Electronic Research Archive

https://www.aimspress.com/journal/ERA

ERA, 33(2): 725–743.
DOI: 10.3934/era.2025033
Received: 15 November 2024
Revised: 03 January 2025
Accepted: 09 January 2025
Published: 11 February 2025

Research article

FF-ResNet-DR model: a deep learning model for diabetic retinopathy grading by frequency domain attention

Chang Yu[1], Qian Ma[2], Jing Li[3], Qiuyang Zhang[4], Jin Yao[4,*], Biao Yan[5,*] and Zhenhua Wang[1,*]

[1] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
[2] General Hospital of Ningxia Medical University, Ningxia 750001, China
[3] Eye Institute and Department of Ophthalmology, Eye and ENT Hospital, Fudan University, Shanghai 201114, China
[4] Department of Ophthalmology and Optometry, The Affiliated Eye Hospital, Nanjing Medical University, Nanjing 210029, China
[5] Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China

* **Correspondence:** Email: jinyao@126.com, yanbiao@sjtu.edu.cn, zh-wang@shou.edu.cn.

Abstract: Diabetic retinopathy (DR) is a major cause of vision loss. Accurate grading of DR is critical to ensure timely and appropriate intervention. DR progression is primarily characterized by the presence of biomarkers including microaneurysms, hemorrhages, and exudates. These markers are small, scattered, and challenging to detect. To improve DR grading accuracy, we propose FF-ResNet-DR, a deep learning model that leverages frequency domain attention. Traditional attention mechanisms excel at capturing spatial-domain features but neglect valuable frequency domain information. Our model incorporates frequency channel attention modules (FCAM) and frequency spatial attention modules (FSAM). FCAM refines feature representation by fusing frequency and channel information. FSAM enhances the model's sensitivity to fine-grained texture details. Extensive experiments on multiple public datasets demonstrate the superior performance of FF-ResNet-DR compared to state-of-the-art models. It achieves an AUC of 98.1% on the Messidor binary classification task and a joint accuracy of 64.1% on the IDRiD grading task. These results highlight the potential of FF-ResNet-DR as a valuable tool for the clinical diagnosis and management of DR.

Keywords: diabetic retinopathy (DR); grading model; classification model; color fundus image; attention mechanism

1. Introduction

Diabetic retinopathy (DR) is a leading cause of irreversible vision loss worldwide and a severe complication of diabetes. DR is classified into two primary stages: non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [1]. NPDR can further be subdivided into mild, moderate, and severe stages, characterized by features such as microaneurysms, hard exudates, and cotton wool spots [2]. PDR, the more advanced stage, involves the growth of abnormal blood vessels, which can lead to severe vision loss. Accurate DR grading is crucial for clinical diagnosis, as it helps doctors assess the type, quantity, and distribution of DR lesions, determine disease severity, and develop suitable treatment strategies. Figure 1 illustrates a schematic diagram of DR grading.

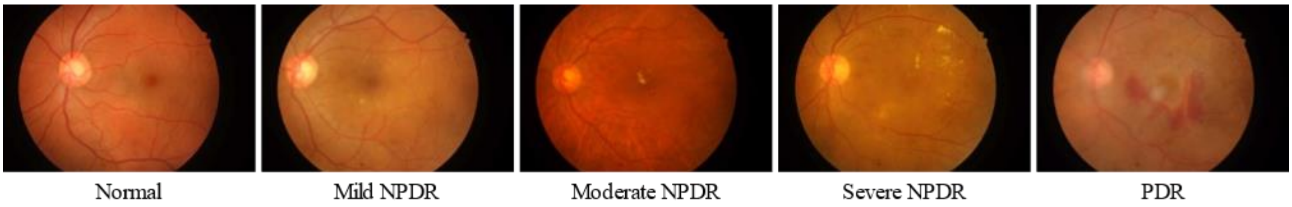


Figure 1. Schematic diagram of DR grading.

With the rapid development of artificial intelligence, deep learning and related technologies have been widely applied in DR image analysis, providing new approaches for early diagnosis and precise treatment of DR. For instance, Pratt et al. [3] combined image enhancement techniques with classic convolutional neural network (CNN) models to grade DR on color fundus images; Wang et al. [4] proposed a dual-stream CNN for multi-model age-related macular degeneration grading, effectively fusing information from color fundus photographs (CFP) and optical coherence tomography (OCT). Although CNNs have achieved excellent performance in DR grading tasks, they still face limitations in handling subtle features such as microaneurysms. The severity of DR is mainly assessed through analyzing CFP. Although these images can clearly show the microstructures of the retina, subtle and diverse DR lesion features, such as microaneurysms and hemorrhages, pose significant challenges for automated grading.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs) [5–9], have significantly improved the performance of automated DR grading systems. Models like Inceptionv3 [10] have shown impressive performance in DR detection. To further improve the model's ability to extract relevant features, researchers have incorporated attention mechanisms. For example, Li et al. applied the attention mechanism to video object segmentation, distance metric learning, video description, and feature selection [11–14]. Attention mechanisms assign different weights to different features, enabling the model to focus on critical information. Attention mechanisms have been widely applied in image classification, object detection, and other fields. In DR grading tasks, He et al. [15]

introduced a category attention block to focus on small, critical lesions. Zhou et al. [16] combined lesion segmentation with DR grading, leveraging pixel-level annotations to provide more detailed information for improved accuracy. Yang et al. [17] designed a two-stage network that can not only locate lesion positions but also perform DR grading. Li et al. [18] employed a joint grading approach to simultaneously evaluate DR and diabetic macular edema (DME), utilizing disease-specific attention modules to learn disease-specific features. Although spatial domain attention mechanisms have improved model performance by enhancing feature focus, they still have limitations in capturing the global context. To address this, frequency domain attention mechanisms have emerged. These transform spatial domain features into frequency domain features, allowing the model to capture richer feature information from a global perspective. For example, Qin et al. [19] proved the equivalence between discrete cosine transform (DCT) and global average pooling (GAP) and integrated this insight into the channel attention mechanism. Zhou et al. [20] embedded fast Fourier transform (FFT) into the transformer, constructing a frequency domain enhancement module to replace self-attention and cross-attention modules. By transforming data into the frequency domain, these models can better capture global correlations, further enhancing performance in DR grading tasks.

Despite significant advancements in deep learning, challenges remain in DR grading, including the limited availability of high-quality annotated data, the diversity and subtlety of DR lesions, and the need for models robust enough for complex clinical environments. To address these challenges, this study proposes FF-ResNet-DR, a deep learning–based model that incorporates frequency domain attention mechanisms. By incorporating both FCAM and FSAM, the model captures both frequency and spatial domain features, allowing it to effectively extract and emphasize subtle lesion features. This approach improves the model’s ability to accurately classify DR, even in face of challenging imaging conditions.

2. Materials and experimental setup

2.1. Datasets

2.1.1. EAM dataset

To comprehensively evaluate the performance of the proposed model and improve its generalization ability, we integrated three datasets, EyePacs, Messidor, and Aptos, to construct the EAM dataset. The EAM dataset contains a total of 92,501 images. Among them, EyePacs, as the largest DR grading dataset, contains 35,126 images with significant variations in image quality, which can effectively test the model’s robustness to different image qualities [21]. The Messidor dataset contains 1200 high-quality color fundus images, which can be used to verify the model’s performance on small-scale datasets [22]. The Aptos dataset provides 3662 high-quality fundus images, which can serve as supplementary training data for the model [23].

2.1.2. IDRiD and DDR datasets

The IDRiD and DDR datasets are competition datasets. The IDRiD dataset provides 516 high-quality fundus images with both DR and DME labels, which can be used to evaluate the model’s multi-task learning ability [24]. The DDR dataset contains 13,673 images covering multiple DR grades, which can be used to evaluate the model’s performance on fine-grained classification [25].

To enhance the model’s generalization ability, data augmentation techniques were applied to the aforementioned datasets, including horizontal flipping, vertical flipping, rotation, cropping, and random adjustment of image brightness and contrast.

2.2. Experimental setup

Table 1 presents the experimental parameters of the proposed model. In this study, ResNet50 [26] is used as the backbone network, and the ImageNet pre-trained model is loaded. The input image size is uniformly set to 512 × 512 pixels. The initial learning rate is set to 0.001, the batch size is 32, and the cosine annealing strategy is used to adjust the learning rate. All models are trained for 20 epochs using the stochastic gradient descent (SGD) optimizer and cross-entropy loss as the objective function. All experimental models are implemented using Python and built on the PyTorch deep learning framework. Experiments are conducted on the Kaggle platform, and GPU P100 (16 GB) is used to accelerate the training process.

Table 1. Experimental parameters of DR grading model.

Experimental parameter	**Value**
Input resolution	512 × 512
Optimizer	SGD
Scheduler	CosineAnnealingLR
Loss function	CrossEntropyLoss
Batch size	32
Initial learning rate	10^{-3}
Minimum learning rate	10^{-5}

2.3. Evaluation metrics

The performance of different DR grading models was analyzed and compared using accuracy (Acc), quadratic weighted kappa coefficient (Kappa) [27], area under the receiver operating characteristic (ROC) curve (AUC), recall, precision, F1-score [28], and joint accuracy as evaluation metrics.

Acc is a performance metric that quantifies the proportion of correct predictions among the total predictions, defined as:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where TP, TN, FP, and FN denote the counts of true positives (correctly predicted positive instances), true negatives (correctly predicted negative instances), false positives (incorrectly predicted positive instances), and false negatives (incorrectly predicted negative instances), respectively.

Kappa further assesses the degree of agreement between the model’s predicted labels and the true labels, defined as:

$$kappa = 1 - \frac{\sum_{i,j} W_{i.j} O_{i.j}}{\sum_{i,j} W_{i.j} E_{i.j}} \tag{2}$$

$$W_{i.j} = \frac{(i-j)^2}{(N-1)^2} \tag{3}$$

$$E_{i.j} = \frac{n_i \cdot n_j}{N} \tag{4}$$

where W is the weight matrix, and $W_{i.j}$ represents the penalty between the predicted class i and the true class j. O is the observed matrix, i.e., the confusion matrix, $O_{i.j}$ represents the number of samples in the predicted class i and the true class j. E is the expected matrix, and $E_{i.j}$ represents the probability of a certain class i predicted as class j under random prediction. A Kappa value greater than 0.8 generally indicates a good level of agreement.

AUC represents the area under the ROC curve, which is a graphical representation of a classifier's performance by plotting the true positive rate (recall) against the false positive rate at various threshold settings, ranging from 0 to 1. An AUC of 0.5 indicates random performance, while an AUC of 1 represents the best possible performance.

Recall, also known as sensitivity, is the proportion of actual positives that are correctly identified, defined as:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Recall measures a model's ability to find all positive instances. The higher the recall, the greater the likelihood that samples with a true positive label will be predicted as positive.

Precision, also known as positive predictive value, is the proportion of predicted positives that are actual positives, defined as:

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Precision measures a model's ability to correctly identify only the relevant positive instances. The higher the precision, the fewer false positives the model produces.

F1 balances both precision and recall, defined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

F1 ranges from 0 to 1, with higher values indicating better model performance.

Joint accuracy, in this context, refers to the proportion of samples where the model correctly predicts both DR and DME grades. Specifically, a sample is considered correct only if its predicted DR and DME grades match exactly the ground truth labels.

3. Diabetic retinopathy grading model

Figure 2 illustrates the architecture of our proposed diabetic retinopathy grading model, FF-ResNet-DR. The model leverages ResNet50 as its backbone to extract deep features from color fundus photographs (CFPs). To enhance the model's ability to identify and localize lesions, we introduce two attention modules: the FCAM and the FSAM.

FCAM refines feature representation by fusing frequency and spatial domain information. It employs a modified multi-spectral attention layer (MSA) to extract frequency domain features, which are then combined with spatial domain features from a channel attention module (CAM) [29]. This approach assigns higher weights to key channels, improving the model's sensitivity to lesion areas.

FSAM enhances the model's ability to capture fine-grained texture details. By simultaneously extracting and multiplying spatial and frequency domain attention maps, FSAM highlights key pixels and emphasizes subtle lesion features.

The FCAM and FSAM modules enhance the FF-ResNet-DR model's ability to capture subtle lesion features by introducing frequency domain information at both channel and spatial levels. The refined features, obtained by applying these attention modules, are then fused with the original features. A fully connected layer classifies the fused features, enabling accurate DR grading.

In Figure 2, the first column presents an overview of the entire network, showcasing the connections between different modules. The second column delves into the internal workings of FCAM and FSAM, providing a detailed explanation of their mechanisms.

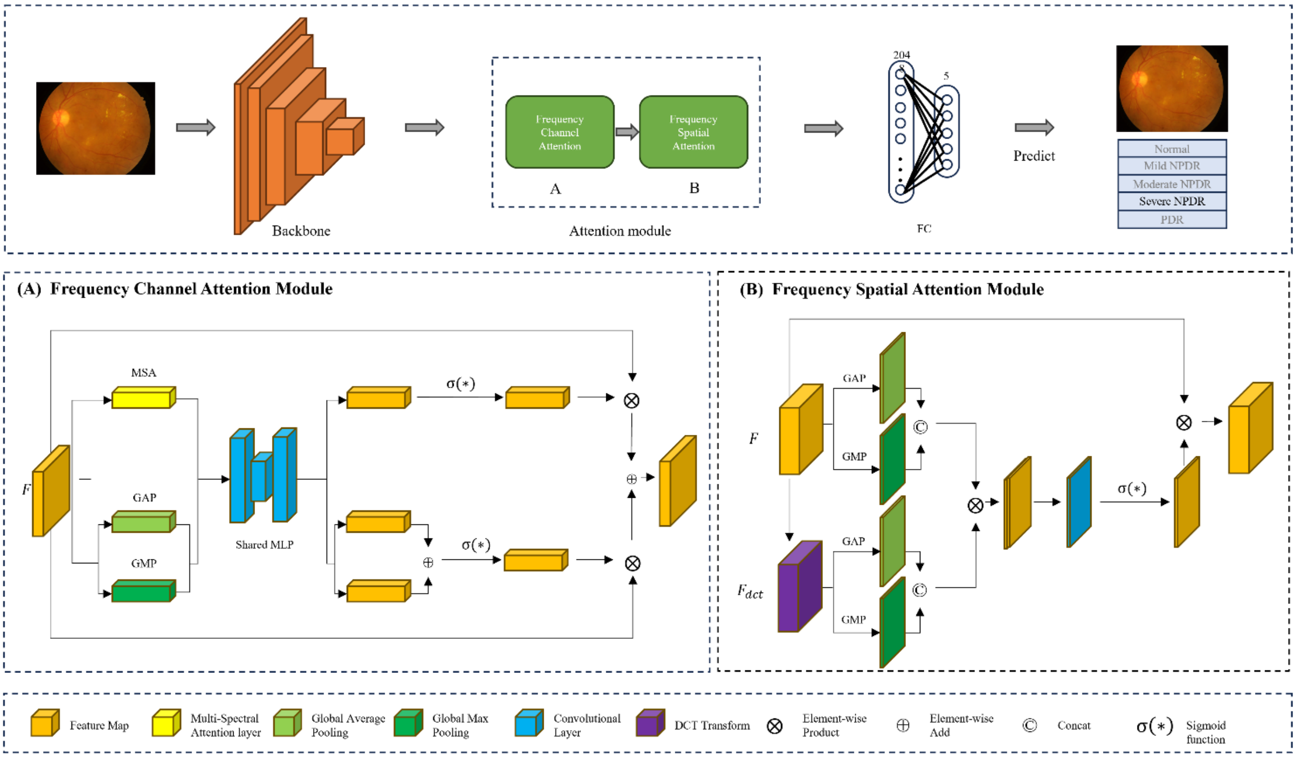


Figure 2. Architecture of DR grading model (FF-ResNet-DR).

3.1. FCAM

Channel attention mechanism (CAM), as an effective feature calibration method, has been widely used in convolutional neural networks. The classic CAM generates channel weights through global pooling and a multi-layer perceptron (MLP) to emphasize key features. However, CAM only focuses on spatial domain information and ignores the rich information that may be contained in the frequency domain of images. To address the limitations of CAM, this paper introduces an FCAM.

FCAM extracts frequency domain features of images through an improved MSA. The MSA layer [19] first transforms the image from the spatial domain to the frequency domain using DCT and then captures the correlation between frequency domain features through a self-attention mechanism. Finally, the obtained frequency domain features are fused with the spatial domain features extracted by CAM to generate a more comprehensive channel attention. This design enables FCAM to simultaneously focus on both spatial and frequency domain feature information, thereby assigning higher weights to key channels and improving the model's sensitivity to lesion areas.

3.1.1. DCT

DCT is an orthogonal transform that converts signals from the time or spatial domain to the frequency domain [30]. DCT has a good energy compaction property, meaning that most of the signal energy is concentrated in the low-frequency part. In the field of image compression, DCT is widely used in the JPEG image compression standard. In this paper, we utilize this property of DCT to transform images from the spatial domain to the frequency domain in order to extract frequency domain features of images.

Given a 2D image $x^{2d} \in R^{H\times W}$, the corresponding 2D DCT image $f^{2d} \in R^{H\times W}$ can be obtained by:

$$f_{h,w}^{2d} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j},$$
$$s.t.\ h \in \{0,1,\dots,H-1\}, w \in \{0,1,\dots,W-1\}. \tag{7}$$

where H and W denote the height and width of the image x^{2d}, respectively. $B_{h,w} \in R^{H\times W}$represents the basic functions of the discrete cosine transform (DCT), which is crucial for extracting frequency information from the image x^{2d}. $x_{i,j}^{2d}$, $B_{h,w}^{i,j}$denotes the pixel value at location (i,j) in the original image, and $f_{h,w}^{2d}$ denotes the pixel value at location (h,w) in the transformed DCT image. $B_{h,w}^{i,j}$ can be expressed as:

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \tag{8}$$

When $h = w = 0$, the DCT transform is equivalent to global average pooling (GAP) as follows:

$$\begin{aligned} f_{0,0}^{2d} &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} B_{0,0}^{i,j} \\ &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} \\ &= gap\left(x^{2d}\right)HW \end{aligned} \tag{9}$$

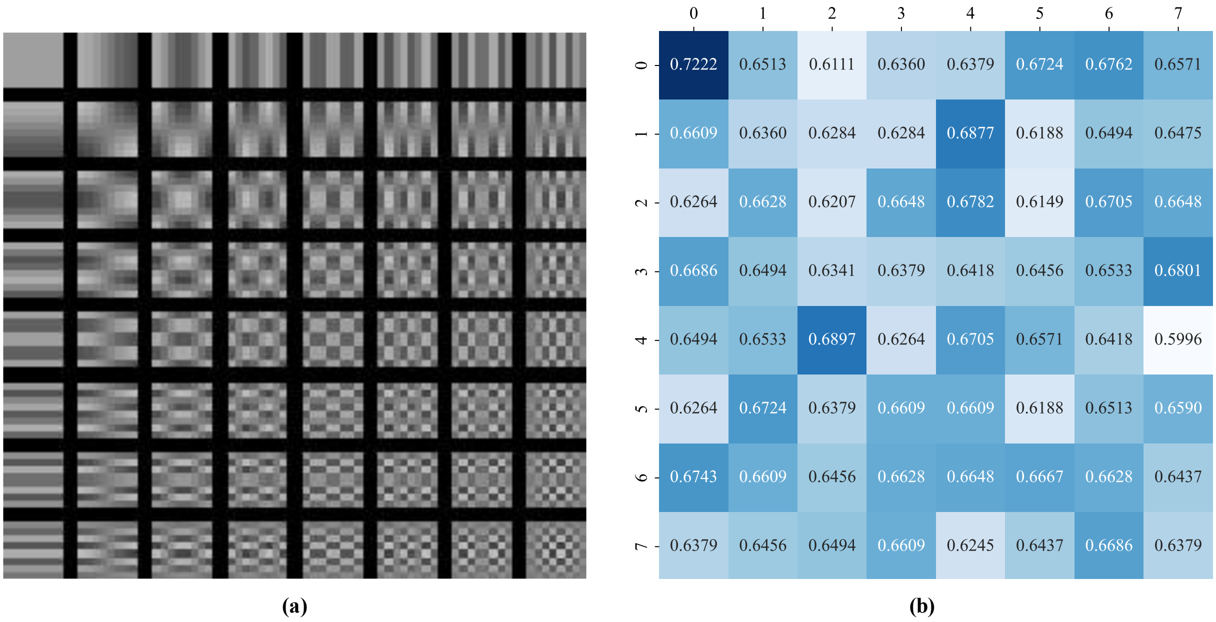
Using DCT basis functions to transform an image is equivalent to performing a pooling operation on the image.

3.1.2. MSA

MSA decomposes an image into multiple frequency bands by employing DCT basis functions with different frequencies, thus enabling the extraction of multi-scale features [19]. DCT basis functions exhibit excellent energy compaction properties, where low-frequency components typically capture global image information, while high-frequency components contain detailed image information. By leveraging this property, MSA can capture rich image features from the frequency domain, thereby enhancing the model's representational capacity. Different combinations of DCT basis functions can exert varying influences on channel attention. FcaNet [19] proposes a basis function selection scheme based on classification performance. However, this scheme is not suitable for DR grading tasks.

To better adapt to DR grading tasks, this paper improves the basis function selection scheme of MSA, including (1) increasing the number of basic functions—to capture high-frequency detail information more comprehensively, the number of DCT basis functions is increased from the original 49 to 64, as shown in Figure 3(a); and (2) basis function selection based on DR datasets—by conducting comparative experiments on public DR datasets, the optimal combination of basic functions for DR grading tasks is selected.

The improved MSA module is termed drMSA. The basis function structure of drMSA will be dynamically adjusted based on the classification accuracy of each basis function on the DR dataset. Figure 3(b) shows the classification accuracy of the 64 basis functions using the Messidor dataset as an example. Subsequent studies can select different DR datasets based on various task scenarios and refer to the results in Figure 3(b) to flexibly choose the optimal combination of basic functions.



	0	1	2	3	4	5	6	7
0	0.7222	0.6513	0.6111	0.6360	0.6379	0.6724	0.6762	0.6571
1	0.6609	0.6360	0.6284	0.6284	0.6877	0.6188	0.6494	0.6475
2	0.6264	0.6628	0.6207	0.6648	0.6782	0.6149	0.6705	0.6648
3	0.6686	0.6494	0.6341	0.6379	0.6418	0.6456	0.6533	0.6801
4	0.6494	0.6533	0.6897	0.6264	0.6705	0.6571	0.6418	0.5996
5	0.6264	0.6724	0.6379	0.6609	0.6609	0.6188	0.6513	0.6590
6	0.6743	0.6609	0.6456	0.6628	0.6648	0.6667	0.6628	0.6437
7	0.6379	0.6456	0.6494	0.6609	0.6245	0.6437	0.6686	0.6379

Figure 3. Basis function selection scheme of the MSA. (a) 64 basis functions generated by discrete cosine transform (DCT) on an 8 × 8 image; (b) Classification accuracy of the 64 basis functions on the Messidor dataset.

After determining the drMSA basis functions, the input feature channels are grouped. Each group of channels is assigned a specific basis function to extract corresponding frequency domain features. In this way, a set of channel vectors containing different frequency features is obtained. Subsequently, an MLP is used to process these channel vectors, generating a weight vector that represents the importance of each channel in the frequency domain.

Given an input feature map $F \in R^{C \times H \times W}$, drMSA produces a DCT frequency domain channel attention map $F_{dct\_att} \in R^{C \times H \times W}$, defined as:

$$F_{dct\_att} = \sigma(W_0 W_1 F_{MSA}) \otimes F \tag{10}$$

where F represents the input feature map, $F_{MSA} \in R^{C\times1\times1}$ denotes the attention weight vector obtained through the DCT basis function transformation, and W_0 and W_1 represent the weights of the MLP. The final output of drMSA is F_{dct_att}.

For the CAM part, given an input feature map $F \in R^{H\times W\times C}$, the channel attention map of this part is $F_{att} \in R^{C\times H\times W}$, defined as:

$$F_{mlp\_out} = W_0 W_1 F_{gap} \oplus W_0 W_1 F_{gmp} \tag{11}$$

$$F_{att} = \sigma(F_{mlp\_out}) \otimes F \tag{12}$$

where $F_{gap} \in R^{C\times1\times1}$ and $F_{gmp} \in R^{C\times1\times1}$ are the results by the weight vectors obtained through GAP and GMP multiplied the input feature map F.

Finally, the output of drMSA $F_{dct_att}$is element-wisely added to the output of CAM F_{att} to allow the model to simultaneously capture features from both the spatial domain and the frequency domain, improving the model's complexity and data diversity and enhancing generalization ability. The output of FCAM $F_{DFCAM} \in R^{C\times H\times W}$is obtained as:

$$F_{FCAM} = F_{dct\_att} \oplus F_{att} \tag{13}$$

3.2. FCAM

The spatial attention module (SAM) in CBAM can effectively capture spatial information of images, helping the model better understand the global structure and semantic information of images. However, SAM only focuses on spatial domain features and ignores the rich information contained in the frequency domain of images.

To fully utilize frequency domain information, this paper proposes a FSAM. FSAM introduces a DCT frequency domain spatial attention mechanism based on SAM. By using the DCT transform, the image is transformed from the spatial domain to the frequency domain to extract the frequency domain features of the image. These frequency domain features contain detailed texture information of the image, and when combined with spatial domain features, they can improve the feature extraction ability and classification accuracy of the model. In addition, the fusion of frequency domain spatial features also makes the model more robust to noise.

Given an input feature map $F \in R^{C\times H\times W}$, its corresponding frequency domain image is $F_{dct} \in R^{C\times H\times W}$. Both are first stacked and concatenated through GAP and GMP operations to obtain two intermediate feature maps $F_{concat} \in R^{2\times H\times W}$.

After concatenation, the spatial domain and frequency domain intermediate feature maps are fused by element-wise multiplication to obtain $F_{mix} \in R^{2\times H\times W}$, which enriches and highlights the most noteworthy details in the image.

Finally, F_{mix} is fed into a convolutional layer and an activation function $F_{att} \in R^{1\times H\times W}$ to obtain the attention map $F_{FSAM} \in R^{C\times H\times W}$, which can be expressed as:

$$F_{concat} = [GAP(F), GMP(F)] \tag{14}$$

$$F_{dct\_concat} = [GAP(F_{dct}), GMP(F_{dct})] \tag{15}$$

$$F_{mix} = F_{concat} \otimes F_{dct\_concat} \tag{16}$$

$$F_{FSAM} = softmax(WF_{mix}) \otimes F \tag{17}$$

where GAP and GMP represent global average pooling and global maximum pooling, respectively, W is the weight of the convolutional filter, and the softmax operation is used to obtain the spatial attention map, which is finally multiplied by the original image to obtain the result F_{FSAM}.

4. Experimental analysis and results comparison

To evaluate the performance of FF-ResNet-DR, our proposed diabetic retinopathy grading model, two sets of experiments were designed. Experiment 1 is an ablation study. This study aimed to determine the individual impact of each improved module within FF-ResNet-DR by comparing its performance with variations using different attention mechanisms. Experiment 2 is a comparative experiment. This study compared FF-ResNet-DR with other state-of-the-art models to evaluate its grading accuracy and identify any performance advantages.

4.1. Ablation study

This chapter conducts an ablation study to assess the impact of each module within the proposed FF-ResNet-DR grading model. This analysis quantifies the specific contributions of individual components to the model's grading accuracy and robustness. The study encompasses: (1) verification and analysis of the improved multi-spectral attention layer (drMSA), which evaluates the effectiveness of the enhanced attention mechanism in capturing relevant features for DR grading; and (2) comparison and analysis of different attention modules, which compares the performance of the drMSA with other attention mechanisms, highlighting its advantages in the context of DR grading.

4.1.1. drMSA

To evaluate the contribution of drMSA to the FF-ResNet-DR grading model, this section integrates both drMSA and MSA modules into the model's FCAM. Comparative experiments are conducted on the EAM, IDRiD, and DDR datasets

To investigate the impact of the number of basis functions, four different configurations are evaluated for both drMSA and MSA: {4, 8, 16, 32} basis functions. These configurations are denoted as drMSA-4, drMSA-8, drMSA-16, drMSA-32, and MSA-4, MSA-8, MSA-16, and MSA-32 respectively.

Table 2 presents the grading performance of drMSA with varying numbers of basis functions. Table 3 compares the performance of drMSA with MSA for each basis function configuration, where "-" represents no promotion. Experimental results demonstrate the following: (1) On the IDRiD dataset, drMSA-8 achieves the best performance, with a 2.22% increase in Kappa index and a 1.55% increase in accuracy compared to the baseline. (2) On the DDR dataset, MSA-16 exhibits the best performance, with a 2.57% increase in Kappa index and a 2.24% increase in accuracy. (3) On the EAM dataset, drMSA-32 yields the best performance, with a 3.12% increase in Kappa index and a 0.9% increase in accuracy. Furthermore, drMSA consistently outperforms MSA across multiple datasets, with the most significant performance improvement observed on the IDRiD dataset. These findings unequivocally demonstrate the effectiveness of drMSA in enhancing DR grading performance.

Table 2. Grading performance of drMSA with varying numbers of basis functions.

Different	**IDRiD**		**DDR**		**EAM**	
configurations	**Kappa%**	**Acc/%**	**Kappa/%**	**Acc/%**	**Kappa/%**	**Acc/%**
ResNet50 [26]	77.28	66.09	73.22	77.58	78.12	86.23
CBAM [29]	80.06	66.86	74.13	78.01	77.80	86.40
drMSA-4	78.60	68.80	69.80	76.70	77.80	86.40
drMSA-8	82.28	68.41	75.47	79.00	79.69	86.99
drMSA-16	78.10	66.30	74.20	77.80	79.70	86.30
drMSA-32	78.47	64.15	75.96	79.17	80.92	87.30
MSA-4	73.71	62.79	71.80	77.83	79.42	86.44
MSA-8	71.20	61.80	76.70	69.90	78.00	86.40
MSA-16	79.63	66.67	76.49	80.25	79.87	86.76
MSA-32	78.24	65.50	72.20	77.22	78.03	86.69

Table 3. Comparison of the performance of drMSA with MSA for each basis function configuration.

Different	**IDRiD**		**DDR**		**EAM**	
configurations	**Kappa/%**	**Acc/%**	**Kappa/%**	**Acc/%**	**Kappa/%**	**Acc/%**
drMSA-4	-	1.94	-	-	-	-
drMSA-8	2.22	1.55	1.34	0.99	1.89	0.59
drMSA-16	-	-	0.07	-	1.90	-
drMSA-32	-	-	1.83	1.16	3.12	0.90
MSA-4	-	-	-	-	1.62	0.04
MSA-8	-	-	2.57	-	0.20	-
MSA-16	-	-	2.36	2.24	2.07	0.36
MSA-32	-	-	-	-	0.23	0.29

4.1.2. Comparison and analysis of different attention modules

To thoroughly evaluate the role of frequency domain attention mechanisms, this paper compares and analyzes the performance of different attention module combinations in the FF-ResNet-DR grading model.

Table 4 presents the performance of various attention module combinations. Figure 4 provides an in-depth analysis of how the proposed frequency domain attention module refines feature extraction, enhancing DR grading performance. This analysis utilizes the gradient-weighted class activation mapping (Grad-CAM) visualization technique.

Results demonstrate that the combination of FCAM and FSAM achieved the highest performance, with a Kappa coefficient of 82.28% and an accuracy of 68.41% on IDRiD, 75.47% and 79.00% on DDR, and 84.21% and 88.36% on EAM. These results indicate that incorporating frequency domain information into both channel and spatial attention mechanisms effectively enhances the model's feature extraction capabilities and generalization ability.

Table 4. Comparison of different attention modules.

Different configurations	IDRiD Kappa/%	Acc/%	DDR Kappa/%	Acc/%	EAM Kappa/%	Acc/%
ResNet50 [26]	77.28	66.09	73.22	77.58	83.57	88.32
CBAM [29]	80.06	66.86	74.13	78.01	82.25	87.56
drMSA+SAM	80.09	66.67	68.42	75.07	83.25	88.10
FCAM+SAM	79.40	67.83	69.60	75.50	83.64	88.37
CAM+FSAM	74.60	63.80	70.30	77.10	83.50	**88.46**
drMSA+FSAM	77.27	66.47	68.70	77.09	82.47	88.16
FCAM+FSAM	**82.28**	**68.41**	**75.47**	**79.00**	**84.21**	88.36

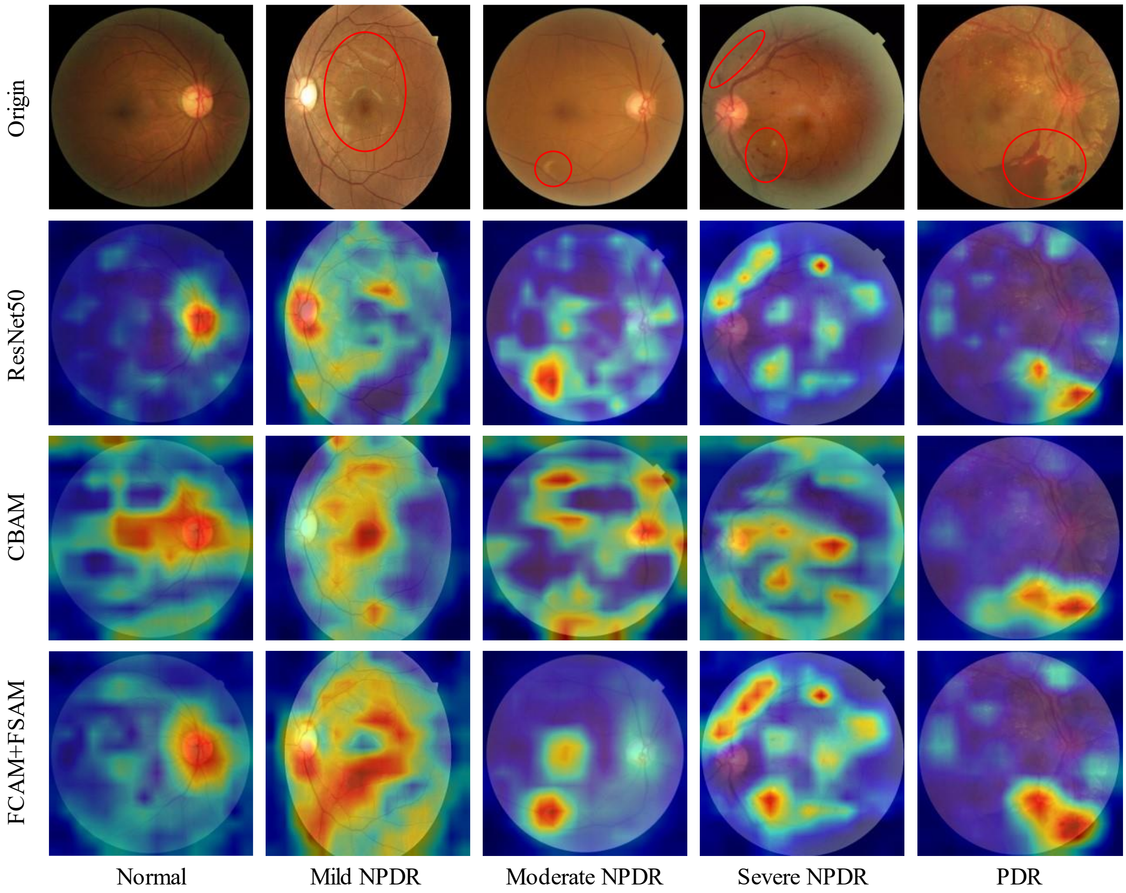


Figure 4. Visualization results of different models on DDR dataset.

As shown in Figure 4, the first row displays the original image, with the red circle highlighting the lesion region. Subsequent rows present heat maps generated by the no-attention model, the CBAM model, and our proposed FF-ResNet-DR model. Based on Figure 4, we can observe the following: (1) Heat maps from the no-attention model appear coarse and fail to effectively focus on the lesion regions. (2) In contrast, our proposed FF-ResNet-DR model effectively emphasizes the lesion locations, as evident by comparing the second and fourth columns. (3) While the CBAM model produces heat maps with more diffuse attention, highlighting numerous irrelevant features, our proposed FF-ResNet-DR model consistently provides more focused attention on the lesion regions, particularly evident in the third and fourth columns.

4.2. A comparative study and application of the FF-ResNet-DR model

This chapter conducts three comparative experiments to evaluate the performance of the proposed FF-ResNet-DR grading model, including (1) binary classification of diabetic retinopathy, classifying images as either *normal* or *abnormal* for diabetic retinopathy; (2) DR and DME classification, distinguishing between DR and DME; and (3) five-stage classification of DR, classifying images into five different stages of diabetic retinopathy severity.

4.2.1. Binary classification of diabetic retinopathy

This experiment utilized the Messidor dataset, a widely used benchmark for diabetic retinopathy (DR) binary classification. Images with grades 0–1 were classified as normal, while those with grades 2–3 were considered abnormal. The dataset comprises 540 normal images and 660 abnormal images. To facilitate a comprehensive comparison, the proposed model was evaluated alongside CANet and CABNet. Figure 5 illustrates the ROC curves of our model, CANet, and CABNet on the Messidor validation set. Table 5 presents a comparison of experimental results for different models on the Messidor dataset.

As depicted in Figure 5 and Table 5, the ROC curve of our proposed model significantly surpasses those of the other models. Our model achieved an AUC of 98.1%, demonstrating a 1.5% improvement over both CABNet and CANet. Our proposed model achieved the highest AUC on the Messidor binary classification task. Compared to the state-of-the-art CABNet model, our model demonstrated improvements of 1.2% in AUC, 0.7% in accuracy, 2.2% in recall, and 1.9% in F1-score. These results indicate that our proposed model can more accurately distinguish between normal and abnormal DR fundus images, highlighting its significant advantage in DR grading tasks.

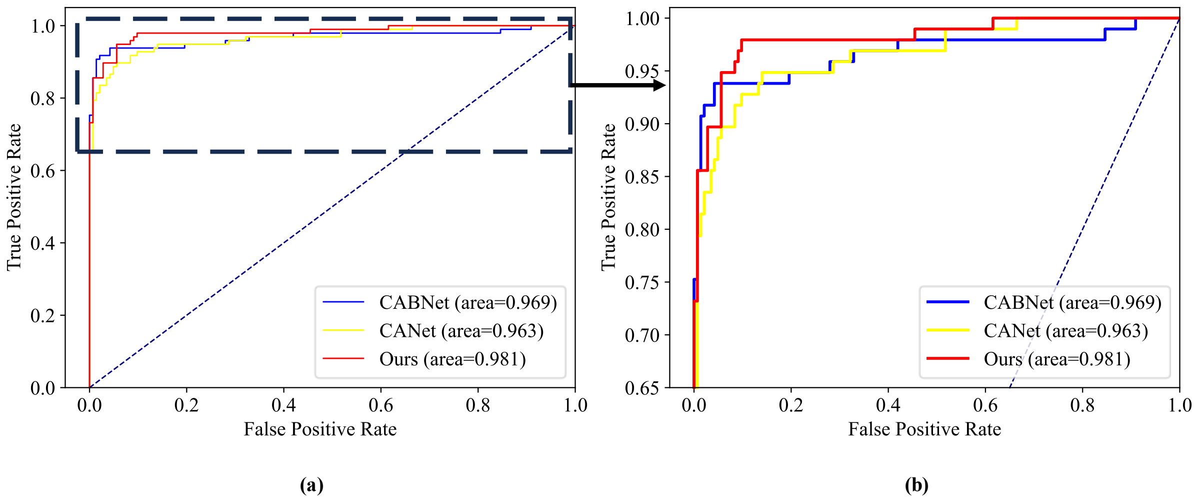


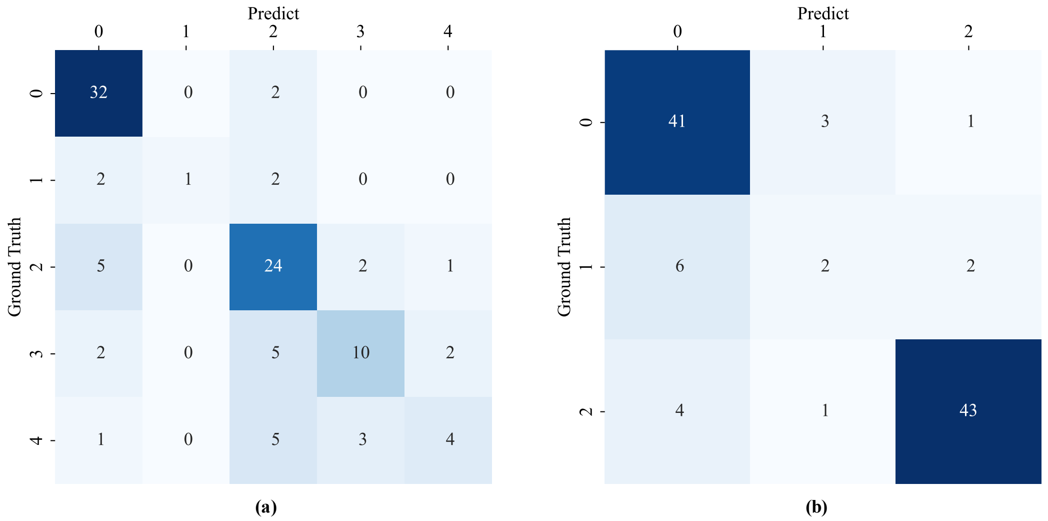
Figure 5. ROC curves of our model, CANet, and CABNet on the Messidor validation set. (a) Full ROC curve; (b) zoomed-in region of interest.

Table 5. Comparison of experimental results for different models on the Messidor dataset.

Models	**AUC/%**	**Acc/%**	**Reall/%**	**F1/%**
CANet [18]	96.3	92.6	92.0	91.2
CABNet [15]	96.9	93.1	90.2	91.5
Ours	**98.1**	**93.8**	**92.4**	**93.4**

4.2.2. DR and DME classification

This experiment utilized the IDRiD dataset to perform joint classification of DR and DME. Table 6 presents the experimental results, while Figure 6 provides a visual representation of the classification performance through a confusion matrix. Our model achieved a joint classification accuracy of 64.1%, surpassing the Lzyuncc model by 1%. Furthermore, our model exhibited promising performance in DME classification.



(a) **(b)**

Figure 6. Confusion matrices of DR and DME classification. (a) DR confusion Matrix; (b) DME confusion matrix.

Table 6. Joint classification results on IDRiD dataset.

Models	**DR Acc**	**DME Acc**	**Joint Acc**
Lzyuncc [18]	**74.8**	80.6	63.1
VRT [18]	59.2	81.6	55.3
Mammoth [18]	54.4	83.5	51.5
Ours	68.9	**83.5**	**64.1**

4.2.3. Five-stage classification of diabetic retinopathy

This experiment employs the IDRiD and DDR datasets for multi-class DR classification, categorizing DR into five stages: No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and PDR.

Table 7 presents experimental results from the different models. On the IDRiD dataset, our model achieved a Kappa of 82.28% and an accuracy of 68.41%, demonstrating an advantage in terms of the Kappa coefficient compared to other methods. On the DDR dataset, although our model's performance on the Kappa coefficient lagged behind CLANet, its accuracy remained relatively close.

Table 7. Comparison of experimental results for different models across different datasets.

Models	IDRiD		DDR	
	Kappa/%	Acc/%	Kappa/%	Acc/%
CABNet [15]	66.21	67.96	78.57	77.73
CLANet [31]	80.12	**71.84**	**80.84**	**79.12**
Ours	**82.28**	68.41	75.47	79.00

Figure 7 shows the visualization results of our proposed models on the IDRiD dataset. As shown in Figure 7, the first row displays the original image, with the red circle highlighting the lesion region. Subsequent rows present heat maps generated by the no-attention model and our proposed FF-ResNet-DR model. We can observe that the heatmaps from the no-attention model fail to effectively focus on the lesion regions, while our proposed FF-ResNet-DR model consistently provides more focused attention on these regions, particularly evident in the third, fourth, and last columns.

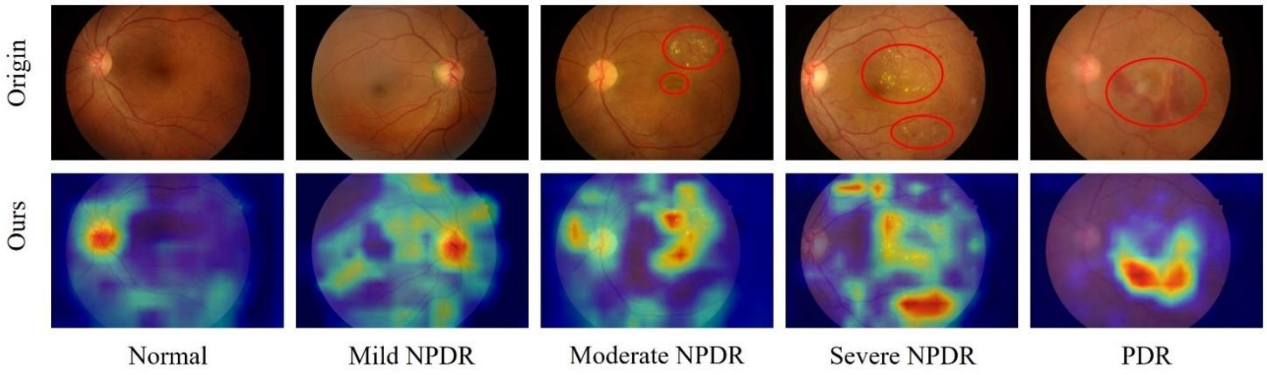


Figure 7. Multi-class DR Classification on IDRiD.

5. Discussion and conclusions

DR is a serious eye condition that necessitates accurate grading for effective treatment planning. While deep learning offers promising solutions for automated DR grading, challenges remain in achieving high classification accuracy and reliably detecting subtle features. To address these challenges, we propose FF-ResNet-DR, a novel deep learning model that leverages a frequency domain attention mechanism.

FF-ResNet-DR integrates spatial and frequency domain information, enabling it to capture both global and local patterns within retinal images. By analyzing images in the frequency domain, the model can identify subtle patterns and textures that may be missed by spatial-based approaches. This enhanced feature representation significantly improves the model's ability to differentiate between DR stages, leading to more accurate and reliable grading.

The proposed model has the potential to significantly impact clinical practice. By providing clinicians with more accurate and reliable grading results, FF-ResNet-DR can support more informed treatment.

To further advance DR grading, we will explore the following directions:

Enhanced frequency domain feature extraction and fusion: While the current model employs discrete cosine transform (DCT) for frequency domain feature extraction, we will investigate alternative methods such as Fourier transform for potentially superior performance. Additionally, we will experiment with different fusion strategies to optimally combine frequency and spatial domain features, leading to more robust and informative representations.

Multi-modal learning: To exploit complementary information from multiple modalities, we will explore the integration of color fundus images with other medical imaging modalities, such as optical coherence tomography (OCT). By leveraging multi-modal learning, we aim to further improve feature extraction and classification accuracy. Furthermore, incorporating temporal sequence information can capture the dynamic progression of DR lesions, which is crucial for early detection and accurate staging.

Clinical application: The proposed model significantly enhances the diagnostic capabilities of computer-aided diagnosis (CAD) systems. By saving the trained model weights, it can be seamlessly integrated into existing medical systems. We envision a CAD system comprising the following components: (1) Data preprocessing: Fundus imaging devices in medical institutions capture color fundus photographs. These images are then preprocessed to match the input requirements of the model. (2) Model inference: The preprocessed images are fed into the trained model, which generates diagnostic results. (3) Result generation: The model provides diagnostic results, which may include attention-based heatmaps highlighting areas of interest. (4) Physician assistance: The generated results assist physicians in identifying pathological regions, improving diagnostic accuracy and efficiency.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was supported by the grants from the National Natural Science Foundation of China (82271107) and the Natural Science Foundation of Jiangsu Province (BK20221186).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. O. Alamoudi, S. M. Allabun, Blood vessel segmentation with classification model for diabetic retinopathy screening, *Comput. Mater. Contin.*, **75** (2023), 2265–2281. https://doi.org/10.32604/cmc.2023.032429

2. J. Lo, T. Y. Timothy, D. Ma, P. Zang, J. P. Owen, Q. Zhang, et al., Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data, *Ophthalmol. Sci.*, **1** (2021), 100069. https://doi.org/ 10.1016/j.xops.2021.100069
3. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, *Procedia Comput. Sci.*, **90** (2016), 200–205. https://doi.org/10.1016/j.procs.2016.07.014
4. W. Wang, X. Li, Z. Xu, W. Yu, J. Zhao, D. Ding, et al., Learning two-stream CNN for multi-modal age-related macular degeneration categorization, *IEEE J. Biomed. Health Inf.*, **26** (2022), 4111–4122. https://doi.org/10.1109/JBHI.2022.3171523
5. Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, X. Wang, Zoom-in-net: Deep mining lesions for diabetic retinopathy detection, in *Medical Image Computing and Computer Assisted Intervention−MICCAI 2017*, (2017), 267–275. https://doi.org/10.1007/978-3-319-66179-7_31
6. L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, J. P. Langlois, Red lesion detection using dynamic shape features for diabetic retinopathy screening, *IEEE Trans. Med. Imaging*, **35** (2015), 1116–1126. https://doi.org/10.1109/TMI.2015.2509785
7. W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, et al., Automated identification and grading system of diabetic retinopathy using deep neural networks, *Knowl. Based Syst.*, **175** (2019), 12–25. https://doi.org/10.1016/j.knosys.2019.03.016
8. R. Zheng, L. Liu, S. Zhang, C. Zheng, F. Bunyak, R. Xu, et al., Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network, *Biomed. Opt. Express*, **9** (2018), 4863–4878. https://doi.org/10.1364/BOE.9.004863
9. X. He, Y. Deng, L. Fang, Q. Peng, Multi-modal retinal image classification with modality-specific attention network, *IEEE Trans. Med. Imaging,* **40** (2021), 1591–1602. https://doi.org/10.1109/TMI.2021.3059956
10. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *JAMA*, **316** (2016), 2402–2410. https://doi.org/10.1001/jama.2016.17216
11. P. Li, Y. Zhang, L. Yuan, H. Xiao, B. Lin, X. Xu, Efficient long-short temporal attention network for unsupervised video object segmentation, *Pattern Recognit.*, **146** (2024), 110078. https://doi.org/10.1016/j.patcog.2023.110078
12. P. Li, G. Zhao, J. Chen, X. Xu, Deep metric learning via group channel-wise ensemble, *Knowl. Based Syst.*, **259** (2023), 110029. https://doi.org/10.1016/j.knosys.2022.110029
13. P. Li, P. Zhang, T. Wang, H. Xiao, Time-frequency recurrent transformer with diversity constraint for dense video captioning, *Inf. Process. Manage.*, **60** (2023), 103204. https://doi.org/10.1016/j.ipm.2022.103204
14. P. Li, J. Chen, L. Yuan, X. Xu, M. Song, Triple-view knowledge distillation for semi-supervised semantic segmentation, preprint, arXiv:2309.12557. https://doi.org/10.48550/arXiv.2309.12557
15. A. He, T. Li, N. Li, K. Wang, H. Fu, CABNet: Category attention block for imbalanced diabetic retinopathy grading, *IEEE Trans. Med. Imaging,* **40** (2020), 143–153. https://doi.org/10.1109/TMI.2020.3023463
16. Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, et al., Collaborative learning of semi-supervised segmentation and classification for medical images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *IEEE*, (2019), 2079–2088. https://doi.org/10.1109/CVPR.2019.00218

17. Y. Yang, T. Li, W. Li, H. Wu, W. Fan, W. Zhang, Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks, in *Lecture Notes in Computer Science*, Springer, (2017), 533–540. https://doi.org/10.1007/978-3-319-66179-7_61
18. X. Li, X. Hu, L. Yu, L. Zhu, C. W. Fu, P. A. Heng, CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading, *IEEE Trans. Med. Imaging*, **39** (2019), 1483–1493. https://doi.org/10.1109/TMI.2019.2951844
19. Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: Frequency channel attention networks, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE*, (2021), 783–792. https://doi.org/10.1109/ICCV48922.2021.00082
20. T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, R. Jin, Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, in *Proceedings of the 39th International Conference on Machine Learning, PMLR*, (2022), 27268–27286. https://doi.org/10.48550/arXiv.2201.12740
21. R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, Y. Zhang, Lesion-aware transformers for diabetic retinopathy grading, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, (2021), 10938–10947. https://doi.org/10.1109/CVPR46437.2021.01079
22. E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, et al., Feedback on a publicly distributed image database: The Messidor database, *Image Anal. Stereol.*, (2014), 231–234. https://doi.org/10.5566/ias.1155
23. M. M. Farag, M. Fouad, A. T. Abdel-Hamid, Automatic severity classification of diabetic retinopathy based on densenet and convolutional block attention module, *IEEE Access*, **10** (2022), 38299–38308. https://doi.org/10.1109/ACCESS.2022.3165193
24. P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, et al., Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research, *Data*, **3** (2018), 25. https://doi.org/10.3390/data3030025
25. T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang, Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, *Inf. Sci.*, **501** (2019), 511–522. https://doi.org/10.1016/j.ins.2019.06.011
26. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
27. J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, *Psychol. Bull.*, **70** (1968), 213. https://doi.org/10.1037/h0026256
28. M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation, in *Australasian Joint Conference on Artificial Intelligence*, (2006), 1015–1021. https://doi.org/10.1007/11941439_114
29. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, preprint, arXiv:1807.06521. https://doi.org/10.48550/arxiv.1807.06521
30. N. Ahmed, T. Natarajan, K. R. Rao, Discrete cosine transform, *IEEE Trans. Comput.*, **100** (1974), 90–93. https://doi.org/10.1109/T-C.1974.223784
31. X. Liu, W. Chi, A cross-lesion attention network for accurate diabetic retinopathy grading with fundus images, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–12. https://doi.org/10.1109/TIM.2023.3322497

32. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 618–626, https://doi.org/10.1109/ICCV.2017.74