A Novel Multi-Scale Network based on Class Attention for Diabetes Retinopathy

HongYu Chen
School of Information Science and Technology
Linyi University
Linyi, China
210854002023@lyu.edu.cn

RongHua Wu
School of Information Science and Technology
Linyi University
Linyi, China
210854002034@lyu.edu.cn

Chen Tao
School of Information Science and Technology
Linyi University
Linyi, China
210854002051@lyu.edu.cn

WenJing Xu
School of Information Science and Technology
Linyi University
Linyi, China
220854042022@lyu.edu.cn

Hui Yu
School of Creative Technologies
University of Portsmouth
Portsmouth, UK
hui.yu@port.ac.uk

HongZhe Liu
Beijing Key Laboratory of Information Service Engineering
Beijing Union University
Beijing, China
liuhongzhe@buu.edu.cn

Cheng Xu
Beijing Key Laboratory of Information Service Engineering
Beijing Union University
Beijing, China
xc-f4@163.com

MuWei Jian*
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
*Corresponding author
jianmuweihk@163.com

***Abstract*—Diabetes Retinopathy (DR) is a common eye disease, which brings irreversible blindness risk to patients in severe cases. Due to the scarcity of professional ophthalmologists, developing computer-aided diagnostic systems to participate in DR grading diagnosis has become increasingly important. However, the current mainstream deep learning methods still cannot accurately classify the severity of DR, and their unreliable results are difficult to serve as a reference for clinicians. To tackle this problem, we propose two novel modules to improve the accuracy of DR classification. Specifically, we designed a multi-scale feature extraction module (MFEM) to capture tiny lesions in fundus images and differentiate similar lesions simultaneously. In addition, we also created a class attention module (CAM) to alleviate the adverse impact of intra-class similarity on DR grading. Experiment on the APTOS2019 blind detection dataset show that our proposed two modules have made significant improvements to the designed model, achieving state-of-the-art performance with 95.98% for ACC and 97.12% for QWK, respectively.**

Keywords—DR grading, Multi-scale, Attention mechanism, Fundus images

I. INTRODUCTION

Diabetes Retinopathy (DR) is one of the main causes of visual impairment for adults worldwide [1]. Since the beginning of the new millennium, the prevalence and blindness rate of diabetes retinopathy have risen rapidly. According to the 2003 International Clinical DR Classification System [2, 3], the severity of DR can be divided into the following stages: No DR, Non-Proliferative DR (NPDR), and proliferative DR (PDR), among which NPDR can be further separated into mild, moderate, and severe DR. The main means of DR diagnosis is to screen patients' color fundus images by professional doctors, identify abnormal lesions in color fundus images (such as microaneurysms, bleeding, exudate and neovascularization), and then judge the severity of DR, in which each severity has corresponding special lesions and diagnostic criteria. Fig. 1 (a) and (b) show the schematic diagrams of fundus images and the morphology of abnormal lesions in the images of five types of DR, respectively. However, due to the time-consuming and laborious process of DR screening, experienced doctors are required to carefully assess fundus images. Limited by medical conditions, most DR patients are unable to receive prompt treatment, which increases the probability of blindness.

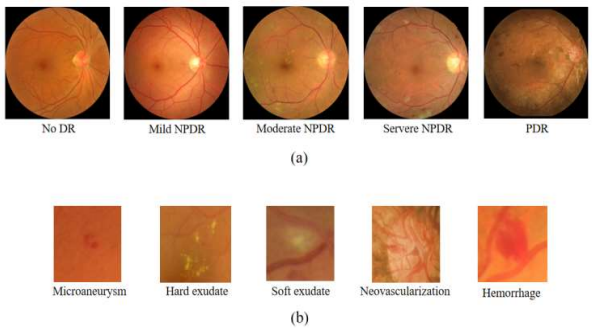

Fig. 1. Examples of fundus images at different stages of diabetes retinopathy and their pathological morphology.

In the past twenty years, an increasing number of researchers have devoted themselves to developing computer-aided diagnosis (CAD) systems to improve the process of DR screening, which can decrease the burden on doctors and provide a second objective opinion, reducing subjectivity in diagnosis [4-9].

Although the CAD system has attained impressive outcomes in DR grading, it remains challenging in clinical practice because of the numerous types of lesions and complex diagnostic criteria. Firstly, in fundus images with higher DR

severity, despite there are lesions different from those in images with lower DR severity, there are also similar lesions, which affects the DR grading task by intra-class similarity and eventually leads to poor performance. Secondly, some lesions (such as microaneurysms) are overly tiny in the fundus image, some of which are only a few pixels in size in the image, making it difficult for the model to detect them. Once missed, it may result in incorrect DR severity classification. Finally, there are visual similarities in shape and color between some lesions or between lesions and normal tissue (such as punctate bleeding and microaneurysms, neovascularization and common blood vessels), and the model is highly likely to confuse them during feature extraction, thereby making for incorrect diagnosis.

Based on the aforementioned issues, we have designed a novel network model for five class of DR grading. Inspired by the process that clinicians use to diagnose DR, that is, to zoom in/out the image to observe the lesions more carefully, we proposed a multi-scale feature extraction module (MFEM), which uses the dilated convolutions with different sizes to extract additional features on the basis of the backbone network, and fuses the extracted features on feature maps of different sizes, We believe that multi-scale features will help the model identify tiny lesions in fundus images and reduce the impact of visual similarity between lesions. At the top of the model, we designed a novel class attention module (CAM) that effectively solves the first issue mentioned above. we first split the feature maps of the last layer of the model into filters with the same number of categories. Then, we separately enhance each filter with self-attention and constrain them by introducing triplet loss to make the features within each filter closer, At the same time, it can also increase the differences between them. Finally, the feature map is subjected to a cross-attention like similarity calculation with each filter, and the calculation results are re-weighted onto the filters. The filters are flattened through global average pooling (GAP) and fully connected (FC) to obtain the outcomes of DR grading.

II. METHODOLOGY

Here, we first provide an overview of the proposed model, which integrates MFEM and CAM, effectively improving the performance of DR grading. Then, we give detailed explanations for each module in the model.

A. Overview of Model

As shown in Fig. 2, Our model takes fundus images as input, uses pre-trained ResNet50 [11] as backbone, and additionally designs a multi-scale architecture to extract features simultaneously with backbone to acquire feature map $F_{m-scale}$ that combines high-level semantic features and multi-scale features. Among them, the multi-scale architecture consists of five diffusion convolution layers (DCLs) responsible for feature extraction of images of different resolutions. Next, $F_{m-scale}$ will enter the CAM for attention calculation to better alleviate the adverse effects of intra-class similarity and obtain the output feature map F_{att}. Finally, we apply the GAP layer and FC layer to perform classification tasks and predict the labels for each fundus image.

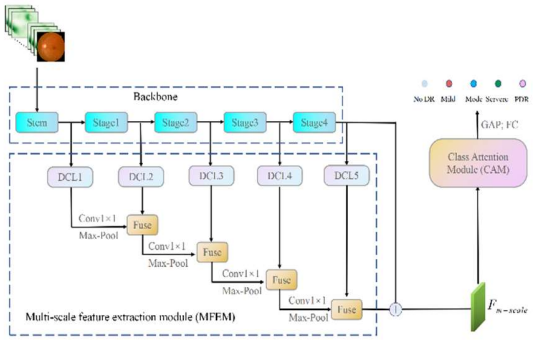

Fig. 2. Overview of the proposed models. The stem and stage 1 to 4 in backbone are both basic structures in ResNet50, where the stem contains a 7×7 convolution. Stage 1 to 4 use two 1×1 convolution kernels and a 3×3 convolution as the basic architecture, and is constructed in a ratio of 3:4:6:3.

B. Multi-scale Feature Extraction Module (MFEM)

Based on observations, we found that clinicians usually conduct scaling operations on fundus images during the diagnosis of DR to observe small lesions and distinguish similar lesions. We designed MEFM to simulate this process. Specifically, MEFM contains five diffusion convolution layers (DCLs), and the internal structure of DCLs is shown in Fig. 3. It will further extract multi-scale features from feature maps F_{S1}, F_{S2}, F_{S3}, F_{S4}, F_{S5} of different resolutions output by stem and stage 1 to 4, respectively.

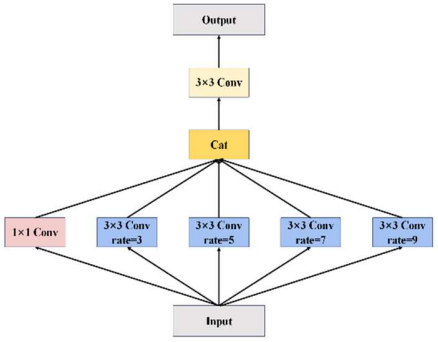

Fig. 3. The structure of diffusion convolution layer.

To start with, the F_{S1} from stem is used as the input of the first DCL, and multi-scale feature extraction is carried out through 3×3 convolution with different dilation rates (different receptive field) and a 1×1 convolution respectively. Then attain feature maps F_1, F_3, F_5, F_7, F_9 at different scales.

Next, we concatenate F_1, F_3, F_5, F_7, and F_9 according to channel dimensions, and use a 3×3 convolution to reduce the number of channels to the same as F_{S1}, thus obtaining a feature map F_{DCL1} with multi-scale information.

Repeat the above operation, we execute multi-scale feature extraction on feature maps of different resolutions to get F_{DCL2},

F_{DCL3}, F_{DCL4} and F_{DCL5}. Finally, integrate multi-scale information from each DCL and fuse it with F_{S5} extracted by backbone with high-level features, namely:

$$g(F'_{DCLi-1}) = f_{1\times1}(MaxPool(F'_{DCLi-1})), \quad (1)$$

$$F'_{DCLi} = [g(F'_{DCLi-1}); F_{DCLi}],$$

$$\text{w.r.t } F'_{DCL1} = F_{DCL1}, \quad (2)$$

$$F_{m-scale} = F_{S5} \oplus f_{1\times1}(F'_{DCL5}), \quad (3)$$

where F'_{DCLi} is the result of adjacent two different DCL concatenated feature maps, and i >1. $\oplus$ denotes element-wise summation.

From this, we obtained the final output $F_{m-scale}$ of MFEM, which contains rich multi-scale information and will be used as input for attention calculation in CAM.

C. Class Attention Module (CAM)

So as to better address the negative impact of intra-class similarity on DR grading, we have created a novel attention module called CAM, as shown in Fig. 4, with the aim of alleviate the problem of category confusion in the classification process.

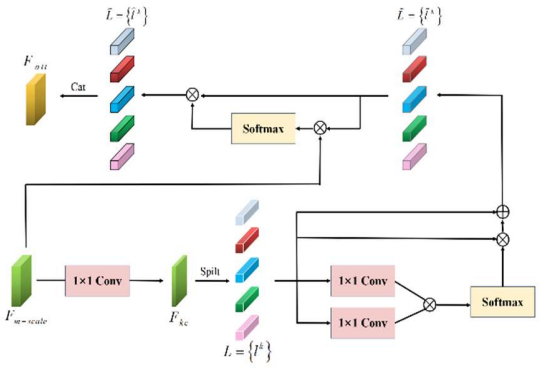

Fig. 4. The structure of class attention module.

Firstly, we decrease the number of channels in multi-scale feature map $F_{m-scale} \in \mathbb{R}^{H\times W\times C}$ to $K\times C'$ by using a convolution with a kernel size of 1, resulting in $F_{kc} \in \mathbb{R}^{H\times W\times KC'}$. K and C' are two hyperparameter, representing the number of categories and the number of channels allocated for each category. Then, split F_{kc} into K filters $L = \{l^k\}$, where $k \in \{1,2,3,4,5\}$, and each $l^k \in \mathbb{R}^{H\times W\times C'}$ represents one of the categories.

Next, we perform inter-class attention calculations on each l^k to make the features in each filter more compact by obtaining contextual information. Taking l^1 as an example, we only used two convolutions with kernel size 1 and obtained the context matrix $M \in \mathbb{R}^{HW\times HW}$ through softmax activation:

$$M = Softmax(f'_{1\times1}(l^1) \otimes f''_{1\times1}(l^1)), \quad (4)$$

where $f'_{1\times1}$ and $f''_{1\times1}$ denote two different 1×1 convolutions.

Fusion matrix M and l^1 to gain a feature dense filter $\tilde{l}^1$:

$$\tilde{l}^1 = l^1 \oplus (l^1 \otimes M), \quad (5)$$

where $\oplus$ represents element-wise summation, and $\otimes$ denotes matrix multiplication.

The calculation process of l^2, l^3, l^4 and l^5 is the same as formulas (4) and (5), and the final result is $\tilde{L} = \{\tilde{l}^k\}$.

To compensate for the potential loss of key features caused by channel reduction on $F_{m-scale}$, we use $F_{m-scale}$ as the key and $\tilde{L} = \{\tilde{l}^k\}$ as the query for attention calculation, and remapping the outcomes to the filters to obtain $\hat{L} = \{\hat{l}^k\}$. Taking $\tilde{l}^1$ as an example, the formula is as follows:

$$\hat{l}^1 = \tilde{l}^1 \otimes (Softmax(\tilde{l}^1 \otimes F_{m-scale})). \quad (6)$$

According to formula (6), we will further obtain $\hat{l}^2$, $\hat{l}^3$, $\hat{l}^4$, and $\hat{l}^5$. Finally, average each $\hat{l}^k$ by channel dimension and concatenate them to acquire category attention feature maps $F_{att} \in \mathbb{R}^{H\times W\times K}$:

$$F_{att} = [\hat{l}^1_{avg}; \hat{l}^2_{avg}; \hat{l}^3_{avg}; \hat{l}^4_{avg}; \hat{l}^5_{avg}], \quad (7)$$

where $\hat{l}^k_{avg}$ is the channel-wise average of the *k-th* filter.

III. EXPERIMENTS

Principally, we first introduce the composition of the dataset. Then, we compare the experimental results with other most advanced models, and verify the progressiveness of the proposed model.

A. Dataset

We used the publicly accessible Kaggle competitive dataset APTOS 2019 blindness detection dataset [10] to train and test the proposed model. This dataset includes 3662 fundus images, all of which are jointly labeled by multiple professional doctors.

Ulteriorly, we removed the blurred images from the dataset, and remaining 3545 for the experiment. They are marked as five levels in the database (i.e. no DR, mild DR, moderate DR, severe DR, PDR). Table I lists the tangible divisions of the data samples.

TABLE I Dataset Sample Partitioning

	No DR	Mild	Moderate	Severe	PDR
Train	1354	278	750	145	222
Test	451	70	180	40	55

B. *Experimental Results*

We validated the performance of the model on the APTOS2019 dataset and compared it with other representative methods, including Inception-V4, ConvNext, EfficientNet, Triple DRNet, etc.

TABLE II Comparison results with other representative methods.

Models	Metric	
	ACC	QWK
Inception-V4 [13]	0.7626	0.7880
EfficientNet [15]	0.8819	0.9081
ConvNext [14]	0.8379	0.8727
Simple-method [12]	0.8480	0.9013
Triple-DRNet [9]	0.9208	0.9362
Ours	**0.9598**	**0.9712**

As shown in Table II, our devised network achieves the best results in terms of ACC and QWK metrics. EfficientNet have excellent feature extraction capabilities, but due to the complexity of lesions in fundus images, they cannot accurately classify them. ConvNext is another advanced CNN model in recent years, which once outperformed Transformer-style models in terms of performance. However, own to its use of depthwise convolution to decrease the amount of parameters, the feature representation ability is weakened, resulting in poor performance in DR grading. Inception-V4 better captures tiny lesions in images by extracting multi-scale features, but in DR classification tasks, it does not address the impact of intra-class similarity. Triple-DRNet divides five types of DR into multiple independent subtasks and categorizes severity from coarse to fine, alleviating the effect of intra-class similarity in some categories, but not extending to all categories. Our proposed model not only relieves the adverse influences of intra-class similarity in all categories, but also extracts more tiny lesions in fundus images and reduces the possibility of confusing similar lesions, thereby improving classification performance.

IV. CONCLUSION

In this paper, we propose a model with two novel modules - MFEM and CAM. This model uses ResNet50 as the backbone to extract features from the image. At the same time, MFEM will further obtain multi-scale information in the image, allowing the model to more accurately capture tiny lesions in the fundus and avoid confusing similar lesions. In addition, we designed CAM and placed it at the top of the model to alleviate the adverse effect of intra-class similarity on DR grading. On the APTOS 2019 dataset, our proposed model is superior to other mainstream models in terms of ACC and QWK.

Furthermore, we found that the overall parameter quantity of the network is relatively large, and it may still be difficult to apply in practical clinical diagnosis. In the future, exploring the balance between lightweight and excellent performance of the model will be our main research content.

ACKNOWLEDGMENT

This work was supported, Taishan Young Scholars Program of Shandong Province; and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

REFERENCES

[1] The Royal College of Ophthalmologists, “Diabetic Retinopathy Guidelines,” Technical Report, 2012.

[2] T. Scully, “Diabetes in numbers,” Nature, vol. 485, no. 7398, pp. S2–S3, May 2012.

[3] C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, and C. D. Agardh, “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,” Ophthalmology, vol. 110, no. 9, pp. 1677–1682, Sep. 2003.

[4] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, “Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning,” Investigative Ophthalmology & Visual Science, vol. 57, no. 13, pp. 5200–5206, Oct. 2016.

[5] T. Araújo, G. Aresta, L. Mendonça, S. Penas, and C. Maia, “DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images,” Medical Image Analysis, vol. 63, p. 101715, Jul. 2020.

[6] E. Saleh, J. Błaszczyński, A. Moreno, A. Valls, P. Romero-Aroca, S. de la Riva-Fernández and R. Słowiński, “Learning ensemble classifiers for diabetic retinopathy assessment,” Artificial Intelligence in Medicine, vol. 85, pp. 50–63, Apr. 2018.

[7] G. Mahendran and R. Dhanasekaran, “Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms,” Computers & Electrical Engineering, vol. 45, pp. 312–323, Jul. 2015.

[8] Z. Han, M. Jian, G. Wang, “ConvUNeXt: An efficient convolution neural network for medical image segmentation,” Knowledge-Based Systems, vol. 253, no. 11, p. 109512, Oct. 2022.

[9] M. Jian, H. Chen, C. Tao, X. Li, and G. Wang, “Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images,” Computers in Biology and Medicine, vol. 155, p. 106631, Mar. 2023.

[10] B. Graham, “Kaggle Diabetic Retinopathy Detection competition report,” University of Warwick, pp. 24-26, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[12] A. Sugeno, Y. Ishikawa, T. Ohshima, and R. Muramatsu, “Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning,” Computers in Biology and Medicine, vol. 137, p. 104795, Oct. 2021.

[13] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng and D. R. Webster, “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.

[14] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

[15] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proceedings of the 36th International Conference on Machine Learning, PMLR, May 2019, pp. 6105–6114.