf{A}_{DR}$, and then multiply these weights with the input feature maps $\mathbf{G}_i$ to select the useful features, which helps to identify the DME disease. After that, we add the selected feature maps with the specific features of DME disease $\mathbf{G}_j$ in an element-wise manner ($\oplus$) to generate the disease-dependent features of DME $\mathbf{G}'_j$:

$$\mathbf{A}_{DR} = \sigma[\mathbf{W}_1^{DR} ReLU(\mathbf{W}_0^{DR}(\mathbf{G}_i))] \quad (5)$$

$$\mathbf{G}'_j = \mathbf{G}_j \oplus \mathbf{A}_{DR} \otimes \mathbf{G}_i. \quad (6)$$

Hence, the network is able to capture the correlation between the DR and DME diseases and improves the overall grading performance for both DR and DME diseases. The detailed structure of disease-dependent attention module is shown in Table I. The input and output channel number of FC 1 and FC 2 are 1024×64 and 64×1024, respectively. We use ReLu activation after the first fully connected layer in the attention module.

D. Network Architecture

We adopted ResNet50 as the backbone network to extract features, followed by a dropout layer with the drop rate of 0.3, and employed two disease-specific attention modules to learn disease-specific features. We employed two loss functions, *i.e.*, $\mathcal{L}'_{DR}$ and $\mathcal{L}'_{DME}$, to learn disease-specific features, and another two loss functions, *i.e.*, $\mathcal{L}_{DR}$ and $\mathcal{L}_{DME}$, for the final DR and DME grading:

$$\mathcal{L} = \mathcal{L}_{DR} + \mathcal{L}_{DME} + \lambda(\mathcal{L}'_{DR} + \mathcal{L}'_{DME}), \quad (7)$$

where $\mathcal{L}'_{DR}$ and $\mathcal{L}'_{DME}$ denote the cross-entropy loss for DR-specific and DME-specific feature learning, respectively; $\mathcal{L}_{DR}$ and $\mathcal{L}_{DME}$ denotes the loss function for the DR and DME grading. $\mathcal{L}_{DR}$ is a binary cross-entropy loss on the Messidor dataset and a 5-class cross-entropy loss on the IDRiD dataset. $\mathcal{L}_{DME}$ is a three-class cross-entropy loss on both Messidor and IDRiD dataset.

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_i^c \log \hat{y}_i{}^c \quad (8)$$

where $\hat{y}_i{}^c$ denotes the probability of voxel i belongs to class/grade c, and y_i^c indicates the ground truth label for

retinal image i. M is three for DME grading and two or five for DR grading (two in the Messidor dataset and five in the IDRiD dataset). Taking DR as an example, we directly apply a fully connected layer on DR-specific features $\mathbf{G}_i$ (batch size $\times$ 1024) for classification. The kernel size of fully connected layer is 1024 $\times$ 2 for Messidor dataset and 1024 $\times$ 5 for IDRiD dataset. λ is the weight in the loss function. When $\lambda = 0.0$, the network is optimized by the refined DR and DME features that include both disease-specific and disease-dependent information. As λ increasing, the framework gives more importance to the disease-specific feature learning. We analyze the effects of different λ in the experiment part, and we empirically set λ as 0.25.

E. Training and Testing Strategies

We normalized the training images and resize images to 350 $\times$ 350 resolution. For data augmentation, we randomly scaled and cropped the images into the patches with a size of 224 $\times$ 224. Random horizontal flip and vertical flip were also used to augment the training data. We optimized the network with Adam optimizer [45]. The initial learning rate was 0.0003 and we decayed the learning rate with a cosine annealing for each batch [46]. We trained the network for 1000 epochs and the batch size is 40. During the training process, we feed the samples of DR and DMR in a random order. The whole framework was built on PyTorch [47] with Titan Xp GPU. The network has 29 M trainable parameters. The training time of the network was five hours and the inference time was 0.02 seconds per image.

To test the grading result, we only used the prediction score after the refined DR and DME features, which include the disease-dependent information. We selected the class with the maximum prediction value in DR and DME, respectively. During inference, we did not use any post-processing operations and model ensemble techniques.

IV. Experiments and Results

A. Datasets

We evaluate the effectiveness of our method by comparing it against existing works on **Messidor dataset** [13][1] and **2018 ISBI IDRiD challenge dataset** [12][2]. To the best of our knowledge, these two datasets are the only two public datasets with both DR and DME severity grading annotations.

1) Messidor Dataset: This dataset has 1200 eye fundus color numerical images of the posterior pole acquired from three ophthalmologic departments. For each image in the dataset, its grading annotations of DR and DME are provided by the medical experts to measure the retinopathy grade and risk of macular edema. Specifically, DR is graded into four classes by the severity scale. Given the fact in the DR screening that the difference between normal images and images of stage 1 is the most difficult task for both the CAD systems and clinical experts, Sánchez *et al*. [48] grouped stages 0 and 1 of the Messidor dataset as referable images and combined stages 2 and 3 as non-referable in their screening work. This two-class setting has been widely used in the existing DR screening methods [35], [49], so that we conducted binary classification for DR grading in the Messidor dataset. To fairly compare with previous works [35], [49], [50], we use 10-fold cross validation on the entire dataset. DME is annotated based on the shortest distance d between the hard exudates location and the macula. The severity of DME is graded to 0 (No visible hard exudate), 1 ($d > 1$ papilla diameter), 2 ($d <= 1$ papilla diameter). The statistics of the DR and DME labels in the Messidor dataset is shown in Table II.

2) IDRiD Dataset: We employed the ISBI 2018 IDRiD sub-challenge 2 dataset. This dataset includes 516 images with a variety of pathological conditions of DR and DME, consisting of 413 training images and 103 test images. In the IDRiD dataset, each image contains both DR and DME severity grading labels. DR grade is annotated into five classes according to the severity scale, and we perform 5 class classification for DR. DME is annotated based on the shortest distance d between the hard exudates location and the macula. The annotation criteria of DME grading is the same as that in the IDRiD dataset. The statistics of the labels in the IDRiD dataset is shown in Table III. The detailed grading criterion for the IDRiD dataset can be found in the provided dataset websites. Note that we report 10-fold cross validation results for the Messidor dataset and use train & test sets split by the challenge organizers for the IDRiD dataset.

B. Evaluation Metrics

To measure the joint grading performance, we employ the IDRiD challenge evaluation metric "Joint Accuracy" (**Joint Ac**). The definition of **Joint Ac** is: If the prediction matches both DR and DME ground-truth label, then it is counted as one, else zero. The total number of true instances is divided by a total number of images to get the final result. We use Joint Ac to select our final model. For the Messidor dataset, we also report the accuracy (Ac), AUC, precision (Pre), recall (Rec), F1-score (F1) for each disease.

For the 2018 ISBI IDRiD dataset, we follow the challenge description and use the challenge evaluation metric ("Joint Ac") for comparison.

C. Analysis of Network Design

1) Compare With Baselines: We first compare our method with two baselines, *i.e*., "Individual training" and "Joint training" on the Messidor dataset. "Individual training (DR)" and "Individual training (DME)" indicates that we trained two individual ResNet50 networks for DR and DME grading, respectively. The "Joint training" denotes that we employed a ResNet50 network for shared feature extraction and two individual fully connected layers for DR and DME grading, respectively.

Table IV reports the 10-fold cross validation results of accuracy, AUC, precision, recall, F1-score for DR and DME respectively, as well as Joint Ac. It is observed that "Individual training (DR)" and "Individual training (DME)" achieve 89.5% AUC and 89.1% AUC for DR and DME, respectively. "Joint training" improves the individual training to 94.2% AUC and 90.5% AUC for DR and DME, respectively. Notably, our method (CANet) with the same backbone (ResNet50) and

[1]http://www.adcis.net/en/third-party/messidor/
[2]https://idrid.grand-challenge.org/Grading/

TABLE IV
QUANTITATIVE RESULTS ON THE MESSIDOR DATASET. THE REPORTED RESULTS ARE THE MEAN VALUES OF 10-FOLD CROSS VALIDATION. AC, PRE, REC, AND F1 DENOTE ACCURACY, PRECISION, RECALL, AND F1-SCORE, RESPECTIVELY (UNIT: %).

Methods	Parameters	Joint Ac	DR					DME				
			AUC	Ac	Pre	Rec	F1	AUC	Ac	Pre	Rec	F1
Individual training (DR)	23.52 M	-	89.5	81.0	76.2	78.9	77.3	-	-	-	-	-
Individual training (DME)	23.52 M	-	-	-	-	-	-	89.1	86.8	62.8	65.2	61.9
Joint training	23.52 M	82.0	94.2	89.1	86.5	88.2	87.2	90.5	90.4	78.7	73.9	75.3
Joint training (complex)	29.04 M	82.8	95.2	91.0	90.5	88.6	89.5	90.3	91.7	73.6	71.2	71.1
Joint training (complex_v2)	29.08 M	82.5	95.4	90.3	87.5	89.8	88.5	89.4	91.0	78.3	72.0	73.4
CANet (d-S only)	29.03 M	84.1	95.8	91.7	91.5	88.8	89.9	90.3	91.0	79.6	71.0	72.5
CANet (d-S; d-D $_{\text{DR}\Rightarrow\text{DME}}$)		84.5	96.0	91.7	91.5	88.7	90.0	89.4	91.9	82.3	72.0	74.8
CANet (d-S; d-D $_{\text{DR}\Leftarrow\text{DME}}$)		84.9	96.3	91.9	90.8	90.1	90.3	91.0	91.9	79.6	73.2	74.6
CANet (λ=0.25; final model)		**85.1**	96.3	92.6	90.6	**92.0**	**91.2**	**92.4**	91.2	76.3	70.8	72.4
CANet (λ=0.00)	29.03 M	84.8	96.3	92.1	92.2	88.6	90.3	91.6	91.5	**82.4**	73.7	75.3
CANet (λ=0.50)		84.7	96.1	91.4	89.9	89.7	89.6	92.2	**92.0**	78.6	**74.9**	75.3
CANet (λ=0.75)		84.8	**96.5**	92.2	**92.3**	88.8	90.3	91.8	**92.0**	79.5	71.4	73.7
CANet (λ=1.00)		84.9	96.3	**92.7**	91.6	90.6	91.0	90.7	91.7	77.4	69.4	71.0

training strategies improves the performance over these two baselines, with 96.3% AUC (DR) and 92.4% AUC (DME). The results show that the effectiveness of our method compared with these two baselines.

From the listed model parameters in Table IV, we can see that our method has more parameters (29.03 M), compared with the Joint training (23.52 M). To validate the effectiveness of our design under the same model complexity, we increase the parameters of "joint training" to 29.04 M by adding several standard components before classification on "joint training". These components include a convolutional layer with kernel size $2048 \times 300 \times 3 \times 3$, batch normalization layer and ReLU activation. secondWe also implemented another complex joint training baseline, *i.e.*, "Joint training (complex_v2)" in Table IV. This architecture is implemented by adding three convolutional layers on "Joint training" baseline. Specifically, the convolutional layer has the filter shapes of $2048 \times 660 \times 1 \times 1$, $660 \times 512 \times 3 \times 3$, and $512 \times 256 \times 3 \times 3$, respectively. Each convolutional layer is followed by a BN and a ReLU activation. Such complex baselines achieve 82.8% and 82.5% on joint Ac, respectively. However, with the same level of network parameters, our method (85.1%) still achieves the best performance, showing the effectiveness of the attention modules.

2) Analyze the Attention Module: We analyze the effects of disease-specific and disease-dependent attention modules. The comparisons are conducted with the same network backbone (ResNet50) and training strategies. The results are reported in Table IV by the 10-fold cross validation on the Messidor dataset. Compared with the "Joint training", adding the disease-specific attention module, *i.e.*, "CANet (d-S only)" enhances the Joint Ac from 82.0% to 84.1%. The accuracy of DR and DME are also improved from 89.1% to 91.7% (DR) and from 90.4% to 91.0% (DME), respectively. These comparisons demonstrate that disease-specific attention module explores more discriminative features for specific disease grading.

Then, we analyze the importance of the disease-dependent attention module for DME, *i.e.*, "CANet (d-S; d-D DR⇒DME)". This experiment indicates that the correlative feature learned on DR is incorporated to DME branch, and vice versa. It is observed that DR⇒DME improves the Joint Ac result to 84.5%, and DME grading results are enhanced on most evaluation metrics. Furthermore, we also analyze the importance of the disease-dependent attention module for DR, *i.e.*, "CANet (d-S; d-D DR⇐DME)". With this dependent attention branch, the joint accuracy is boosted to 84.9%, and DR grading results are also increased on most evaluation metrics. When we incorporate the disease-dependent attention module into both branches, our method "CANet ($\lambda = 0.25$; final model)" achieves the highest results, with joint Ac of 85.1%. These results validate that the disease-specific and disease-dependent attention module are both effective to utilize the disease-specific and disease-dependent information for better joint grading.

3) Analyze the Weight λ in the Loss Function: We analyze the effect of the weight λ in our method. The bottom part of Table IV shows the results with different weights in the loss function. When $\lambda = 0.00$, that whole framework is trained with the final refined DR and DME features that include the both specific and dependent information. When λ increases, the network is trained with the additional supervision for disease-specific attention. As shown in the Table IV, the variance of results with different λ is little, which indicates that our method is not very sensitive to the weight in the loss function. Our method reaches the best "Joint Ac" result (85.1%), when $\lambda = 0.25$. Therefore, we choose this model as our final model.

4) Analysis on Architectures: To analyze the effectiveness of backbone models, we perform experiments on "Joint training" to select the proper backbone architecture. The "Joint training" denotes that we employed a backbone network for shared feature extraction and two individual fully connected layers for DR and DME grading, respectively. We implemented with ResNet50 [42], ResNet34 [42], and DenseNet161 [51] and the results is showed in Table VI. We can see that ResNet50 achieves better results and finally we use ResNet50 as the backbone model.

TABLE V
COMPARISON WITH OTHER MULTI-TASK LEARNING METHODS ON THE MESSIDOR DATASET. THE REPORTED RESULTS ARE THE MEAN OF 10-FOLD CROSS VALIDATION (UNIT: %)

Methods	Joint Ac	DR					DME				
		AUC	Ac	Pre	Rec	F1	AUC	Ac	Pre	Rec	F1
Multi-task net [9]	82.4	94.8	89.9	89.7	85.7	87.5	90.5	90.5	79.1	70.8	72.2
MTMR-Net [11]	83.1	94.9	90.3	90.0	86.7	88.1	90.6	90.4	**79.8**	**73.2**	**75.4**
CANet (ours)	**85.1**	**96.3**	**92.6**	**90.6**	**92.0**	**91.2**	**92.4**	**91.2**	76.3	70.8	72.4

TABLE VI
RESULTS OF DIFFERENT BACKBONE ARCHITECTURES ON THE MESSIDOR DATASET (UNIT: %)

Methods	Joint Ac
ResNet50	82.0
ResNet34	81.4
DenseNet161	78.9

D. Compare With Other Multi-Task Learning Methods

To the best of our knowledge, there is no previous work for joint DR and DME grading. To show the effectiveness of our method for joint grading, we compare our method with two recent multi-task learning methods in the medical imaging community. Chen *et al*. [9] designed a method for the Age-related Macular Degeneration disease grading, while Liu *et al*. [11] proposed a network for both lung nodule classification and attribute score regression tasks. Since these works are not tailored for DR and DME grading, we did not directly use their methods for joint DR and DME grading. Instead, we adapted their key ideas to our task with the same network backbone and training strategies for fair comparison. For [9], after the ResNet50 feature extractor, we use the average pooling operation. Then, we use another one fully connected layer to reduce the channel number to 1024, followed by three fully connected layers (channel number: 1024, 256, 128) for DR and DME grading, respectively. The dropout layer is also employed. For [11], we use a fully connected layer with channel size 256 to concatenate the information from one task to another task, then two individual fully connected layers are employed for final DR and DME grading, respectively.

We report the performance of these two methods in Table V. It is observed that our method clearly outperforms these multi-task learning based methods on the Joint Accuracy metric. Compared with [11], our method achieves 1.4% (AUC) and 2.3% (Ac) improvement for DR; 1.8% (AUC) and 0.8% (Ac) improvement for DME. These results show the superiority of our framework for joint DR and DME grading.

E. Comparisons on the Messidor Dataset

We also compare our method with other DR grading models and DME grading models reported on the Messidor dataset in Table VII. As described in section II, there are two main branches for DR grading: employing both image-level and lesion location information as the supervision [32], [33], [53], [54], and employing only image-level supervision [35], [48], [49]. As for DME grading, some works [6], [7], [38] utilized macular or lesion location information features to help the grading of DME. For fair comparison, we only compare with those methods with only image-level supervision.

TABLE VII
RESULTS OF DIFFERENT METHODS ON THE MESSIDOR DATASET. OUR RESULT IS UNDER 10-FOLD CROSS VALIDATION. OTHER RESULTS ARE COPIED FROM ORIGINAL PAPERS. "–" INDICATES NO REPORTED RESULT

Methods	DR		DME	
	AUC	Ac	AUC	Ac
Lesion-based [50]	76.0	-	-	-
Fisher Vector [50]	86.3	-	-	-
VNXK/LGI [49]	88.7	89.3	-	-
CKML Net/LGI [49]	89.1	89.7	-	-
Comprehensive CAD [48]	91.0	-	-	-
DSF-RFcara [20]	91.6	-	-	-
Clinical B [48]	92.0	-	-	-
Clinical A [48]	94.0	-	-	-
Zoom-in-net [35] †	95.7	91.1	-	-
DME classifier [39]	-	-	-	88.8
CANet (ours)	**96.3**	**92.6**	**92.4**	**91.2**

"†": denotes using additional dataset EyePACS [52] as the pretrain.

For DR grading models, the combined kernels with multiple losses network (CKML) [49] and VGGNet with extra kernels (VNXK) [49] aims to employ multiple filter sizes to learn fine-grained discriminant features. Moreover, clinical experts [48] were also invited to grade on the Messidor dataset. It is worth mentioning that our method outperforms the clinical experts by 2.3% abd 4.3% on the AUC metric. Note that the clinical experts are provided by specific expert in [48]. Recently, Wang *et al*. [35] proposed the gated attention model and combined three sub-networks to classify the holistic image, high-resolution crops and gated regions. It is worth noticing that they first pretrain their model on EyePACS dataset [52] and then fine tune on the Messidor dataset, while we only use the Messidar dataset to train our model. Our method with cross-disease attention module further pushes the result, which obtains 1.5% Ac and 0.6% AUC gain over Zoom-in-net. For DME grading, our model excels the other reported results [39] by 2.4% improvement on Ac metric.

F. Results on the IDRiD Challenge Leaderboard

Table VIII shows the results of our method and other challenge participation methods on the IDRiD challenge dataset.[3]

[3]Challenge results are in https://idrid.grand-challenge.org/Leaderboard/

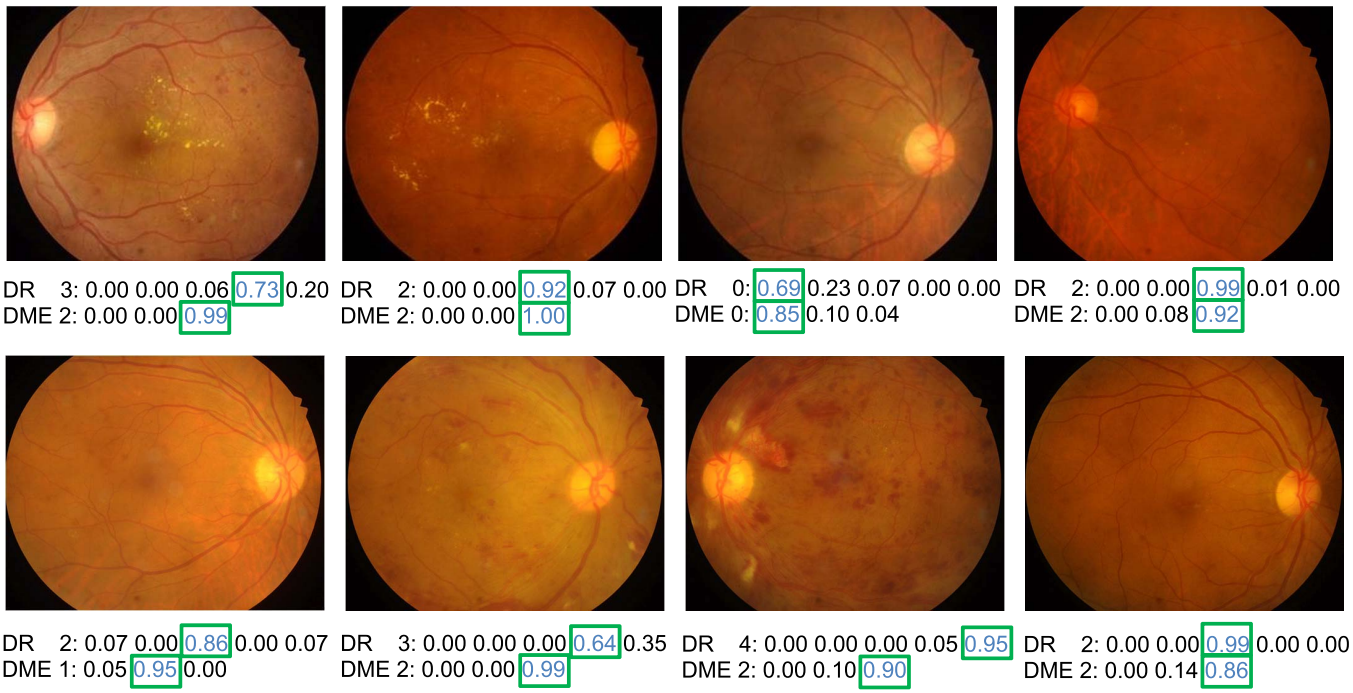
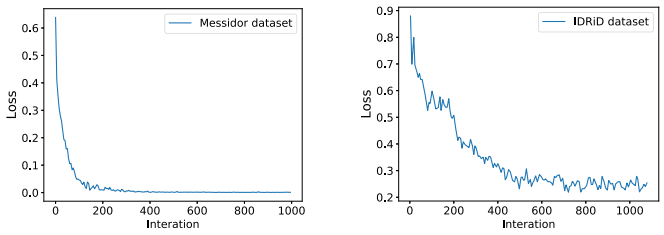


Fig. 5. Visual results of our method on the test set in the IDRiD dataset. We list the ground-truth, followed by the prediction score for different severity for each individual disease in a sequential order (0-4 for DR and 0-2 for DME). Blue indicates our predicted grade and green box indicates the ground-truth.

TABLE VIII
COMPARISON WITH THE REPORTED RESULTS ON THE IDRID LEADERBOARD. (UNIT: %)

Methods	Joint Ac	Rank
CANet (ours)	**65.1**	**1**
lzyuncc	63.1	2
VRT	55.3	3
Mammoth	51.5	4
HarangiM1	47.6	5
AVSASVA	47.6	5
HarangiM2	40.8	6



(a) (b)

Fig. 6. The learning curves of our method on the Messidor dataset (a) and IDRiD dataset (b).

Our model is trained with only the data in the Sub-challenge 2 (image-level supervision). It is observed that our model achieves a joint accuracy of 65.1%, which is higher than the top-ranked result by LzyUNCC (an unpublished work) on the leaderboard, with a relative 2.0% improvement on the joint accuracy. Lastly, it is worth noting that we trained our model using only the data in Sub-challenge 2 in the IDRiD dataset, while others (unpublished works) may use model ensembles or other supervision provided in other Sub-challenges.

We also analyze the effect of each attention module on the IDRiD dataset, and the results are shown in Table IX. With only disease-specific attention modules (CANet (d-S only)), our method excels the joint training baseline by 1%. Two disease-dependent modules "CANet (d-S, d-D DR⇒DME)" and "CANet (d-S, d-D DR⇐DME)" both further improve the joint grading performance by exploring the dependence between these two diseases. We can also observe that DME has much influences for the grading of DR, and this observation is consistent with that in the Messidor dataset in Table IV. With both direction dependent attention modules, our method achieves the best performance with Joint Ac 65.1%. Finally, we visualize some examples of the disease prediction score of our method on the IDRiD dataset in Figure 5. We can see that our method clearly differentiates the severity for DR and DME, respectively. secondAs shown in Figure 6, we visualize the learning curves of our method on the Messidor dataset and IDRiD dataset, respectively.

V. DISCUSSION

Recently, with the advances of deep learning techniques, automatic grading of DR and DME has been widely studied in the research community [6], [7], [32], [33], [35], [38], [53]–[55]. Although the large improvements have been achieved on these tasks, to the best of our knowledge, there is no previous works that jointly grade these two diseases and model the relationship between them. In this work, we investigate the importance of the

TABLE IX
RESULTS ON THE IDRID DATASET WITH DIFFERENT ATTENTION MODULES SETTING. (UNIT: %)

Methods	Joint Ac
Joint training	59.2
CANet (d-S only)	60.2
CANet (d-S, d-D $_{\text{DR}\Rightarrow\text{DME}}$)	62.1
CANet (d-S, d-D $_{\text{DR}\Leftarrow\text{DME}}$)	63.1
CANet (final model)	**65.1**

relationship between DR and DME for the joint grading task, and propose a cross-disease attention network (CANet) to capture the relationship between these two diseases. One method consists of two kinds of attention modules: one to learn disease-specific features and another to learn disease-dependent features. Results shown on two public benchmark datasets, *i.e.*, the Messidor dataset and 2018 ISBI IDRiD dataset, demonstrated the effectiveness of our method.

Although the good performance achieves, the limitation of our method still exists. The whole network is trained with only image-level supervision, making it very challenging to find the accurate abnormal signs, such as soft exudates, hard exudates, microaneurysms, and hemorrhage. The lesion masks or bounding boxes would provide the location information of these abnormal signs, which would be largely beneficial to the grading tasks [32], [33], [56], since the severity is usually based on the lesions. However, we are not aware of any public datasets containing both DR, DME grading labels, as well as the lesion or abnormal region segmentation masks. One solution is to collect the datasets with massive annotations, *i.e.*, lesion masks and the grading labels of multi-diseases. Another feasible solution is to explore how to utilize the lesion segmentation information from additional datasets to help the joint DR and DME grading. The dataset with lesion masks and dataset with DR & DME grading labels may have domain shifts, and generative adversarial networks [57]–[59] will be beneficial for this task.

Our method is feasible to extend to more correlated diseases. The attention mechanism aims to learn the attentional weights among multiple diseases. If we have multiple correlated diseases, the architecture will have multiple outputs, and each of them is optimized by an individual loss function to obtain the disease-specific features. Moreover, the disease-dependent attention module can be added to these diseases. For example, if there are five correlated diseases, 20 disease-dependent attention modules should be designed, and each module learns the correlation between every two diseases. Due to the high computational cost, the limitations would be the effective design of such attention blocks.

The future direction we would like to work on is to better model the relationship between DR and its complication DME, and also explore the relationship of multi-diseases occurred in one image. One potential research direction is to use the graph convolutional neural network [60] to model the relationship among different diseases. Through this, we hope to leverage the correlative information to improve joint grading performance. Also, it might bring some new insights to help doctors in understanding the diseases and their complications.

VI. CONCLUSION

In this work, we present a cross-disease attention network (CANet) to jointly grade DR and DME, and explore the individual diseases and also the internal relationship between two diseases by formulating two attention modules: one to learn disease-specific features and another to learn disease-dependent features. After that, the network leverages these two features simultaneously for DR and DME grading to maximize the overall grading performance. Experimental results on the public Messidor dataset demonstrate the superiority of our network over other related methods on both the DR and DME grading tasks. Moreover, our method also achieves the best results on the IDRiD challenge dataset. In the future, we plan to train our network jointly with the lesion annotations to further improve the DR and DME grading performance.

REFERENCES

[1] N. Cho *et al.*, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018.
[2] A. Das, P. G. McGuire, and S. Rangasamy, "Diabetic macular edema: Pathophysiology and novel therapeutic targets," *Ophthalmology*, vol. 122, no. 7, pp. 1375–1394, Jul. 2015.
[3] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," in *Proc. Int. Conf. Mach. Learn., Image Process., Netw. Secur. Data Sci.*, Dec. 2018.
[4] K. Zhou *et al.*, "Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2724–2727.
[5] J. Krause *et al.*, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, Aug. 2018.
[6] F. Ren, P. Cao, D. Zhao, and C. Wan, "Diabetic macular edema grading in retinal images using vector quantization and semi-supervised learning," *Technol. Health Care*, vol. 26, no. S1, pp. 389–397, 2018.
[7] A. M. Syed, M. U. Akram, T. Akram, M. Muzammal, S. Khalid, and M. A. Khan, "Fundus images-based detection and grading of macular edema using robust macula localization," *IEEE Access*, vol. 6, pp. 58784–58793, 2018.
[8] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.
[9] Q. Chen *et al.*, "A multi-task deep learning model for the classification of age-related macular degeneration," *AMIA Summits Transl. Sci. Proc.*, vol. 2019, pp. 505–514, May 2019.
[10] C. Tan, L. Zhao, Z. Yan, K. Li, D. Metaxas, and Y. Zhan, "Deep multi-task and task-specific feature learning network for robust shape preserved organ segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1221–1224.
[11] L. Liu, Q. Dou, H. Chen, I. E. Olatunji, J. Qin, and P.-A. Heng, "MTMR-Net: Multi-task deep learning with margin ranking loss for lung nodule analysis," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 74–82.
[12] P. Porwal *et al.*, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," vol. 3, no. 3, p. 25, Sep. 2018.
[13] E. Decencière *et al.*, "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
[14] N. Silberman, K. Ahrlich, R. Fergus, and L. Subramanian, "Case for automated detection of diabetic retinopathy," in *Proc. AAAI Spring Symp. Ser.*, Mar. 2010.
[15] A. Sopharak, U. Bunyarit, and S. Barman, "Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy C-means clustering," *Sensors*, vol. 9, no. 3, pp. 2148–2161, 2009.

[16] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014.

[17] U. R. Acharya, C. M. Lim, E. Y. K. Ng, C. Chee, and T. Tamura, "Computer-based detection of diabetes retinopathy stages using digital fundus images," *Proc. Inst. Mech. Eng., H, J. Eng. Med.*, vol. 223, no. 5, pp. 545–553, Jul. 2009.

[18] M. U. Akram, S. Khalid, A. Tariq, S. A. Khan, and F. Azam, "Detection and classification of retinal lesions for grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 45, pp. 161–171, Feb. 2014.

[19] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowl.-Based Syst.*, vol. 60, pp. 20–27, Apr. 2014.

[20] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. P. Langlois, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1116–1126, Apr. 2016.

[21] N. Kumar, A. V. Rajwade, S. Chandran, and S. P. Awate, "Kernel generalized-Gaussian mixture model for robust abnormality detection," in *Proc. MICCAI*. Berlin, Germany: Springer, 2017, pp. 21–29.

[22] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.

[23] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2017, pp. 1–11.

[24] A. Kori, S. S. Chennamsetty, M. S. K. P., and V. Alex, "Ensemble of convolutional neural networks for automatic grading of diabetic retinopathy and macular edema," 2018, *arXiv:1809.04228*. [Online]. Available: https://arxiv.org/abs/1809.04228

[25] D. Xiao, A. Bhuiyan, S. Frost, J. Vignarajan, M.-L. Tay-Kearney, and Y. Kanagasingam, "Major automatic diabetic retinopathy screening systems and related core algorithms: A review," *Mach. Vis. Appl.*, vol. 30, no. 3, pp. 423–446, Apr. 2019.

[26] P. Cao, F. Ren, C. Wan, J. Yang, and O. Zaiane, "Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis," *Computerized Med. Imag. Graph.*, vol. 69, pp. 112–124, Nov. 2018.

[27] M. T. Hagos and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," 2019, *arXiv:1905.07203*. [Online]. Available: https://arxiv.org/abs/1905.07203

[28] D. S. W. Ting, L. Carin, and M. D. Abramoff, "Observations and lessons learned from the artificial intelligence studies for diabetic retinopathy screening," *JAMA Ophthalmology*, vol. 137, no. 9, pp. 994–995, 2019.

[29] M. J. Van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1273–1284, May 2016.

[30] L. Dai *et al.*, "Retinal microaneurysm detection using clinical report guided multi-sieving CNN," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2017, pp. 525–532.

[31] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in *Proc. MICCAI*. New York, NY, USA: Springer, 2017, pp. 533–540.

[32] Z. Lin *et al.*, "A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2018, pp. 74–82.

[33] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. CVPR*, Jun. 2019, pp. 2079–2088.

[34] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[35] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *Proc. MICCAI*. New York, NY, USA: Springer, 2017, pp. 267–275.

[36] M. M. Fraz, M. Badar, A. W. Malik, and S. A. Barman, "Computational methods for exudates detection and macular edema estimation in retinal images: A survey," *Arch. Comput. Methods Eng.*, vol. 26, no. 4, pp. 1193–1220, Sep. 2019.

[37] M. U. Akram, A. Tariq, S. A. Khan, and M. Y. Javed, "Automated detection of exudates and macula for grading of diabetic macular edema," *Comput. Methods Programs Biomed.*, vol. 114, no. 2, pp. 141–152, Apr. 2014.

[38] U. R. Acharya *et al.*, "Automated diabetic macular edema (DME) grading system using DWT, DCT Features and maculopathy index," *Comput. Biol. Med.*, vol. 84, pp. 59–68, May 2017.

[39] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Taee, B. M. Williams, and Y. Zheng, *Diabetic Macular Edema Grading Based on Deep Neural Networks*. Des Moines, IA, USA, Univ. of Iowa, 2016.

[40] P. Moeskops *et al.*, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Proc. MICCAI*, 2016, pp. 478–486.

[41] W. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Med. Image Anal.*, vol. 43, pp. 54–65, Jan. 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[43] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, pp. 3–19.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Jun. 2018, pp. 7132–7141.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Dec. 2015.

[46] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, May 2017.

[47] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. NIPS Workshop Autodiff Program Chairs*, Oct. 2017.

[48] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. Suttorp-Schulten, M. D. Abramoff, and B. van Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Investigative Ophthalmology Vis. Sci.*, vol. 52, no. 7, pp. 4866–4871, Jun. 2011.

[49] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2016, pp. 209–215.

[50] R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Beyond lesion-based diabetic retinopathy: A direct approach for referral," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 193–200, Jan. 2017.

[51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.

[52] *Kaggle Diabetic Retinopathy Detection Competition*. Accessed: 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[53] N. Ramachandran, S. C. Hong, M. J. Sime, and G. A. Wilson, "Diabetic retinopathy screening using deep neural network," *Clin. Experim. Ophthalmology*, vol. 46, no. 4, pp. 412–416, May/Jun. 2018.

[54] J. Sahlsten *et al.*, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," 2019, *arXiv:1904.08764*. [Online]. Available: https://arxiv.org/abs/1904.08764

[55] Y.-W. Chen, T.-Y. Wu, W.-H. Wong, and C.-Y. Lee, "Diabetic retinopathy detection based on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1030–1034.

[56] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, "An ensemble deep learning based approach for red lesion detection in fundus images," *Comput. Methods Programs Biomed.*, vol. 153, pp. 115–127, Jan. 2018.

[57] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*. [Online]. Available: https://arxiv.org/abs/1609.03126

[58] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, Jul. 2017, pp. 4681–4690.

[59] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, Jul. 2017, pp. 214–223.

[60] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: https://arxiv.org/abs/1812.08434