CABNet: Category Attention Block for Imbalanced Diabetic Retinopathy Grading

Along He, Tao Li, Ning Li, Kai Wang, *Member, IEEE*, and Huazhu Fu, *Senior Member, IEEE*

***Abstract*— Diabetic Retinopathy (DR) grading is challenging due to the presence of intra-class variations, small lesions and imbalanced data distributions. The key for solving fine-grained DR grading is to find more discriminative features corresponding to subtle visual differences, such as microaneurysms, hemorrhages and soft exudates. However, small lesions are quite difficult to identify using traditional convolutional neural networks (CNNs), and an imbalanced DR data distribution will cause the model to pay too much attention to DR grades with more samples, greatly affecting the final grading performance. In this article, we focus on developing an attention module to address these issues. Specifically, for imbalanced DR data distributions, we propose a novel Category Attention Block (CAB), which explores more discriminative region-wise features for each DR grade and treats each category equally. In order to capture more detailed small lesion information, we also propose the Global Attention Block (GAB), which can exploit detailed and class-agnostic global attention feature maps for fundus images. By aggregating the attention blocks with a backbone network, the CABNet is constructed for DR grading. The attention blocks can be applied to a wide range of backbone networks and trained efficiently in an end-to-end manner. Comprehensive experiments are conducted on three publicly available datasets, showing that CABNet produces significant performance improvements for existing state-of-the-art deep architectures with few additional parameters and achieves the state-of-the-art results for DR grading. Code and models will be available at https://github.com/he2016012996/CABnet.**

***Index Terms*— Diabetic retinopathy grading, attention mechanism, category attention block (CAB), global attention block (GAB).**

I. Introduction

DR IS a microvascular complication caused by diabetes and it is the leading cause of blindness and visual impairment in the world [1], with one in three diabetics presenting signs of DR [2]. The main pathological features of DR include microaneurysms, hemorrhages, soft exudates, and hard exudates. According to the type and number of lesions in fundus images, DR can be divided into five stages: no DR, mild DR, moderate DR, severe DR and proliferative DR [3]. Part of the blood vessels in the retina will be blocked if the pathological state is maintained for a long time, which will eventually lead to severe visual impairment or even blindness. Therefore, it is important to grade the severity of DR, so that DR patients can receive correct and timely treatment at an early stage.

The blindness caused by DR can be prevented through regular fundus examinations. In clinical diagnosis, DR screening mainly relies on ophthalmologists examining colored fundus images. However, the large number of patients with DR brings a great burden for limited number of ophthalmologists. As the number of diabetic people increases, the amount of fundus images is increasing and becoming more and more difficult to be real-time analyzed manually. Thus, it is necessary to use computer aided diagnosis to reduce the burden on ophthalmologists and examination time, making patients keep abreast of their illness.

CNNs have made great progress in recent years and have been widely applied in the field of computer vision. CNNs integrate feature extraction with classification in an end-to-end manner, and have achieved great breakthroughs in tasks such as image classification [4], object detection [5] and semantic segmentation [6]. Thanks to the powerful capability of high-level feature extraction and representational ability of CNNs, they have also been widely used in medical image analysis tasks such as retinal blood vessel segmentation [7], optic disc segmentation [8] and glaucoma screening [9], [10]. In DR grading, Li *et al.* [11] adopted pre-trained CNNs based on transfer learning for DR grading. Specifically, they considered pre-trained CNNs as feature extractors, and used the outputs of the last fully connected layer as features in combination with a support vector machine (SVM) to solve the grading task. Yang *et al.* [12] proposed a two-stage CNN to detect lesions and grade DR from fundus images, achieving good results. This method solves DR grading in a two-stage way, which is more complex than a one-stage strategy.

Although CNN-based DR grading methods have achieved good results, their practical clinical application is still challenging due to the complexity of the task. First, the five DR grades are very similar in color and texture, and thus it is easy to confuse them in the grading task. This adversely affects the inter-class diversity. Second, some lesions in fundus images are very small and made up of only a few pixels, as shown

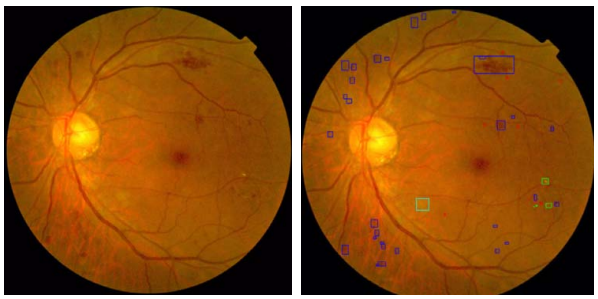
Manuscript received August 15, 2020; accepted September 7, 2020. Date of publication September 11, 2020; date of current version December 29, 2020. This work was supported in part by the National Natural Science Foundation under Grant 61872200, in part by the Natural Science Foundation of Tianjin under Grant 19JCZDJC31600 and Grant 18YFYZCG00060, and in part by the Open Project Fund of the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, under Grant CARCH201905. *(Corresponding author: Kai Wang.)*

Along He, Tao Li, Ning Li, and Kai Wang are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: healong2020@163.com; litao@nankai.edu.cn; lining1994@mail.nankai.edu.cn; wangk@nankai.edu.cn).

Huazhu Fu is with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates (e-mail: hzfu@ieee.org).

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

Digital Object Identifier 10.1109/TMI.2020.3023463



(a) Original Image (b) Annotated Image

Fig. 1. An example of a fundus image with small lesions. The original image and the annotated one are on the left and right, respectively. Green, red, blue, and light blue bounding boxes represent hard exudates, microaneurysms, hemorrhages and soft exudates, respectively.

in Fig. 1. These are likely to be overlooked during convolution and thus are harmful to the final DR grading results. Third, the DR data distribution among different grades is extremely imbalanced since abnormal fundus images only make up a small portion. Fig. 2 shows the data distribution of three DR grading datasets. The imbalanced data distributions will cause the model to focus on DR grades with more samples (DR 0 and DR 2) and ignore those with a small number of samples (DR 1, DR 3 and DR 4), which will affect the generalization ability of the model. Fourth, we expect each channel of the feature maps to represent a certain feature pattern and behave differently on different DR grades, thus benefiting the distinction of different categories. However, some channels in existing CNN models either lack the inter-class diversity, or are redundant in the feature maps [13], which will limit the representational ability of CNNs. These four problems make DR grading remain a challenging task with only image-level supervision.

Motivated by the above observations, we introduce a novel attention module that generates more features for DR grading using only image-level supervision. For the first and second problems, we propose the Global Attention Block (GAB) inspired by [14], which can learn the class-agnostic global attention features and preserve detailed lesion information in fundus images, while suppressing useless information (similar color and texture). For the last two issues, the simple yet effective Category Attention Block (CAB) can learn class-specific features and enlarge the distance between different DR grade levels. More specifically, we assume a fixed number of feature channels to represent each DR grade, and all feature channels belonging to the same class are discriminative, which will reduce the feature channel redundancy. If there is a limited amount of data for a certain DR grade, CAB can put emphasis on this DR grade, and its class-specific feature channels can learn discriminative features, tackling the problem of an imbalanced data distribution. We embed the two proposed blocks into a backbone network, producing a new model for DR grading, called CABNet, which can help CNNs learn discriminative representations for DR grading with few extra parameters. Our main contributions are summarized as follows:

(1) A novel Category Attention Block (CAB) is proposed to explore different discriminative region-wise features for each DR grade in a class-specific way and treats each DR grade equally, reducing the feature redundancy and the impact of an imbalanced data distribution on DR grading. Moreover, CAB focuses on the category attention, which is complementary to the channel and spatial attentions, so CAB can be combined with other non-category attention blocks to further improve the performance of the DR grading.

(2) By combining two complementary blocks, i.e. GAB and CAB, CABNet is proposed for DR grading. CABNet can capture the detailed small lesion features that are helpful for DR grading and alleviate the problem of imbalanced data distributions.

(3) The proposed attention module is a universal plug and play CNN module, which could be utilized in any backbone to improve the DR grading performance significantly.

(4) Extensive experiments are conducted on three public datasets, i.e. DDR, Messidor and EyePACS, to verify the effectiveness of the proposed CABNet. Ablation studies show that CAB not only significantly improves the performance of DR grading, but is also highly complementary to GAB and other attention blocks including SE (Squeeze-and-Excitation), CBAM (Convolutional Block Attention Module) and GC (Global Context). Further, in comparative experiments with other DR grading methods, CABNet achieves the state-of-the-art results for both binary-class and multi-class DR grading tasks.

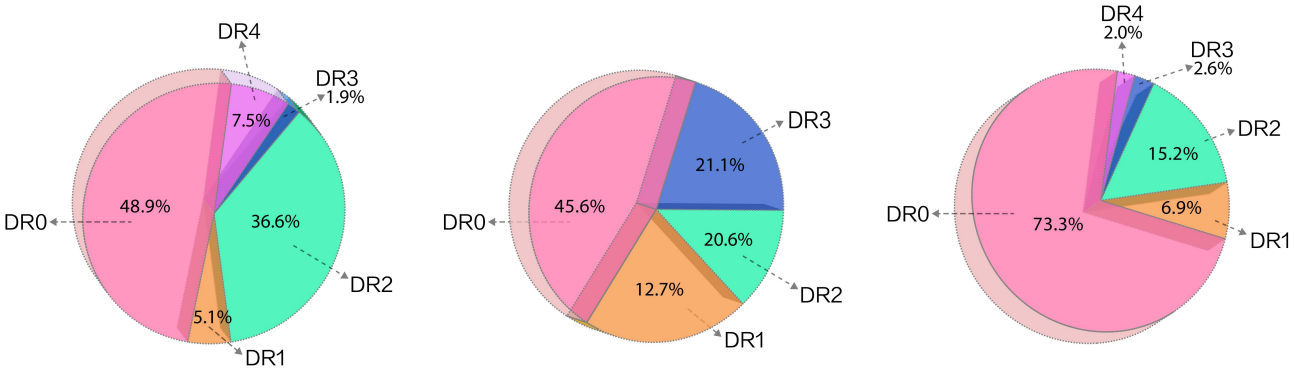
The remainder of this article is organized as follows. In Section II, we analyze related works on DR grading and attention mechanism. The proposed CABNet for DR grading is described in detail in Section III. Extensive experiments are conducted in Section IV to evaluate the DR grading performance of CABNet, and CABNet is compared with other methods on three datasets. The discussion and conclusion are given in Section V and Section VI, respectively.

II. Related Works

In this section, we briefly review the recent works on DR grading and attention mechanisms in deep learning.

A. Deep Learning for DR Grading

In recent years, remarkable progress has been made in the field of medical analysis through the use of deep learning algorithms. Deep learning, especially with CNNs, provides powerful support for DR grading [15-17]. Specifically, a CNN can be trained as a feature extractor in an end-to-end way, which can discern subtle features directly from fundus images without human effort or specific domain knowledge. Van Grinsven *et al.* [18] proposed a selective sampling method to speed-up the training for the detection of hemorrhages in colored fundus images by dynamically selecting misclassified negative samples during training. Vo and Verma [19] designed a new deep network by using multiple filter sizes to learn fine-grained features for DR recognition on two public datasets and achieved good results. Gulshan *et al.* [20] applied



(a) DDR Dataset (b) Messidor Dataset (c) EyePACS Dataset

Fig. 2. The imbalanced data distribution of three DR grading datasets: DDR, Messidor and EyePACS datasets.

Inception-v3 architecture for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus images, achieving good results. Previous works demonstrated that deep learning based methods are effective on DR grading tasks. However, it is still challenging since the small lesions such as microaneurysms and hard exudates are difficult to capture with traditional deep learning models.

B. Attention Mechanism in Deep Learning

Attention mechanisms are helpful to capture the fine-grained features in most computer vision tasks and have been widely used for image classification [21], [22], object localization [23], and semantic segmentation [24], [25]. Zhou *et al.* [26] proposed a multi-cell architecture that gradually increases the depth of the CNNs and the resolution of the input image, which reduces the training time and improves the classification performance. In order to further enhance the screening performance, they utilized attention mechanism for DR screening from fundus images. Wang *et al.* [27] proposed a CNN based algorithm with an attention mechanism for simultaneously diagnosing DR and highlighting suspicious regions based on small high resolution patches with image-level labels. Li *et al.* [28] made attention maps an explicit and natural component of end-to-end training, which can seamlessly bridge the gap between using weak and extra supervision, and their attention module achieved good results in semantic segmentation. In fine-grained image classification, Ding *et al.* [29] applied the Attention Pyramid Convolutional Neural Network with a top-down feature pathway and a bottom-up attention pathway to enhance feature representation and accurately locate discriminative regions. These proposed methods can be trained end-to-end without additional supervision and learn both high-level semantic and low-level detailed feature representations for an input image. However, while they have achieved good results, they do not take imbalanced data distributions into account. Motivated by these previous works, we propose an attention module that aggregates channel-wise, spatial-wise and category-wise attention for fine-grained DR grading. Our attention module is relatively simple but effective and works well on the DR grading task.

Two recent works focused on the issue of imbalanced data distribution. Dai *et al.* [30] proposed a multi-sieving convolutional neural network (MS-CNN) to improve the identification of the microaneurysms regions, which fuses low-level image features and high-level text information of the diagnostic report. The experimental results indicate that it benefits to alleviate the issue of imbalanced data distribution by integrating expert domain knowledge. For imbalanced lesion detection tasks, Zhuang *et al.* [31] proposed CARE to enhance the attention on minor classes, which can improve the classification performance on rare diseases. However, to make CARE more attentive on the lesion regions for minor classes, the extra annotation of bounding boxes is needed for minority classes during training. Compared to MS-CNN and CARE, the proposed CABNet has two advantages. On one hand, CABNet only needs image-level supervision information, but no extra information such as clinical report in MS-CNN and bounding boxes annotation in CARE. On the other hand, an attention module is proposed to handle the imbalanced data distribution in CABNet, which could be used for any backbone to improve the DR grading performance.

III. Methodology

The structure of CABNet, shown in Fig. 3, consists of four parts, i.e. the backbone, GAB, CAB and the classifier. GAB and CAB form the attention module, and CABNet is trained in an end-to-end manner. In this section, we first give an overview of the CABNet, which effectively integrates GAB and CAB to improve the performance of the DR grading. And then we illustrate the proposed GAB and CAB in detail.

A. Overview of CABNet

1) Structure: As shown in Fig. 3, CABNet takes a fundus image as input and the backbone network is used as a feature extractor to obtain the global feature maps. We can adopt any CNNs pre-trained on ImageNet [32] as the backbone network to extract feature maps $F \in R^{H \times W \times C}$ from the last convolutional layer, which contain high-level semantic features of fundus images, where H, W and C denote height, width and number of channels in the feature maps, respectively.

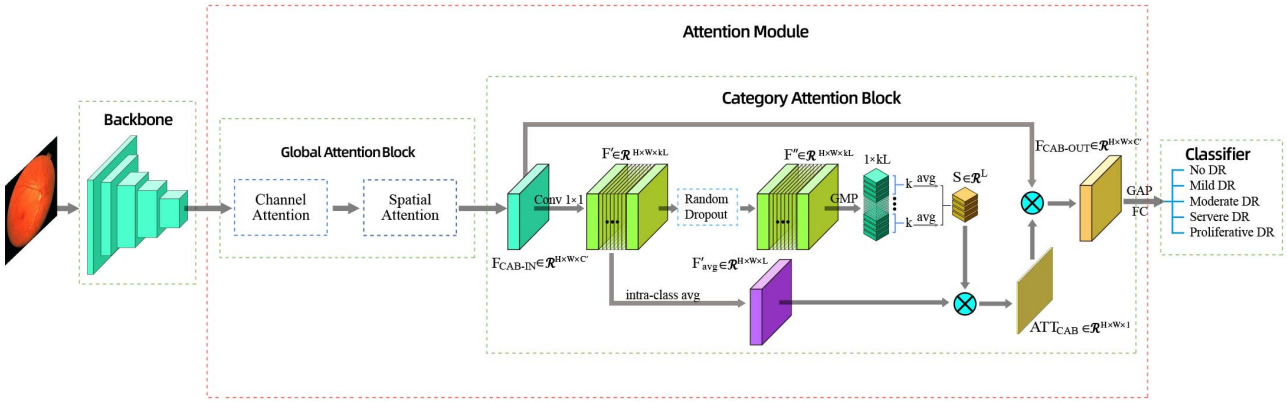
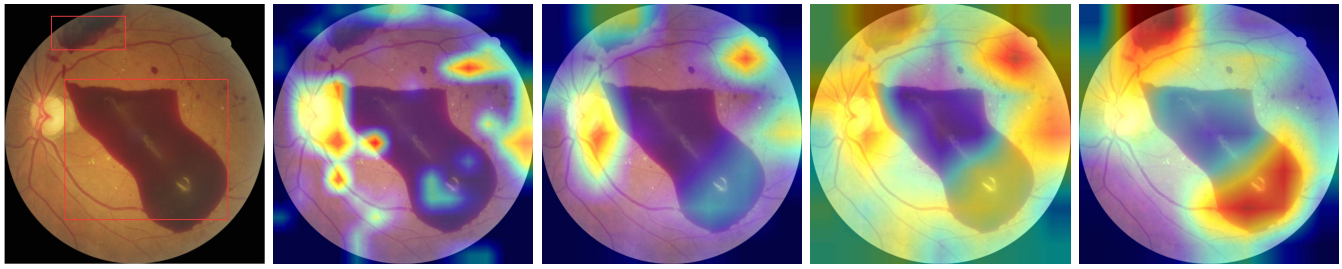


Fig. 3. The overall structure of CABNet. It consists of four parts, i.e. the backbone, the Global Attention Block, the Category Attention Block and a classifier. Note that the feature map from the last convolutional layer of the backbone network is first fed into a 1×1 convolutional layer to reduce the input channels and obtain F_{reduce} (we omit this simple operation in the structure for simplicity). Then, F_{reduce} is taken as the input of GAB, and the output of spatial attention is the input of CAB. Finally, the output of CAB is fed into classifier for DR grading.



(a) Original Image (b) Heatmap w/o Attention (c) Heatmap with CBAM (d) Heatmap with GAB (e) Heatmap with CABNet

Fig. 4. Visualization of heatmaps for a fundus image from DDR dataset. Red bounding boxes indicate the lesion regions.

Next, to reduce the computational cost and memory usage, we apply a 1×1 convolutional layer on F for channel reduction and obtain $F_{reduce} \in R^{H\times W\times C'}$, where $C' = C/2$, which is the input of GAB. Then, the GAB learns channel-wise and spatial-wise attention feature maps $F_{GAB-OUT}$ to preserve more detailed small lesion information and suppress less useful information. Next, $F_{GAB-OUT}$ is taken as the input of CAB, forcing the network to learn different discriminative region-wise features for each DR grade, and producing the output $F_{CAB-OUT}$. CAB is compatible with GAB and they are concatenated in the attention module for CABNet. Finally, we use a global average pooling (GAP) layer and a fully connected (FC) layer to perform the classification task, predicting the class label for each input image.

2) Analysis: The CABNet can be constructed by incorporating GAB and CAB into any backbone network. It has three main characteristics for CABNet.

First, GAB and CAB are two completely different attention blocks, where GAB focuses on the channel attention and spatial attention, and CAB focuses on the category attention. Therefore, by combining the two blocks, CABNet can improve the performance of DR grading.

Second, CABNet uses CAB to generate class-specific attention feature maps in a category-wise manner. Therefore, CABNet can capture more detailed (GAB) and category-wise (CAB) features, which is more suitable for the imbalanced datasets in the DR grading task.

Third, CABNet uses the single-branch GAB designed for the DR grading task, and CAB is applied based on the attention feature maps produced by the GAB. With the better category-agnostic global attention feature maps, CAB brings a significant improvement to DR grading.

To intuitively observe the advantage of CABNet, heatmaps are compared in Fig. 4. We can see that the backbone without attention ignores much lesion information. After adopting CBAM, the network can capture more lesion information, but it still misses some information (the center part). When the backbone is combined with GAB, the network preserves detailed global attention features and thus CABNet can highlight more informative regions for DR grading.

B. Global Attention Block

1) Structure: GAB consists of channel attention and spatial attention, and it adopts a single-branch structure instead of the two-branch structure in CBAM [14]. The experimental results show that compared with CBAM [14] GAB is more suitable for the DR grading task. The structure is shown in Fig. 5. GAB takes the reduced features $F_{reduce} \in R^{H\times W\times C'}$ as input and learns category-agnostic global attention feature maps.

First, we calculate the channel attention feature maps $F_{c\_att} \in R^{H\times W\times C'}$ by the following formula:

$$F_{c\_att} = (\sigma(Conv2(GAP(F_{GAB-IN})))) \otimes F_{GAB-IN}, \quad (1)$$

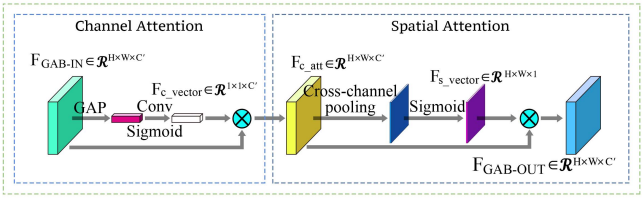


Fig. 5. The overall structure of the Global Attention Block (GAB).

where σ denotes the Sigmoid function, GAP is the global average pooling layer, $Conv2$ indicates two 1×1 convolutional layers, F_{GAB-IN} is the F_{reduce}, and $\otimes$ denotes element-wise multiplication.

Then, we calculate the output of GAB, i.e. the spatial attention feature maps $F_{GAB-OUT}$, by the following formula:

$$F_{GAB-OUT} = F_{c\_att} \otimes (\sigma(C\_GAP(F_{c\_att}))), \tag{2}$$

where C_GAP denotes the cross channel average pooling. $F_{GAB-OUT}$ is taken as the input of CAB, i.e. F_{CAB-IN}, to generate the category attention feature maps.

2) Analysis: For channel attention, it acts as a feature selector in the channel-wise by learning the channel-wise attention weights, which indicates the importance of each feature channel and suppressing less informative channels. For spatial attention, it points out the importance of each spatial position through the learning of spatial attention weights, which is complementary to the channel attention. For the sequential arrangement of the attention module, we put GAB before CAB to extract global detailed lesion information and preserve more small lesion regions, which can reduce the loss of information. The CAB pays more attention to discriminative regions and it refines the features produced by GAB. If we switch the order of the two blocks, CAB will lose some fine detailed information, which will have an adverse effect on the final results.

C. Category Attention Block

1) Structure: The structure of CAB is shown in Fig. 3. CAB is designed to learn the discriminative regions from fundus images to enhance the fine-grained DR grading task with L classes. For an incoming feature map $F_{CAB-IN} \in R^{H\times W\times C'}$, we first feed it to a 1×1 convolutional layer to produce feature maps $F' \in R^{H\times W\times kL}$, where k is the number of channels needed to detect discriminative regions for each class. In order to force all k feature maps within a class to learn different discriminative regions, we randomly remove half of the features during training and set their values to zero, and then we can get $F'' \in R^{H\times W\times kL}$, which retains half of the features from each feature map. During inference, we remove the dropout operation and all features in the k feature maps are used for prediction. Then, we calculate the scores $S = \{S_1, S_2, \ldots, S_L\}$ for each class by the following formula:

$$S_i = \frac{1}{k}\sum_{j=1}^{k} GMP(f''_{i,j}), \quad i \in \{1, 2, .., L\}, \tag{3}$$

where GMP denotes global max pooling, and $f''_{i,j}$ represents the j-th feature map for the i-th class from F''. S_i measures the importance of the feature maps for each class.

Note that CAB should be compatible with the global attention block, so that the DR grading can be better completed by combining the two complementary types of blocks. S is not directly used as the category attention, and a category-wise cross channel average pooling operation is applied on F' to get the feature map for each class:

$$F'_{i\_avg} = \frac{1}{k}\sum_{j=1}^{k} f'_{i,j}, \quad i \in \{1, 2, \ldots, L\}, \tag{4}$$

where $f'_{i,j}$ represents the j-th feature map for the i-th class from F', and $F'_{i\_avg} \in R^{H\times W\times1}$ denotes the semantic feature map for the i-th class.

Then, the category attention $ATT_{CAB} \in R^{H\times W\times1}$ is obtained as follows:

$$ATT_{CAB} = \frac{1}{L}\sum_{i=1}^{L} S_i F'_{i\_avg} \tag{5}$$

ATT_{CAB} highlights the discriminative regions that are informative for DR grading. Finally, the feature maps F_{CAB-IN} can be converted to the feature maps $F_{CAB-OUT}$ by the category attention ATT_{CAB}:

$$F_{CAB-OUT} = F_{CAB-IN} \otimes ATT_{CAB}, \tag{6}$$

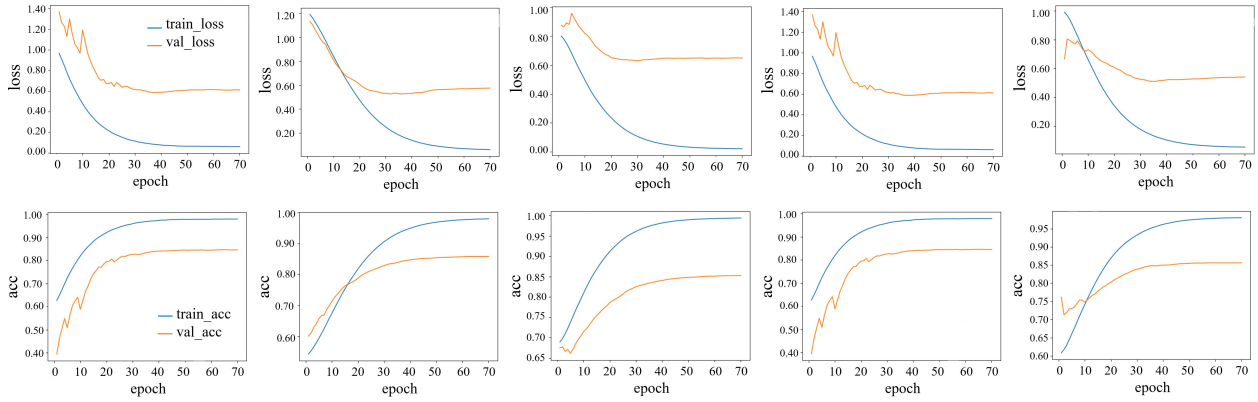
where $\otimes$ denotes element-wise multiplication, and $F_{CAB-OUT}$ is the output feature maps of CAB, which enhances the discriminative regions in F_{CAB-IN} for DR grading.

2) Analysis: CAB has three characteristics.

First, CAB learns the attention in a category-wise manner, and each DR category is treated equally. In traditional CNN, all the feature maps are stacked together without distinction, which may lead to confused information among different categories and less attention on the categories with less samples. For the proposed CAB, it assigns a certain number of feature channels to each DR category, and guarantee each DR grade having the equal feature channels, benefitting to avoid channel bias and enlarge the distance among different DR categories. Therefore, CAB can effectively alleviate the problem of imbalanced data distribution that widely exists in DR grading datasets, as shown in Fig. 2. This is further supported by experimental results.

Second, CAB mines more discriminative regions than other CNNs for each DR category, which reduces the feature redundancy and makes it completely different from other global attention blocks, such as channel attention and spatial attention. Therefore, CAB is complementary to these global attention blocks, and combining them can be beneficial for improving the performance of DR grading.

Third, CAB generates attention features with a small number of parameters, which can reduce the computational cost and memory usage. It integrates the discriminative regions for each category in a single feature map and has the same width and height as the incoming feature maps. Therefore, it is easy to implement the combination of CAB and other global attention blocks.



(a) MobileNet1.0 (b) VGG-16 (c) ResNet-50 (d) DenseNet-121 (e) Xception

Fig. 6. Training curves on DDR dataset (top: loss vs. epoch, bottom: accuracy vs. epoch). train_loss and val_loss denote the losses on training set and validation set, respectively. And train_acc and val_acc denote the accuracies on training set and validation set, respectively. It can be seen from the loss vs. epoch curves, all models can converge within 70 epochs, and we obtain the optimal model with the minimum loss on the validation set.

To intuitively observe the advantage of CAB, heatmaps generated by models without/with CAB are provided in Fig. 4 (d) and (e), respectively. The red bounding boxes indicate lesion regions. As can be seen, the CAB learns discriminative regions that are associated with the lesion information and highlight more accurate and informative regions for DR grading.

IV. Experiments

In this section, we first introduce the commonly used DR grading datasets, implementation details and evaluation metrics, and then present the qualitative and quantitative results of the proposed method on three DR grading datasets.

A. Datasets

DDR DataSet [33]: This dataset contains 13,673 fundus images, including 6,835 training images, 2,733 validation images and 4,105 test images. These images are graded into six classes by seven trained graders according to the International Classification of Diabetic Retinopathy. Poor-quality images without clearly visible lesions are considered ungradable. Thus, the six levels: no DR, mild DR, moderate DR, severe DR, proliferative DR and ungradable. In our experiments, we only focus on the five-class classification task for DR grading; that is, we do not use images belonging to the ungradable class. As a result, the training, validation and test images are 6,320, 2,503 and 3,759, respectively.

Messidor DataSet [34]: There are 1,200 fundus images in this dataset and DR is graded into four classes in terms of the severity, which is inconsistent with international standards. It also contains DME (Diabetic Macular Edema) labels and it is graded into three classes to measure the risk of macular edema. To fairly compare with previous works, we conducted a binary classification for DR grading in this dataset. We classified DR 0 and DR 1 as referable and grouped DR 2 and DR 3 as non-referable, following previous works [19,33].

EyePACS DataSet [36]: This is the largest DR grading dataset, which consists of 35,126 training images and 53,576 test images. Each image is labeled with a DR grade from 0 to 4. The challenge of this dataset is its large variation in resolution, intensity, and quality.

B. Implementation Details

Our CABNet is based on a backbone network pre-trained on a large-scale image dataset. We apply random horizontal flips, vertical flips, and random rotation as forms of data augmentation to reduce overfitting, and the input resolution of our network is 512×512. The initial learning rate is set to 0.005 and is decayed by a factor of 0.8 if the performance on the validation dataset cannot be improved within three epochs. All models are trained for 70 epochs with the Adam optimizer and the cross-entropy loss function. And for each backbone, the model that performs best (has the minimum loss) on the validation set is selected as the final model. Training curves in Fig. 6 demonstrate that the model can converge within 70 epochs. The batch-size is 16 and k in CAB is set to 5. Our framework is implemented with Keras using Tensorflow backend. All the experiments are performed on NVIDIA GTX 1080Ti GPUs with 11 GB of memory.

C. Evaluation Metrics

To evaluate the performance of the proposed method, we employ accuracy, quadratic weighted Kappa score [36], area-under-the-curve (AUC), Precision, Recall, and F1-Score. For binary DR grading task, we use accuracy, AUC, Precision, Recall and F1-Score metrics, and for multi-class DR grading, we use accuracy and quadratic weighted Kappa metrics.

D. Ablation Studies on DDR Dataset

We conduct ablation studies to better understand the impact of each component of CABNet. We first analyze the effects of CAB and GAB on the DDR dataset with MobileNet1.0 [46] as the backbone, and then we discuss the choice of k in CABNet, the effect of CABNet on different imbalanced ratios, and the choice of dropout rate in CABNet, respectively. Finally, we adopt other state-of-the-art CNN architectures as the backbones to evaluate the generality of CABNet.

TABLE I
ABLATION STUDY OF CABNET ADOPTING MOBILENET1.0 AS THE BASELINE ON DDR DATASET. ACC AND KAPPA SCORE ARE REPORTED IN THIS TABLE. '#Para' DENOTES THE NUMBER OF PARAMETERS

Method	Acc	Kappa	δacc/%	#Para
baseline	0.7318	0.6683	-	3.23M
baseline+SE [37]	0.7384	0.7294	↑0.66	4.29M
baseline+CBAM [14]	0.7398	0.6979	↑0.80	4.29M
baseline+GC [38]	0.7368	0.7212	↑0.50	4.29M
baseline+GAB	**0.7422**	**0.7305**	↑**1.04**	4.29M
baseline+CAB	0.7635	0.7333	↑3.17	3.77M
baseline+SE [37]+CAB	0.7640	0.7342	↑3.22	4.31M
baseline+CBAM [14]+CAB	0.7584	0.7431	↑2.66	4.31M
baseline+GC [38]+CAB	0.7717	0.7489	↑3.99	4.31M
baseline+GAB+CAB	**0.7813**	**0.7699**	↑**4.95**	4.31M

1) Analysis of Category Attention Block: We investigate the effectiveness of CAB in Table I. The experimental results indicate that the CAB can improve the performance by a large margin compared with the baseline (the improvement in Acc and Kappa are 3.17% and 0.065, respectively). We also combine CAB with other state-of-the-art attention blocks under the same model parameters for fair comparison. The DR grading results can be improved consistently after incorporating CAB into other attention blocks, which shows the superiority of CAB.

Another advantage of CAB is that the increase in model parameters is small (0.54M), providing a good trade-off between performance and computational cost. In order to conduct a quantitative analysis of the model's inference speed, we perform forward propagation on the DDR test set, which contains 3,759 fundus images with the size of 512×512. The model is evaluated on a NVIDIA GTX 1080Ti GPU and it can process 75 images per second with MobileNet1.0 as backbone, meeting the real-time requirements.

2) Analysis of CABNet: To investigate the relationship between CAB and GAB, we aggregate the two blocks to form the attention module. Compared with baseline+CAB, baseline+CAB+GAB (CABNet) can consistently improve the accuracy and Kappa score, and achieves the best results on the DDR dataset, which demonstrates the effectiveness of the two blocks, and shows that they complement each other.

We also compare our GAB with other attention networks to demonstrate its effectiveness. For fair comparison, we keep the same model parameters for each attention block. The results in Table I show that GAB achieves the best DR grading performance on the DDR dataset compared with SE, CBAM and GC attention blocks, demonstrating its effectiveness. Importantly, while CBAM, with its strong fitting ability, it behaves well on the training set, it leads to overfitting and thus performs slightly worse than GAB on the test dataset.

3) The Choice of k in CABNet: In this part, we analyze the effect of the hyper-parameter k in CABNet, and k is the number of feature channels for each DR category. The DR grading performance of CABNet on DDR dataset with different k are reported in Table II. The results show that the DR grading performance of CABNet is improved with the increasing of k, and it achieves the best grading results when

TABLE II
THE DR GRADING RESULTS OF DIFFERENT k IN CABNET ADOPTING MOBILENET1.0 AS BACKBONE ON DDR VALIDATION SET

	k = 3	k = 4	k = 5	k = 6	k =7
Acc	0.8425	0.8496	**0.8569**	0.8501	0.8483
Kappa	0.8542	0.8632	**0.8794**	0.8678	0.8597

TABLE III
THE RESULTS OF DR GRADING USING DIFFERENT IMBALANCED RATIOS ADOPTING MOBILENET1.0 AS BACKBONE ON DDR DATASET. M IS THE UNION OF DR 0 AND DR2 WITH MORE SAMPLES, AND L IS THE UNION OF DR 1, DR 3 AND DR 4 WITH LESS SAMPLES. #M AND #L DENOTE THE NUMBER OF SAMPLES IN M AND THE NUMBER OF SAMPLES IN L, RESPECTIVELY. THE IMBALANCED RATIO IS REPRESENTED BY (#M/#L):1. R_M AND R_L DENOTE THE RECALL FOR M AND L, RESPECTIVELY

(#M/#L):1	Method	Acc	Kappa	R_M	R_L	δR_M/%	δR_L/%
6:1	baseline	0.7318	0.6683	0.7930	0.3515	-	-
	CABNet	0.7813	0.7699	0.8396	0.4346	↑4.66	↑8.31
7:1	baseline	0.7265	0.6507	0.7906	0.3379	-	-
	CABNet	0.7796	0.7603	0.8363	0.4321	↑4.57	↑9.42
10:1	baseline	0.7194	0.6422	0.7903	0.2899	-	-
	CABNet	0.7723	0.7586	0.8316	0.4172	↑4.13	↑12.73
15:1	baseline	0.7102	0.6312	0.7901	0.2321	-	-
	CABNet	0.7701	0.7505	0.8304	0.4125	↑4.03	↑18.04

k is set to 5. However, when we further increase the value of k, the classification performance of the model decreases, mainly because of overfitting and feature redundancy in CABNet. Therefore, we set k = 5 in our CABNet for better performance.

4) The Effect of CABNet on Different Imbalanced Ratios: To verify the effect of CABNet on different imbalanced data distributions, we make the problem of the imbalanced data distribution more serious by decreasing the the number of the training samples for the categories with less samples. As shown in Fig. 2, for the DDR dataset, there are more samples in DR 0 and DR 2 but less samples in DR 1, DR 3 and DR 4. We use M to denote the union of DR 0 and DR 2 with more samples, and use L to denote the union of DR 1, DR 3 and DR 4 with less samples. The imbalanced ratio is represented by (#M/#L):1, where #M and #L denote the number of samples in M and the number of samples in L, respectively. To increase the imbalanced ratio, we keep the number of training samples in DR 0 and DR 2 unchanged. And for DR 1, DR 3 and DR 4, we use 100%, 80%, 60% and 40% of training samples for model training, which approximately correspond to the imbalanced ratios of 6:1, 7:1, 10:1 and 15:1, respectively. For the fair comparison, we use the same training samples for the baseline and CABNet. As shown in Table III, with the increasing imbalanced ratio, the performance decreases for both baseline and CABNet. However, compared to the baseline, CABNet has a smaller decline. Specifically, compared to the results with the imbalanced ratio of 6:1, when the imbalanced ratio is 15:1, the performance of baseline drops 2.16% in Acc and 0.0371 in Kappa, while CABNet only drops 1.12% in Acc and 0.0194 in Kappa, which demonstrates that the proposed CABNet works well on different imbalanced ratios. Moreover, according to δR_M and δR_L, by using CABNet, the performance improvements on the categories with less samples are more significant than on the

TABLE IV
THE RESULTS OF DR GRADING USING DIFFERENT DROPOUT RATES IN CABNET ADOPTING MOBILENET1.0 AS BACKBONE ON DDR VALIDATION DATASET. 'r' DENOTES THE DROPOUT RATE, AND 'r = 0.0' MEANS ALL FEATURES ARE USED DURING TRAINING WITHOUT DROPOUT

	r = 0.0	r = 0.25	r = 0.5	r = 0.75
Acc	0.8465	0.8545	**0.8569**	0.8521
Kappa	0.8565	0.8738	**0.8794**	0.8645

categories with more samples for all imbalanced ratios. And the significance becomes more obvious when the imbalanced ratio increases, which demonstrates that the proposed CABNet is suitable for extremely imbalanced data.

5) The Choice of Dropout Rate in CABNet: In this part, we analyze the effect of the dropout rate r in CABNet, and the results are reported in Table IV. The results demonstrate that CABNet achieves the best grading results when r is set to 0.5. The feature dropping operation during training has two advantages. On one hand, it can reduce overfitting caused by insufficient training data but too many features. On the other hand, the loss of some information will lead to a larger loss, so the network will force the rest of the feature maps to learn more discriminative features to reduce the loss. As a result, each of the feature maps will learn different discriminative features and thus improve the representation ability of CABNet. This strategy works like a simple model ensemble, which can effectively enhance the DR grading performance of the model.

6) Different Backbones on DDR Dataset: To demonstrate the generality of the proposed CABNet, we adopt different state-of-the-art CNN architectures, which can be grouped into plain networks, residual networks, densely-connected networks, and depth-wise separable convolutional networks. These architectures include VGG-16 [39], ResNet-50 [4], DenseNet-121 [40], and Xception [41].

From Table V, we can see that the baseline model integrated with CAB and GAB blocks can achieve significant performance improvement, while DenseNet-121 with the attention module achieves the best results on the DDR dataset. For VGG-16, the attention block can greatly improve the grading performance, since VGG-16 does not contain BN layers, making it difficult to converge during training. The attention block can solve this problem. The results demonstrate that the proposed attention module can be applied to a wide range of backbone networks and consistently improve the DR grading performance with a small increase in model parameters.

E. Results on Other DR Grading Datasets

In addition to the DDR dataset, we also verify the effectiveness of the proposed method on other DR grading datasets, including Messidor and EyePACS.

1) Results on Messidor Dataset: Due to the limited number of fundus images in the Messidor dataset (1200 images), we only adopt this dataset to test the models trained on the EyePACS dataset, which has the following advantages: On one hand, the problem of overfitting caused by insufficient training data can be avoided; on the other hand, we can verify the

TABLE V
THE RESULTS OF CURRENT STATE-OF-THE-ART CLASSIFICATION NETWORKS ADOPTING PROPOSED ATTENTION BLOCKS ON DDR DATASET

Backbone	Method	Acc	Kappa	δacc/%	#Para
VGG-16 [39]	baseline	0.6288	0.5392	-	14.71M
	baseline+GAB	0.7653	0.7429	↑13.65	15.52M
	baseline+CAB	0.7563	0.7449	↑12.75	15.26M
	CABNet	**0.7701**	**0.7502**	↑**14.13**	15.53M
ResNet-50 [4]	baseline	0.7557	0.7427	-	23.59M
	baseline+GAB	0.7696	0.7573	↑1.39	25.17M
	baseline+CAB	0.7685	0.7504	↑1.28	24.92M
	CABNet	**0.7773**	**0.7857**	↑**2.16**	25.19M
DenseNet-121 [40]	baseline	0.7669	0.7438	-	7.04M
	baseline+GAB	0.7770	0.7726	↑1.01	8.10M
	baseline+CAB	0.7845	0.7783	↑1.76	7.58M
	CABNet	**0.7898**	**0.7863**	↑**2.29**	8.12M
Xception [41]	baseline	0.7494	0.7494	-	20.87M
	baseline+GAB	0.7555	0.7595	↑0.61	22.45M
	baseline+CAB	0.7698	0.7562	↑2.04	22.20M
	CABNet	**0.7757**	**0.7693**	↑**2.63**	22.46M

TABLE VI
THE FOUR-CLASS DR GRADING RESULTS ON MESSIDOR DATASET OF THE PROPOSED METHOD ADOPTING DIFFERENT STATE-OF-THE-ART BACKBONES. THESE NETWORKS ARE TRAINED ON EYEPACS AND TESTED ON THIS DATASET

Backbone	Method	Acc	Kappa	δacc/%
MobileNet1.0	baseline	0.7791	0.7887	-
	baseline+GAB	0.7841	0.7970	↑0.50
	baseline+CAB	0.7916	0.8067	↑1.25
	CABNet	**0.8325**	**0.8408**	↑**5.34**
VGG-16	baseline	0.7625	0.7756	-
	baseline+GAB	0.7975	0.8093	↑3.50
	baseline+CAB	0.8033	0.8123	↑4.08
	CABNet	**0.8275**	**0.8298**	↑**6.50**
ResNet-50	baseline	0.7650	0.7786	-
	baseline+GAB	0.8083	0.8125	↑4.33
	baseline+CAB	0.8016	0.8123	↑3.66
	CABNet	**0.8375**	**0.8456**	↑**7.25**
DenseNet-121	baseline	0.7900	0.8014	-
	baseline+GAB	0.8116	0.8247	↑2.16
	baseline+CAB	0.8283	0.8324	↑3.83
	CABNet	**0.8408**	**0.8723**	↑**5.08**
Xception	baseline	0.7891	0.7985	-
	baseline+GAB	0.8041	0.8124	↑1.50
	baseline+CAB	0.8125	0.8214	↑2.34
	CABNet	**0.8400**	**0.8546**	↑**5.09**

generalization ability of the CABNet trained on one dataset and tested on another. We present the four-class DR grading results on the Messidor dataset in Table VI. The CABNet can consistently improve the DR grading performance, and the model adopting DenseNet-121 as the backbone achieves the best results, with 84.08% Acc and 0.8723 Kappa score.

2) Results on EyePACS Dataset: The EyePACS dataset is a large dataset and it is challenging due to the class imbalance between the DR grades and its large variation in resolution, intensity, and quality. The results reported in Table VII show that the proposed CABNet can consistently improve the DR grading performance on EyePACS compared to the baseline.

In Tables V, VI and VII, we verify the effectiveness of our proposed method on five different backbone networks and three DR grading datasets. The results show that CABNet can be applied to a wide range of backbones and achieve state-of-the-art performance on DDR, Messidor and EyePACS datasets,

TABLE VII
THE DR GRADING RESULTS OF OUR METHOD ADOPTING STATE-OF-THE-ART BACKBONES ON EYEPACS DATASET

Backbone	Method	Acc	Kappa	δacc/%
MobileNet1.0	baseline	0.8421	0.8320	-
	baseline+GAB	0.8460	0.8409	↑0.39
	baseline+CAB	0.8589	0.8576	↑1.68
	CABNet	**0.8668**	**0.8607**	↑**2.47**
VGG-16	baseline	0.8159	0.8102	-
	baseline+GAB	0.8371	0.8212	↑2.12
	baseline+CAB	0.8464	0.8320	↑3.05
	CABNet	**0.8588**	**0.8479**	↑**4.29**
ResNet-50	baseline	0.8285	0.8192	-
	baseline+GAB	0.8339	0.8242	↑0.54
	baseline+CAB	0.8356	0.8302	↑0.71
	CABNet	**0.8589**	**0.8509**	↑**3.04**
DenseNet-121	baseline	0.8261	0.8201	-
	baseline+GAB	0.8482	0.8498	↑2.21
	baseline+CAB	0.8514	0.8535	↑2.53
	CABNet	**0.8618**	**0.8678**	↑**3.57**
Xception	baseline	0.8193	0.8197	-
	baseline+GAB	0.8445	0.8498	↑2.52
	baseline+CAB	0.8594	0.8574	↑4.01
	CABNet	**0.8659**	**0.8673**	↑**4.66**

demonstrating the good generalization ability of the attention module. In terms of the number of parameters and DR grading performance, DenseNet-121 achieves the best result, so we use this as the backbone when compared with other methods.

F. Comparisons With Other State-of-the-Art Methods

1) Binary-Class Task (Messidor Dataset): To further measure the grading performance of our method, we conduct binary DR grading on the Messidor dataset to compare with other state-of-the-art DR grading methods. Note that we do not pretrain our model on the EyePACS dataset for fair comparisons. We can see that for previous works, CANet [44] achieves the best results on the Messidor dataset. This model is designed for joint DR and DME grading, and uses both DR and DME labels for training. Its promising performance in DR grading is a result of its ability to capture the internal relationship between the two diseases. Therefore, we also perform joint training to improve the performance of our CABNet. "Joint training" indicates that we adopt CABNet as the backbone for shared feature extraction and add two individual FC layers for DR and DME grading, respectively. We can see from Table VIII that our method adopting a joint training strategy achieves the best performance (AUC: 96.9%, Acc: 93.1%, F1-Score: 91.5%) on the Messidor dataset, and outperforms the CANet by 0.6%, 0.5% and 0.3% in AUC, Acc and F1-Score, respectively. This shows that the DR and DME grading tasks are complementary and can enhance the learning ability of deep models.

2) Multi-Class Task (EyePACS Dataset): We also compare our method with other state-of-the-art DR grading methods on the EyePACS dataset. The results in Table IX show that our CABNet can achieve state-of-the-art results compared with other methods with only image-level labels. AFN and Zoom-in Net adopt attention mechanisms to learn lesion maps and improve the performance of DR grading. Our method considers not only attention mechanisms, but also imbalanced data distributions, introducing CAB to solve this problem, which greatly enhances the performance. It can be observed from Table IX, our CABNet obtains the best results in terms of Kappa score, which demonstrates its effectiveness.

TABLE VIII
BINARY DR GRADING RESULTS OF DIFFERENT METHODS ON MESSIDOR DATASET (%). "-" INDICATES NO RESULTS REPORTED IN THEIR PAPER. "JT" DENOTES JOINT TRAINING. P, R AND F1 DENOTE PRECISION, RECALL AND F1-SCORE, RESPECTIVELY

Method	AUC	Acc	P	R	F1	Backbone Network
Lesion Based [43]	76.0	-	-	-	-	-
Fisher Vector [43]	86.3	-	-	-	-	-
VNXK [19]	88.7	89.3	-	-	-	-
CKML [19]	89.1	89.7	-	-	-	-
Zoom-in Net [27]	**95.7**	91.1	-	-	-	Inception-Resnet
Dynamic feature [35]	91.6	-	-	-	-	-
CABNet (ours)	94.6	**92.1**	91.0	89.2	89.9	DenseNet-121 (8.12M)
CANet (JT) [44]	96.3	92.6	90.6	**92.0**	91.2	ResNet-50 (29.03M)
CABNet (JT, ours)	**96.9**	**93.1**	**92.9**	90.2	**91.5**	DenseNet-121 (8.12M)

TABLE IX
DR GRADING COMPARISONS WITH OTHER STATE-OF-THE-ART METHODS ON EYEPACS DATASET WITH ONLY IMAGE-LEVEL LABELS

Method	Acc	Kappa
Min-Pooling	-	0.8490
o_O	-	0.8450
RG	-	0.8390
Zoom-in Net [27]	-	0.8540
AFN [45]	-	0.8590
CABNet (ours)	0.8618	**0.8678**

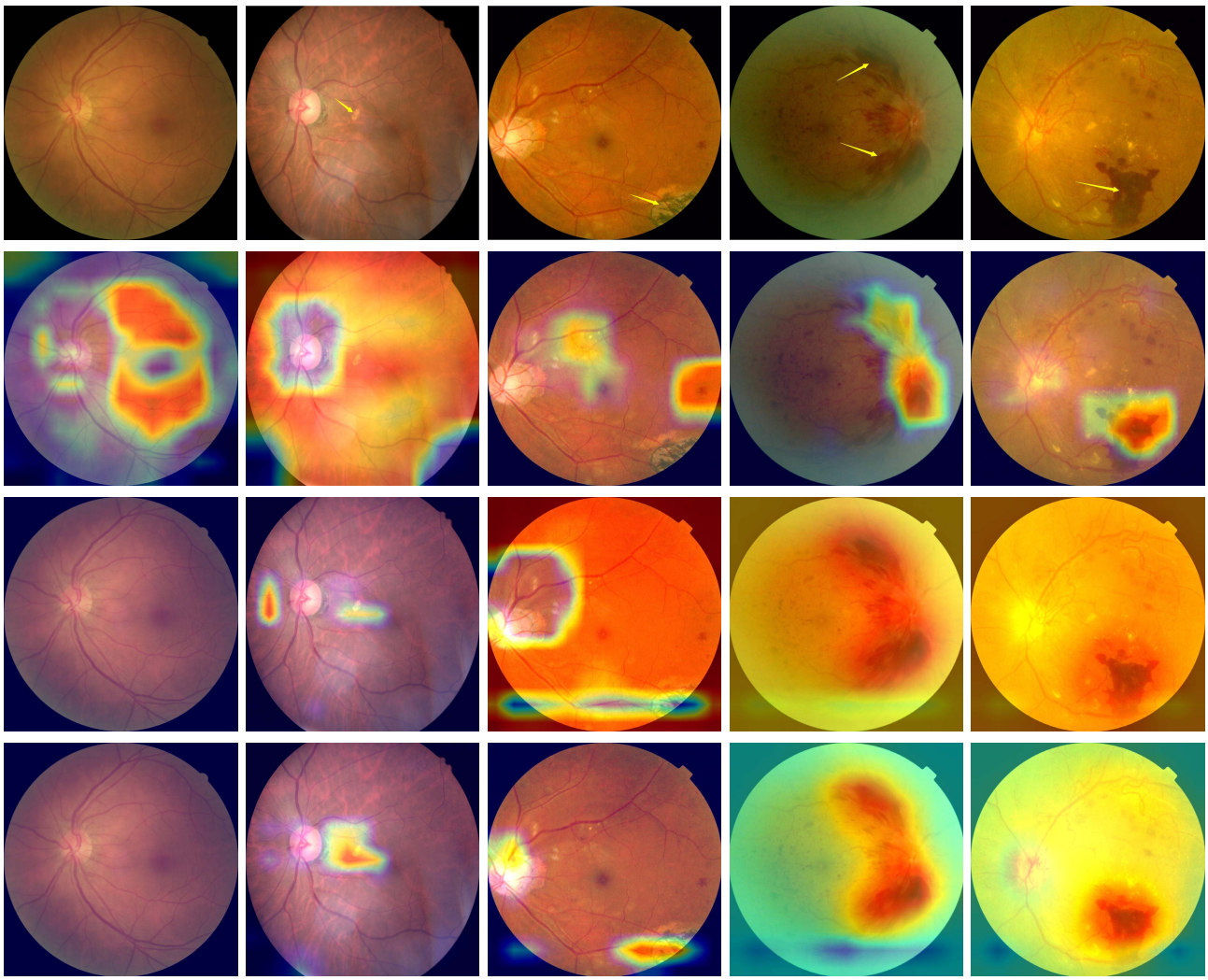
G. Visualization of CABNet

To verify the interpretability of the model and better understand the effect of CAB and CABNet, we visualize the results using Grad-CAM [42]. As shown in Fig. 7, from left to right in the top row, we provide five images from the DDR dataset corresponding to the five severity levels from DR 0 to DR 4, i.e. no DR, mild DR, moderate DR, severe DR, and proliferative DR.

The results of GAB are visualized in the third row, which is better than that generated by the model without attention in the second row. For the model without attention, it may highlight some unrelated regions as shown in the heatmaps of DR 0 and DR 1 in the second row, and it may not cover the lesion regions as shown in the heatmaps of DR 3 and DR 4 in the second row. As can be seen, even though GAB can obtain global attention maps, it still produces some useless features, shown in the second column of the third row. GAB puts more emphasis on global attention features rather than region-wise features.

After being refined by CAB, as shown in the bottom row of Fig. 7, the attention maps can be corrected to focus on some obvious discriminative lesion regions and coarsely locate the suspicious lesion regions, which provides a good interpretability for the classification results and will be helpful for the clinical diagnosis.

Specifically, for the challenging cases such as DR 1 shown in the second column of Fig. 7, the lesion heatmap produced by CABNet can still cover the lesion region even though the lesion is small, demonstrating that CABNet can capture small lesions.



(a) No DR (b) Mild DR (c) Moderate DR (d) Severe DR (e) Proliferate DR

Fig. 7. Some visualization results of GAB and CAB on DDR dataset. We show five DR grade levels (0-4 from left to right, i.e. No DR, Mild DR, Moderate DR, Severe DR and Proliferate DR, respectively). The top row provides original images where yellow arrows indicate the lesion regions. The second row provides the heatmaps without attention, the third row provides the heatmaps of GAB, and the bottom row shows the heatmaps refined by CAB.

V. Discussion

With the development of fundus image acquisition equipment and deep learning algorithms, automatic DR screening has become a hot topic in the field of medical image processing. Though deep learning methods have achieved good performance on the DR grading task, there is still a certain gap for application in clinical practice. In this work, we proposed the CABNet to tackle small lesions and the problem of imbalanced data distributions. In order to increase the interpretability of the model, we further obtain the location map of suspicious-looking lesions in fundus images, so that the results produced by the model can help ophthalmologists make better diagnoses. Experimental results show that our method can be applied to a wide range of backbones and achieve state-of-the-art performance on three public datasets.

The good performance of our method in the DR grading task can be attributed to two main components, as analyzed in the previous experimental section, i.e. the Category Attention Block and CABNet, which combines CAB and GAB. Although our method achieves good performance, there is still room for improvement. First, the whole network is trained with only image-level supervision, making it very challenging to accurately locate some small lesion regions. Second, from the perspective of clinical application, our model can provide a grading score and the coarse location of the suspicious lesion regions, but not the type of DR lesion, such as soft exudate, hard exudate, microaneurysm, and hemorrhage, which is important for DR screening and should thus be addressed in future work.

VI. Conclusion and Future Work

In this article, we present a novel CABNet that combines CAB and GAB. CABNet can be trained in an end-to-end manner for fine-grained DR grading and learn discriminative features by the attention module. Extensive experiments on three datasets demonstrate that CABNet can achieve superior DR grading performance with different backbone networks, which shows the generality of our method. Our future work is to use generative adversarial networks (GANs) for synthesizing high-quality fundus images with labels. This is critical in the medical field since it is expensive to obtain annotated images. We could thus design a more effective model that can

not only provide a grading score, but also indicate the lesion type. By using these synthetic datasets to pretrain the deep model and then fine-tuning on real retinal fundus datasets, we may further improve the DR grading performance.

REFERENCES

[1] N. H. 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] J. Ding and T. Y. Wong, "Current epidemiology of diabetic retinopathy and diabetic macular edema," *Current Diabetes Rep.*, vol. 12, no. 4, pp. 346–354, Aug. 2012.

[3] C. P. Wilkinson *et al.*, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, Sep. 2003.

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

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 91–99.

[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2980–2988.

[7] T. Birgui Sekou, M. Hidane, J. Olivier, and H. Cardot, "From patch to image segmentation using fully convolutional networks—Application to retinal images," 2019, *arXiv:1904.03892*. [Online]. Available: http://arxiv.org/abs/1904.03892

[8] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[9] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, and U. R. Acharya, "Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images," *Inf. Sci.*, vol. 441, pp. 41–49, May 2018.

[10] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018.

[11] 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. CISP-BMEI*, Shanghai, China, Oct. 2017, pp. 1–11.

[12] 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*, Quebec City, QC, Canada, 2017, pp. 533–540.

[13] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," 2019, *arXiv:1911.11907*. [Online]. Available: http://arxiv.org/abs/1911.11907

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

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

[16] W. Zhang *et al.*, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowl.-Based Syst.*, vol. 175, pp. 12–25, Jul. 2019.

[17] J. de la Torre, A. Valls, and D. Puig, "A deep learning interpretable classifier for diabetic retinopathy disease grading," *Neurocomputing*, vol. 396, pp. 465–476, Jul. 2020.

[18] M. J. J. P. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sanchez, "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.

[19] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. ISM*, San Jose, CA, USA, Dec. 2016, pp. 209–215.

[20] 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.

[21] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 4476–4484.

[22] C. Cao *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. ICCV*, Santiago, Chile, Dec. 2015, pp. 2956–2964.

[23] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 2219–2228.

[24] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 3640–3649.

[25] Q. Tang, F. Liu, J. Jiang, and Y. Zhang, "Attention-guided chained context aggregation for semantic segmentation," 2020, *arXiv:2002.12041*. [Online]. Available: http://arxiv.org/abs/2002.12041

[26] K. Zhou *et al.*, "Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading," in *Proc. EMBC*, Honolulu, HI, USA, Jul. 2018, pp. 2724–2727.

[27] 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*, Quebec City, QC, Canada, 2017, pp. 267–275.

[28] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 9215–9223.

[29] Y. Ding *et al.*, "Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," 2020, *arXiv:2002.03353*. [Online]. Available: http://arxiv.org/abs/2002.03353

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

[31] J. Zhuang, J. Cai, R. Wang, J. Zhang, and W. Zheng, "CARE: Class attention to regions of lesion for classification on imbalanced data," in *Proc. MIDL*, 2019, pp. 588–597.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[33] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, Oct. 2019.

[34] E. Decencière *et al.*, "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[35] 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.

[36] *Kaggle Diabetic Retinopathy Detection Competition*. Accessed: Apr. 11, 2020. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[38] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. ICCV*, Seoul, South Korea, Oct. 2019, pp. 1971–1980.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

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

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 1251–1258.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[43] 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 Informat.*, vol. 21, no. 1, pp. 193–200, Jan. 2017.

[44] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, May 2020.

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

[46] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861