International Journal of Network Dynamics and Intelligence

Article

Multi-Scale Class Attention Network for Diabetes Retinopathy Grading

Hongyu Chen [1], Ronghua Wu [1], Chen Tao [1], Wenjing Xu [1], Hongzhe Liu [3], Cheng Xu [3], and Muwei Jian [1,2,*]

[1] School of Information Science and Technology, Linyi University, 276000, China
[2] School of Computer Science and Technology, Shandong University of Finance and Economics, 250014, China
[3] Beijing Key Laboratory of Information Service Engineering, 100101, Beijing, China
[*] Correspondence: jianmuweihk@163.com

Received: 1 June 2023
Accepted: 21 July 2023
Published: 26 June 2024

Abstract: Diabetes retinopathy (DR) is a universal eye disease, which brings irreversible blindness risks to patients in severe cases. Due to the scarcity of professional ophthalmologists, it has become increasingly important to develop computer-aided diagnostic systems for DR grading diagnosis. However, the current mainstream deep learning methods face challenges in accurately classifying the severity of DR, making it difficult for them to provide a reliable reference for clinicians. To tackle this problem, we propose two novel modules to improve the accuracy of DR classification. Specifically, we design a multi-scale feature extraction module to capture tiny lesions in fundus images and differentiate similar lesions simultaneously. In addition, we develop a class attention module to alleviate the adverse impact of intra-class similarity on DR grading. Experimental results show that our proposed modules attain significant performance improvement on the APTOS2019 blind detection dataset, with accuracy and quadratic weighted Kappa metrics achieving 95.98% and 97.12%, respectively.

Keywords: Diabetes retinopathy grading; multi-scale; attention mechanism; fundus images

1. Introduction

Diabetes retinopathy (DR) is a complication caused by diabetes and is one of the main causes of visual impairment in adults worldwide [1]. Since the beginning of the new millennium, the prevalence rate and blindness rate of diabetes retinopathy have risen rapidly. According to the prediction of the World Health Organization (WHO), the number of DR patients will grow to 552 million by 2030, and DR will become the main cause of blindness among people of working age [2].

According to the 2003 international clinical DR classification system [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 1) screen patients' color fundus images by professional doctors; 2) identify abnormal lesions in color fundus images (such as microaneurysms, bleeding, exudate, and neovascularization); and 3) judge the severity of DR where each severity has corresponding special lesions and diagnostic criteria. Figure 1(a) and Figure 1(b) show, respectively, the schematic diagrams of fundus images and the morphology of abnormal lesions in the images of five types of DR. In the no DR stage, patients have no obvious lesions. In the pathological stage, the severity increases in a progressive manner. In the mild stage of DR, patients only exhibit microaneurysms caused by leakage from retinal microvessels. Without intervention, mild DR will further progress to moderate DR characterized by a small amount of bleeding (dot and blot haemorrhages) in the fundus images. This progress may also be accompanied by hard exudates (HE), resulting from the aggregation of lipids, proteins, and lipoproteins. Severe DR is the most serious stage of NPDR, and there will be much more microaneurysms and bleeding in the eyes at this stage than at the mild and moderate DR stages. Moreover, such a stage may be accompanied by white or gray soft exudate. PDR is the most severe stage among the five categories of DR, and its most obvious lesion is the appearance of neovascularization generated by vitreous contraction in the fundus image. At this stage, patients usually experience acute visual loss until complete blindness. In clinical practice, timely identification of abnormal lesions from fundus images, accurate classification of the severity of DR, and targeted treatment of DR are the keys to avoiding blindness in patients. 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.

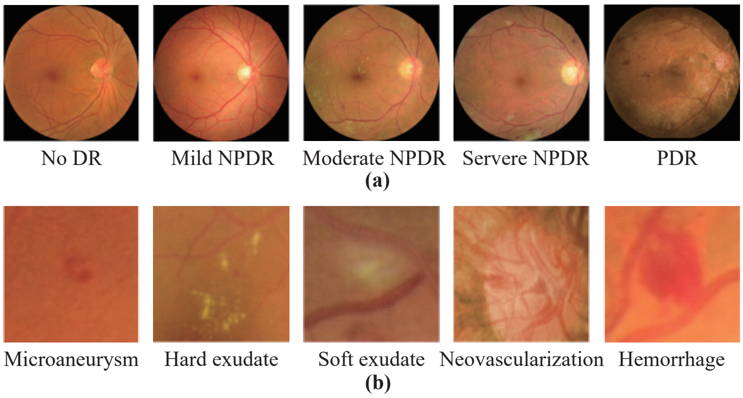

Figure 1. Examples of fundus images at different stages of diabetes retinopathy(DR) and their pathological morphology.

In the past twenty years, research on computer-aided diagnosis (CAD) systems has developed rapidly for improving the process of DR screening [4, 5]. This helps to decrease the burden on doctors, provide a second objective opinion, and reduce subjectivity in diagnosis [6, 7]. Early researchers have tried to design models by manually extracting features. For example, Saleh et al. [8] have chosen to use a fuzzy random forest and domain-based rough set balanced rule ensemble to extract features from fundus images and estimate the risk of developing DR. Mahendran et al. [9] have used a neighbourhood-based segmentation technique to detect lesion from the low contrast images, and proposed the probabilistic neural network and support vector machine (SVM) to assess the severity of the disease, ultimately achieving good results. However, manual feature extraction relies heavily on prior knowledge and is not effective in handling complex situations that arise in images. This limits the performance of CAD systems.

The development of convolutional neural networks (CNNs) has enabled to design more efficient CAD systems [10, 11]. The convolutional kernel in CNNs can effectively extract features from images without depending on prior knowledge, which has gained increasing attention. For instance, Xu et al. [12] have used eight convolutional layers to calculate the image, sequentially obtained low-level and high-level features in the image, and finally performed a five-class DR grading through a fully connected layer, achieving the accuracy of 94.5% on private datasets. In order to further improve the performance and efficiency of the model, Li et al. [13] have combined the transfer learning technology, loaded the pre-trained weights on the large-scale dataset into CNNs, and carried out fine-tuning training on the fundus image dataset, finally reaching excellent classification accuracy.

Although the CNN-based 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 lesions in images with lower DR severity, there are also similar lesions affecting the DR grading task by intra-class similarity. This eventually leads to poor performance. Secondly, some lesions (e.g. 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 is the visual similarity in shape and color among some lesions or between lesions and normal tissues (such as punctate bleeding and microaneurysms, neovascularization, and common blood vessels), and the model is very likely to be confused during feature extraction, leading to false diagnosis.

Based on the aforementioned issues, we design a novel network model for five classes of DR grading. Inspired by the process (of zooming in/out the image to observe the lesions carefully) used by clinicians for DR diagnosis, we propose a multi-scale feature extraction module (MFEM). The MFEM 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 can 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 design 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 subject to the cross-attention like similarity calculation at 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) layers to obtain the outcomes of DR grading. A preliminary paper has been presented in the proceedings of UIC 2023 [14].

The contributions of this article are summarized as follows.

1) We propose a novel CAM that separates feature maps into learnable filters and introduces triplet loss constraints. This module can effectively solve the problem of difficult classification of certain categories in DR grading tasks caused by intra-class similarity.

2) Inspired by the habit of clinicians who scale images to examine lesions during diagnosis, we design an additional MFEM in the model to alleviate ignorance of tiny lesions as well as the adverse grading effects caused by visual similarity between lesions.

3) Extensive experiments conducted on the publicly available APTOS 2019 blindness detection dataset show that our method obtains satisfactory results and the state-of-the-art performance in DR grading with five categories.

2. Related Work

In this section, we mainly review the closely related technologies and introduce the previous work on DR grading tasks.

2.1. Multi-scale Architecture

In the development process of CNNs, researchers have gradually explored the three elements that effectively improve the performance of network models, namely depth, width, and resolution. The multi-scale architecture is an improved structure of the model based on resolution. Zhang et al. [15] have proposed a multi-scale input network, whose main characteristic is to input images of different resolution into different subnetworks and fuse the outputs of each subnetwork in a cascaded manner, thereby obtaining results with multi-scale information. In addition, multi-scale ideas can also be implemented within the model, with the most influential architecture being the inception module designed by Szegedy et al. [16]. The inception module uses three convolutional layers with different kernel sizes and a 3 × 3 maximum pooling layer is used for feature extraction in parallel. Then, the features extracted from these four branches are fused as inputs for the next layer, thereby achieving multi-scale information within the model. Moreover, researchers find that multi-scale information can also be fused in the model prediction phase. Based on this, Liu et al. [17] have built a multi-scale feature prediction fusion network, which uses VGG as the backbone network. After each convolution layer, a classifier is added to get the prediction outcomes under different scale feature maps, and then the final prediction value is acquired by adopting a fast non-maximum suppression algorithm. The advantage lies in reaching a balance between the efficiency and performance of the model.

2.2. Attention Mechanism in Deep Learning

In recent years, attention mechanisms have become progressively important in deep learning, as their essence is to enable models to process more important information when computing resources are restricted [18]. In 2017, the transform model has been proposed by Vaswani et al. [19] to solve machine translation tasks. Later, Wang et al. [20] have combined the attention mechanism with convolution and designed a non-local convolution module to overcome the shortcomings of traditional CNNs which can only perform local operations in feature extraction. Such a module achieves impressive performance in the field of image and video processing. Although this type of attention mechanism further enhances the performance of the model, it brings about excessive computational complexity. To tackle this issue, researchers have designed numerous variants of attention mechanisms [21–23] to find a balance between the computational speed and the model performance. Additionally, Hu et al. [24] have found a unique path by modelling channel relationships in feature maps to automatically obtain the importance of each channel, highlight useful features for the task and suppress useless features. This method is simple and effective, achieving outstanding effects in solving image processing tasks. Also, this method inspires further development of variant attention mechanisms [25–28].

2.3. DR Grading

Due to the vigorous development of deep learning, especially CNNs, a large number of CAD systems for DR screening have been developed to provide strong support for DR grading tasks. Krause et al. [29] have used inception V4 to automatically detect lesions in fundus images and predict the severity of DR. However, accounting for the complexity of DR grading tasks, the performance of solely using CNN models for prediction is not ideal. Therefore, researchers have applied many different methods to improve the model to adapt to DR classification. For example, Sugeno et al. [30] have employed a Laplacian filter with a core size of 5 on the fundus image, filtered the blurred image and deleted it by calculating the standard deviation output by the Laplace operator. The clear fundus image is used as a dataset in the EfficientNet model for DR grading. Compared with previous work, this method significantly improves the performance of the model. Bellemo et al. [31] have combined the idea of ensemble learning and used two different models (VGG and ResNet) to perform DR classification and prediction on fundus images. The final prediction result is determined by averaging the prediction scores of the two models, which leads to better outcomes than the single CNN model on a private dataset with 76370 fundus images. Notwithstanding the method of integrating multiple models can improve the prediction accuracy, it may lead to excessive parameters of the overall model. To tackle this issue, Tymchenko et al. [32] have proposed different methods. Specifically, a single model is used as the feature extractor, and three classifiers are added at the top of the model to conduct supervised learning by different loss functions. Finally, the predicted values of different classifiers are integrated to gain classification results. This method effectively alleviates the drawbacks brought by multiple models, while achieving similar performance. The above research indicates that the method based on CNNs works well for DR classification, but the accuracy of the model still cannot meet the needs of clinical diagnosis practice. With that in mind, He et al. [33] and Li et al. [34] both have designed special attention mechanisms for DR grading tasks, and improved the accuracy of the model in DR grading to a new level. Recently, Wu et al. [35] and Jian et al. [36] have divided the DR grading task into different subtasks, and screened the severity of DR from coarse to fine. This novel method is more in line with the clinical diagnosis process of experts in reality.

The above work makes outstanding contributions to solving the DR grading task, but most of them fail to achieve high-precision results. In this article, we 1) present a novel category attention module for intra-class similarity in DR grading; and 2) propose an MFEM for extremely tiny and similar lesions in fundus images to improve the accuracy of DR grading and provide more reliable references for clinical diagnosis in reality.

3. Methodology

We first provide an overall overview of the proposed model, which integrates MFEM and CAM to effectively improve the performance of DR grading. Then, we furnish detailed explanations for each module in the model.

3.1. Overview of Model

As shown in Figure 2, our model takes fundus images as the input, uses pre-trained ResNet50 as the backbone, and additionally designs a multi-scale architecture to extract features simultaneously with the backbone. Then, the feature map $F_{m-scale}$ is acquired which combines high-level semantic features and multi-scale features. Among them, the multi-scale architecture consists of five diffusion convolution layers (DCLs) and is responsible for feature extraction of images of different resolution. 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.

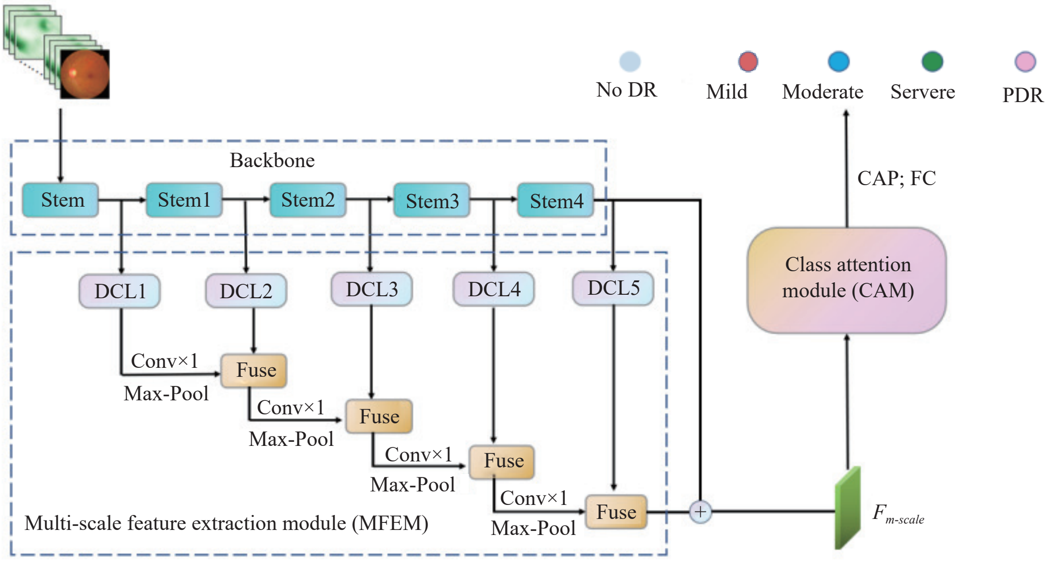

Figure 2. Overview of the proposed models. The stem and stage 1 to 4 in backbone are both basic structures in ResNet50, where the stem contains a 7 × 7 convolution. Stage 1 to 4 use two 1×1 convolution kernels and a 3 × 3 convolution as the basic architecture, and is constructed in a ratio of 3: 4:6:3.

3.2. Multi-Scale Feature Extraction Module (MFEM)

Based on observations, we find that clinicians usually conduct scaling operations on fundus images during the diagnosis of DR to observe small lesions and distinguish similar lesions. We design a multi-scale feature extraction module (MFEM) to simulate this process. Specifically, the MFEM contains five DCLs, and the internal structure of the DCL is shown in Figure 3. The MFEM will further extract multi-scale features from feature maps F_{S1}, F_{S2}, F_{S3}, F_{S4}, F_{S5} of different resolution which are output by the stem and stages 1 to 4, respectively.

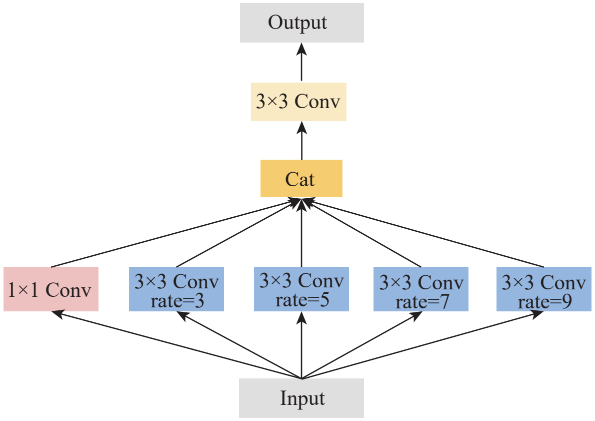

Figure 3. The structure of diffusion convolution layer.

To start with, the F_{S1} from the stem is used as the input of the first DCL, and multi-scale feature extraction is carried out through 1 × 1 convolution and 3 × 3 convolution with different dilation rates (different receptive fields). Then, feature maps F_1, F_3, F_5, F_7, F_9 are obtained at different scales. The formula is expressed as follows:

$$f_i(F_{S1}) = \begin{cases} F_3, i=3 \\ F_5, i=5 \\ F_7, i=7 \\ F_9, i=9 \end{cases}, \tag{1}$$

$$F_1 = f_{1\times1}(F_{S1}), \tag{2}$$

where f_t represents 3 × 3 convolution with different rates, and $f_{1\times1}$ denotes 1×1 convolution.

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

$$F_{DCL1} = \sigma(BN(f_{3\times3}([F_1;F_3;F_5;F_7;F_9]))), \tag{3}$$

where $f_{3\times3}$ denotes a 3 × 3 convolution, σ is ReLU activation function, and BN represents batch normalization.

Based on Equations (1)-(3), we execute multi-scale feature extraction on feature maps of different resolution to get F_{DCL2}, F_{DCL3}, F_{DCL4} and F_{DCL5}. Finally, we integrate multi-scale information from each DCL and fuse the information with F_{S5} extracted by the backbone with high-level features, namely:

$$g\left(F'_{DCLi-1}\right) = f_{1\times1}\left(\text{MaxPool}\left(F'_{DCLi-1}\right)\right), \tag{4}$$

$$F'_{DCLi} = \left[g\left(F'_{DCLi-1}\right); F_{DCLi}\right], w.r.t\ F'_{DCL1} = F_{DCL1}, \tag{5}$$

$$F_{m-scale} = F_{S5} \oplus f_{1\times1}\left(F'_{DCL5}\right), \tag{6}$$

where F'_{DCLi} is the result of two adjacent DCLs' with concatenated feature maps, and i>1. ⊕ denotes element-wise summation.

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

3.3. CAM

To address the negative impact of intra-class similarity on DR grading, we develop a novel CAM as shown in

Figure 4, with the aim to alleviate the problem of category confusion in the classification process.

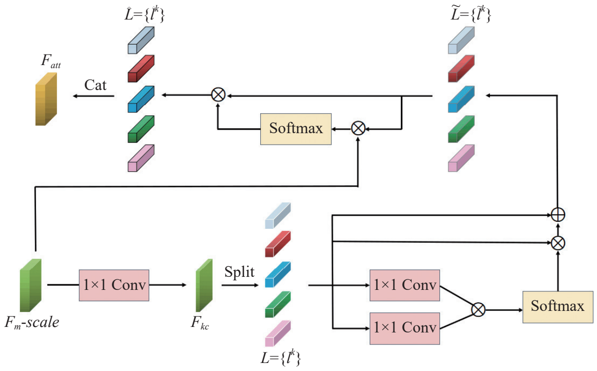

Figure 4. The structure of the class attention module (CAM).

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 convolution with a kernel size of 1, resulting in $F_{kc} \in \mathbb{R}^{H\times W\times KC'}$. K and C' are two hyperparameters representing the number of categories and the number of channels allocated for each category. Then, split F_{kc} into K filters $L=\left\{l^k\right\}$, 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 use two convolution layers with kernel size 1 and obtain the context matrix $M\in\mathbb{R}^{HW\times HW}$ through softmax activation:

$$M = Softmax\left(f'_{1\times1}\left(l^1\right)\otimes f''_{1\times1}\left(l^1\right)\right), \tag{7}$$

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

The fusion matrix M and l^1 are used to gain a feature dense filter $\tilde{l}^1$:

$$\tilde{l}^1 = l^1 \oplus \left(l^1 \otimes M\right), \tag{8}$$

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 that in Equations (7) and (8), and the final result is $\tilde{L}=\left\{\tilde{l}^k\right\}$.

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

$$\widehat{l}^1 = \tilde{l}^1 \otimes \left(\text{Softmax}\left(\tilde{l}^1 \otimes F_{m-tcalb}\right)\right). \tag{9}$$

According to Equation (9), we will further obtain $\widehat{l}^2$, $\widehat{l}^3$, $\widehat{l}^4$, and $\widehat{l}^5$. Finally, average each $\widehat{l}^k$ by the channel dimension and concatenate them to acquire category attention feature maps $F_{att}\in\mathbb{R}^{H\times W\times K}$:

$$F_{att} = [\widehat{l}^1_{avg};\widehat{l}^2_{avg};\widehat{l}^3_{avg};\widehat{l}^4_{avg};\widehat{l}^5_{avg}] \tag{10}$$

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

4. Experiment

We first introduce the composition of the dataset and the selection of evaluation indicators. Then, we compare the experimental results with other most advanced models, and verify the progressiveness of the proposed model. Finally, we conduct a series of ablation studies to demonstrate the effectiveness of each module.

4.1. Dataset

We use the publicly accessible Kaggle competitive dataset, i.e. the APTOS 2019 blindness detection dataset [37], to train and test the proposed model. This database has two parts: a training set and a testing set. This training set includes 3662 fundus images, all of which are jointly labelled by multiple professionals.

Ulteriorly, we remove the blurred images from the dataset, and use 3545 images for the experiment. These images are marked as five levels in the database (i.e. no DR, mild DR, moderate DR, severe DR, PDR). Table 1 lists the tangible divisions of the data samples.

Table 1 Dataset Sample Partitioning

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

4.2. Evaluation Metrics

For the five-class DR grading, we adopt accuracy (ACC) and quadratic weighted Kappa metrics (QWK) to evaluate the validity of the designed models which are further compared with other typical models. ACC and QWK are defined as follows:

$$ACC = \frac{(TP+TN)}{(TP+TN+FP+FN)}, \tag{11}$$

$$QWK = 1 - \frac{\sum_{i,j} \varpi_{i,j} P_{i,j}}{\sum_{i,j} \varpi_{i,j} \tilde{P}_{i,j}}, \tag{12}$$

$$w.r.t\ \varpi_{i,j} = \frac{(i-j)^2}{(C-1)^2}, \tag{13}$$

where TP and TN denote the number of correctly classified positive samples and correctly classified negative samples, respectively. FP and FN are, respectively, the number of negative samples wrongly classified as positive and the number of positive samples wrongly classified as negative. $P_{i,j}$ and $\tilde{P}_{i,j}$ are the observed and expected probabilities.C denotes the total number of categories. i and j represent certain classes.

4.3. Implementation Details

In this work, we use ResNet50 [38] without the classifier as the backbone, and the size of the input image is adjusted to 448 × 448.

During the training phase, we use the Adam optimizer, while the batch size is set to 16. The initial learning rate is adjusted to 1e-3 and dynamically decreased. Our framework is implemented based on Python 3.8 and PyTorch 1.10.0. All experiments are conducted on two NVIDIA RTX 3090, 24G GPUs.

4.4. Experimental Results

We validate the performance of the model on the APTOS 2019 dataset and present the results in Figure 5. As shown in Figure 5, our model achieves results comparable to those of clinicians in the five classes of DR grading. Among them, excellent differentiation has been achieved for easily confused categories such as no DR/mild DR, mild DR/moderate DR, and moderate DR/severe DR.

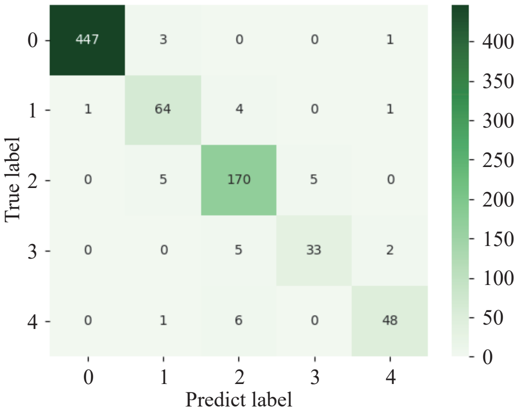

Figure 5. Confusion matrix of diabetes retinopathy(DR) grading outcomes.

Moreover, we compare the proposed model with other representative methods for further evaluation. These methods include Inception-V4, ConvNext, EfficientNet, CBANet, Triple-DRNet, etc.

As shown in Table 2, our devised network achieves the best results in terms of ACC and QWK metrics. ResNext and EfficientNet both have excellent feature extraction capabilities, but due to the complexity of lesions in fundus images, they cannot accurately classify different lesions. ConvNext is another advanced CNN model which outperforms transformer-style models in terms of performance. However, due to its use of depthwise convolution to decrease the number of parameters, the feature representation ability is weakened, resulting in poor performance in DR grading. Both SE-ResNet and CBANet introduce attention mechanisms to improve the performance of the backbone, but fundamentally, they do not pay better attention to tiny lesions, making it difficult to further improve DR grading performance. Inception-V4 can better capture 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. This alleviates the effect of intra-class similarity in some categories, but cannot be extended to all categories. The proposed model not only relieves the adverse influence of intra-class similarity in all categories, but also extracts more tiny lesions in fundus images to reduce the possibility of confusing similar lesions, thereby improving classification performance.

Table 2 Comparison results with other representative methods

Models	Metric	
	ACC	QWK
Inception-V4	0.7626	0.7880
ResNext [39]	0.8681	0.9024
EfficientNet [40]	0.8819	0.9081
SE-ResNet [41]	0.8178	0.8620
ConvNext [42]	0.8379	0.8727
Simple-method	0.8480	0.9013
CBANet	0.8869	0.9282
Triple-DRNet	0.9208	0.9362
Ours	**0.9598**	**0.9712**

4.5. Ablation Study

To evaluate the effectiveness of the MFEM and CAM, we conducted a series of ablation studies where ResNet50 is used as the backbone. The research results are shown in Table 3.

Table 3 Ablation studies of the impact of different modules on model performance

Backbone	Method	Metric	
		ACC	QWK
ResNet50	baseline	0.8631	0.9075
	baseline + MFEM	0.8867	0.9212
	baseline + CAM	0.9373	0.9416
	baseline + CAM + MFEM	**0.9598**	**0.9712**

As we can observe from Table 3, both modules can gradually improve the performance of the model. Wherein, after adding CAM, the results of DR classification are improved significantly. This indicates that our designed CAM greatly alleviates the adverse impact of intra-class similarity on the model, making the differentiation between categories more pronounced. The MFEM module provides the model with the ability to identify tiny lesions and reduce the possibility of confusing similar lesions, further improving the performance of DR grading.

5. Conclusion

In this article, we have proposed a model with two novel modules, the MFEM and CAM. This model uses ResNet50 as the backbone to extract features from the image. At the same time, the MFEM will further obtain multi-scale information in the image, allowing the model to 1) capture tiny lesions more accurately in the fundus and 2) avoid confusing similar lesions. In addition, we have designed the CAM at the top of the model to alleviate the adverse effect of intra-class similarity on DR grading. Experiments conducted on the APTOS 2019 dataset have shown that the proposed model is superior to other mainstream models in terms of ACC and QWK. The effectiveness of the MFEM and CAM has been verified through ablation studies.

In future research, we plan to develop a lightweight multi-scale feature extraction model that can be more conveniently applied to clinical diagnosis.

Author Contributions: Hongyu Chen: conceptualization, visualization, methodology, software, writing —

original draft preparation; **Ronghua Wu**: validation, visualization, methodology, software, writing—original draft preparation; **Chen Tao**: validation, visualization, methodology, software, writing—original draft preparation; **Wenjing Xu**: formal analysis, visualization, methodology, software, writing—original draft preparation; **Hongzhe Liu**: data curation, visualization, investigation; **Cheng Xu**: supervision, software, validation; **Muwei Jian**: supervision; validation, funding acquisition, methodology, project administration, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Taishan Young Scholars Program of Shandong Province and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

Data Availability Statement: Preprocessed data are available on request from the first author.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: We would like to thank the editor and anonymous reviewers for their helpful remarks.

References

1. The Royal College of Ophthalmologists. Diabetic Retinopathy Guidelines. Technical Report, 2012.
2. Scully, T. Diabetes in numbers. *Nature*, **2012**, *485*: S2−S3.
3. Wilkinson, C.P.; Ferris, F.L.; Klein, R.E.; *et al*. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, **2003**, *110*: 1677−1682.
4. Jian, M.W.; Wang, J.J.; Yu, H.; *et al*. Integrating object proposal with attention networks for video saliency detection. *Inf. Sci.*, **2021**, *576*: 819−830.
5. Yin, Y.C.; Han, Z.M.; Jian, M.W.; *et al*. AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation. *Comput. Biol. Med.*, **2023**, *162*: 107120.
6. Abràmoff, M.D.; Lou, Y.Y.; Erginay, A.; *et al*. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest. Ophthalmol. Visual Sci.*, **2016**, *57*: 5200−5206.
7. Araújo, T.; Aresta, G.; Mendonça, L.; *et al*. DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Med. Image Anal.*, **2020**, *63*: 101715.
8. Saleh, E.; Błaszczyński, J.; Moreno, A.; *et al*. Learning ensemble classifiers for diabetic retinopathy assessment. *Artif. Intell. Med.*, **2018**, *85*: 50−63.
9. Mahendran, G.; Dhanasekaran, R. Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms. *Comput. Electr. Eng.*, **2015**, *45*: 312−323.
10. Jian, M.W.; Wu, R.H.; Chen, H.Y.; *et al*. Dual-branch-UNet: A dual-branch convolutional neural network for medical image segmentation. *Comput. Model. Eng. Sci.*, **2023**, *137*: 705−716.
11. Han, Z.M.; Jian, M.W.; Wang, G.G. ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowl. Based Syst.*, **2022**, *253*: 109512.
12. Xu, K.L.; Feng, D.W.; Mi, H.B. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. *Molecules*, **2017**, *22*: 2054.
13. Li, X.G.; Pang, T.T.; Xiong, B.; et al. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *2017 10th International Congress on Image and Signal Processing*, *BioMedical Engineering and Informatics* (*CISP-BMEI*), *Shanghai*, *China*, *14–16 October 2017*; IEEE: New York, 2017; pp. 1–11. doi:10.1109/CISP-BMEI.2017.8301998
14. Chen, H. A novel multi-scale network based on class attention for diabetes retinopathy. In *The 20th IEEE International Conference on Ubiquitous Intelligence and Computing* (*UIC 2023*), UK.
15. Zhang, Y.; Lv, P.H.; Lu, X.B.; *et al*. Face detection and alignment method for driver on highroad based on improved multi-task cascaded convolutional networks. *Multimed. Tools Appl.*, **2019**, *78*: 26661−26679.
16. Szegedy, C.; Liu, W.; Jia, Y.Q.; et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Boston*, *MA*, *USA*, *7–12 June 2015*; IEEE: New York, 2015; pp. 1–9. doi:10.1109/CVPR.2015.7298594
17. Liu, W.; Anguelov, D.; Erhan, D.; et al. SSD: Single shot MultiBox detector. In *14th European Conference on Computer Vision*, *Amsterdam*, *The Netherlands*, *11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37. doi:10.1007/978-3-319-46448-0_2
18. Niu, Z.Y.; Zhong, G.Q.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing*, **2021**, *452*: 48−62.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *Long Beach*, *CA*, *USA*, *4 December 2017*; Curran Associates: Red Hook, NY, USA, 2017; pp. 6000–6010. doi:10.5555/3295222.3295349
20. Wang, X.L.; Girshick, R.; Gupta, A.; et al. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *UT*, *USA*, *18–23 June 2018*; IEEE: New York, 2018; pp. 7794–7803. doi:10.1109/CVPR.2018.00813
21. Huang, Z.L.; Wang, X.G.; Huang, L.C.; et al. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, *Seoul*, *Korea* (*South*), *27 October 2019–2 November 2019*; IEEE: New York, 2019; pp. 603–612. doi:10.1109/ICCV.2019.00069
22. Zhu, Z.; Xu, M.D.; Bai, S.; et al. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, *Seoul*, *Korea* (*South*), *27 October 2019–2 November 2019*; IEEE: New York, 2019; pp. 593–602. doi:10.1109/ICCV.2019.00068
23. Cao, Y.; Xu, J.R.; Lin, S.; et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision Workshop, *Seoul*, *Korea* (*South*), *27–28 October 2019*; IEEE: New York, 2019; pp. 1971–1980. doi:10.1109/ICCVW.2019.00246

24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *UT*, *USA*, *18–23 June 2018*; IEEE: New York, 2018; pp. 7132–7141. doi:10.1109/CVPR.2018.00745
25. Li, X.; Wang, W.H.; Hu, X.L.; et al. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Long Beach*, *CA*, *USA*, *15–20 June 2019*; IEEE: New York, 2019; pp. 510–519. doi:10.1109/CVPR.2019.00060
26. Woo, S.; Park, J.; Lee, J.Y.; et al. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision* (*ECCV*), *Munich*, *Germany*, *8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19. doi:10.1007/978-3-030-01234-2_1
27. Wang, X.L.; Sun, Y.; Ding, D.R. Adaptive dynamic programming for networked control systems under communication constraints: A survey of trends and techniques. *Int. J. Netw. Dyn. Intell.*, **2022**, *1*: 85−98.
28. Yu, N.X.; Yang, R.; Huang, M.J. Deep common spatial pattern based motor imagery classification with improved objective function. *Int. J. Netw. Dyn. Intell.*, **2022**, *1*: 73−84.
29. Krause, J.; Gulshan, V.; Rahimy, E.; *et al*. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, **2018**, *125*: 1264−1272.
30. Sugeno, A.; Ishikawa, Y.; Ohshima, T.; *et al*. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Comput. Biol. Med.*, **2021**, *137*: 104795.
31. Bellemo, V.; Lim, Z.W.; Lim, G.; *et al*. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study. *Lancet Digital Health*, **2019**, *1*: e35−e44.
32. Tymchenko, B.; Marchenko, P.; Spodarets, D. Deep learning approach to diabetic retinopathy detection. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, *Valletta*, *Malta*, *22–24 February 2020*; ICPRAM, 2020; pp. 501–509.
33. He, A.L.; Li, T.; Li, N.; *et al*. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Trans. Med. Imaging*, **2021**, *40*: 143−153.
34. Li, X.M.; Hu, X.W.; Yu, L.Q.; *et al*. CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE Trans. Med. Imaging*, **2020**, *39*: 1483−1493.
35. Wu, Z.; Shi, G.L.; Chen, Y.; *et al*. Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. *Artif. Intell. Med.*, **2020**, *108*: 101936.
36. Jian, M.W.; Chen, H.Y.; Tao, C.; *et al*. Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images. *Comput. Biol. Med.*, **2023**, *155*: 106631.
37. Graham, B. Kaggle Diabetic Retinopathy Detection Competition Report. University of Warwick, pp. 24-26, 2015.
38. He, K.M.; Zhang, X.Y.; Ren, S.Q.; et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *NV*, *USA*, *27–30 June 2016*; IEEE: New York, 2016; pp. 770–778. doi:10.1109/CVPR.2016.90
39. Xie, S.N.; Girshick, R.; Dollár, P.; et al. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Honolulu*, *HI*, *USA*, *21–26 July 2017*; IEEE: New York, 2017; pp. 5987–5995. doi:10.1109/CVPR.2017.634
40. Tan, M.X.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, *Long Beach*, *CA*, *USA*, *9–15 June 2019*; ICML, 2019; pp. 6105–6114.
41. Chai, R.M.; Chen, D.; Ma, X.; et al. Diabetic retinopathy diagnosis based on transfer learning and improved residual network. In *2022 IEEE 11th Data Driven Control and Learning Systems Conference* (*DDCLS*), *Chengdu*, *China*, *3–5 August 2022*; IEEE: New York, 2022; pp. 941–946. doi:10.1109/DDCLS55054.2022.9858557
42. Liu, Z.; Mao, H.Z.; Wu, C.Y.; et al. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *New Orleans*, *LA*, *USA*, *18–24 June 2022*; IEEE: New York, 2022; pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167

Citation: Chen, H.; Wu, R.; Tao, C.; *et al*. Multi-Scale Class Attention Network for Diabetes Retinopathy Grading. *International Journal of Network Dynamics and Intelligence*. 2024, 3(2), 100012. doi: 10.53941/ijndi.2023.100012