IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 14 April 2023, accepted 2 May 2023, date of publication 8 May 2023, date of current version 12 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3273613

RESEARCH ARTICLE

Dual-Path and Multi-Scale Enhanced Attention Network for Retinal Diseases Classification Using Ultra-Wide-Field Images

FANGSHENG CHEN[1,2], SHAODONG MA[2], JINKUI HAO[2], MENGTING LIU[3], YUANYUAN GU[2], QUANYONG YI[4], JIONG ZHANG[2], AND YITIAN ZHAO[2], (Member, IEEE)

[1]School of Mechanical and Engineering, Zhejiang University of Technology, Hangzhou 310014, China
[2]Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo 315399, China
[3]School of Biomedical Engineering, Sun Yat-sen University, Shenzhen 510275, China
[4]The Affiliated Ningbo Eye Hospital of Wenzhou Medical University, Ningbo 325035, China

Corresponding authors: Jiong Zhang (zhangjiong@nimte.ac.cn) and Yitian Zhao (yitian.zhao@nimte.ac.cn)

This work was supported in part by the National Natural Science Foundation Program of China under Grant 62103398 and Grant 61906181; in part by the Zhejiang Provincial Natural Science Foundation under Grant LQ23F010002, Grant LZ23F010002, Grant LR22F020008, Grant LZ19F010001, and Grant LQ21F010007; in part by the Ningbo Natural Science Foundation under Grant 2022J143 and Grant 20221JCGY010608; in part by the Youth Innovation Promotion Association CAS under Grant 2021298; in part by the Research Startup Fund of Ningbo University of Technology under Grant 2022KQ29; and in part by 2025 S&T Megaprojects under Grant 2019B10033, Grant 2019B10061, and Grant 2021Z054.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Review Board of Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences.

ABSTRACT Early computer-aided early diagnosis (CAD) based on retinal imaging is critical to the timely management and treatment planning of retina-related diseases. However, the inherent characteristics of retinal images and the complexity of their pathological patterns, such as low image contrast and different lesion sizes, restrict the performance of CAD systems. Recently, ultra-wide-field (UWF) retinal images have become a useful tool for disease detection due to the capability of capturing much broader view of retina (i.e., up to 200°), in comparison with the most commonly used retinal fundus images (45°). In this paper, we propose an attention-based multi-branch network for the diseases classification of four different subject groups. The proposed method consists of a multi-scale feature fusion module and a dual attention module. Specifically, small-scale lesions are identified using the features extracted from the multi-scale feature fusion module. To better explore the obtained features, the dual attention module with a global attention graph is incorporated to enable the network to recognize the salient objects of interest. Comprehensive validations on both private and public datasets were carried out to verify the effectiveness of the proposed model.

INDEX TERMS Diseases classification, UWF image, multi-scale, attention, retina.

NOMENCLATURE

DR:	Diabetic retinopathy.
AMD:	Age-related macular degeneration.
RD:	Retinal detachment.
FOV:	Field-of-view.
DL:	Deep learning.
UWF:	Ultra-wide-field.
CNN:	Convolutional Neural Network.

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

I. INTRODUCTION

Visual impairment of most ocular diseases, including diabetic retinopathy (DR), age-related macular degeneration (AMD), and retinal detachment (RD), is becoming a serious issue for public health, particularly in aging populations [1], [2], [3]. Early screening through ophthalmologic examinations is important to the timely detection of early symptoms. However, the diagnostic efficiency is strongly affected by the imbalance between the limited ophthalmic resources and the ever-growing patient population. Conventionally, funuds photography with a regular field-of-view (FOV) is commonly used in clinical practice for screening multiple retinal-related diseases [4], [5]. A single image with a 30-50° FOV only partially shows the retina, and is not sufficient to observe pathological changes in the exterior regions. As such, the advent of wide-angle retinal imaging systems has become increasingly common in clinical applications. These systems can provide an angle of view of up to 200° and cover nearly the entire retina field.

Automatic localization and detection of lesions using deep learning (DL) algorithms have been shown as effective means for the detection of a variety of ocular diseases. For retinal fundus image analysis tasks such as retinal vessel segmentation [6], [7], optic disc segmentation [8], [9], artery/vein classification [10], and DR grading [11], [12], [13], DL-assisted diagnosis has been extensively studied. Several attempts were also made on both color fundus and ultra-wide-field (UWF) images by Convolutional Neural Network (CNN). For instance, Wang et al. [14] designed a dual-stream CNN specifically for multimodal AMD classification. Wang et al. [15] used a multi-stream network structure to identify 36 retinal diseases using three sub-networks to extract features from the macula, optic disc and entire fundus region, respectively. For the recognition of UWF fundus images, Oh et al. [16] designed a model for the DR detection task and demonstrated that automatic detection of DR on UWF images is feasible. Zhang et al. [17] used six image preprocessing technologies to investigate their effects on fundus abnormality prediction performance and three fundus disease models. Li et al. [18] devised a deep learning system that employs UWF images to automatically identify retinal exudates and/or drusen with high reliability.

Previous works on ocular disease classification have found that the diagnostic accuracy of DL models trained from retinal fundus images is highly dependent on the image quality and the scale of training set. Despite the advantage of larger FOV in UWF images, they however have lower image contrast and visibility, as well as uneven illuminations as shown in Figure 1. Those limitations increase the difficulty in disease classification since many lesions can be observed in the peripheral areas of UWF images, where their appearances are relatively small and even harder to be detected. Therefore, by paying a special attention on the subtle lesion detection is essential for improving disease classification performance with UWF images. In standard CNN [19], [20], the shallow convolution layer of the network has relatively higher spatial resolution but less semantic knowledge, which is still possible to learn the subtle lesion features with low visibility. On the contrary, the deep convolution layer has rich semantic information, but it is easy to ignore subtle features. Previous disease classification methods [21], [22] paid little attention to the shallow features and may result in poor performance. Thus, appropriate strategy needs to be developed to better combine shallow features with deep semantics, for a higher discrimination ability of the classification model.

Currently, the concept of visual spatial attention mechanism is applied in medical image analysis to detect representative features from images with multiple lesions, noise, and artifacts [23], [24]. By aggregating different levels of features, the attention model shows strong capability for disease classification. Inspired by the above idea, we propose a new attention-based dual-path and multi-scale feature fusion network to focus on the lesion object and thus to extract the most distinctive features for multiple disease classification tasks on UWF images.

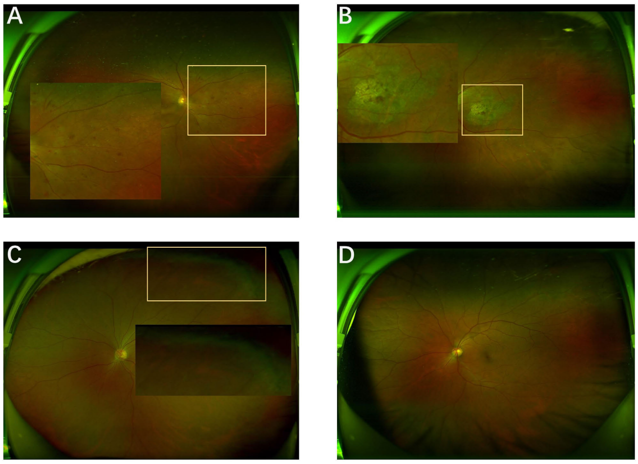


FIGURE 1. Ultra-wide-field fundus images showing typical Examples. The pathological results are highlighted with a yellow border. (A)DR, (B)AMD, (C)RD, (D)Normal.

The proposed network structure uses Resnet34 for both paths. One path is designed to use the multi-scale feature fusion module to fuse features at different scales and direct the network to focus on fine-grained lesion features. The other path works similarly but pays more attention to the global lesion features. The local and global features are combined after average pooling to provide comprehensive knowledge for the decision-making stage of the classification model. In addition, a dual attention module with both channel and spatial attention enhancement is added to the dual paths so that the model can focus on the appropriate level of detail at low contrast. Overall, the contributions of this paper can be summarized as follows:

- We introduce a dual-path attention mechanism, enabling the model to effectively focus on both small lesions features and global pathological distribution features.
- We propose a multi-scale feature fusion strategy, which can fully exploit and fuse different levels semantic information while suppressing irrelevant information.
- Both of these strategies can be used in a plug-and-play approach for any other backbone. We validate the proposed model on multiple datasets and experimental results show that our method outperforms others in classifying these different diseases.

II. RELATED WORK

A. DEEP LEARNING BASED CLASSIFICATION ON UWF IMAGES

Deep learning algorithms, especially CNN, have achieved considerable success in video and image recognition.

In particular, CNN has unique advantages in large-scale data processing, and also has excellent performance in medical image applications [25], [26], [27]. In the field of ophthalmology, several tasks have demonstrated that they can achieve comparable accuracy to clinical experts [28], [29]. Color fundus imaging is a widely used conventional imaging technique for diagnosing retinal diseases. In recent years, the UWF images have also attracted increasing attention from many researchers to further explore CNN-based image analysis and disease diagnosis techniques. Hiroki et al. [30] used a simple network to detect the pathological features of glaucoma and estimate the associated visual field defects. Nagasawa et al. [31] used a dataset of 378 UWF images in conjunction with the VGG-16 model to identify the presence of early-stage proliferative diabetic retinopathy. Ju et al. [32] devised a modified cyclic generative adversarial network (CycleGAN) to generate additional UWF images for training. Li et al. [33] designed a DL system using UWF images to detect retinal detachment and identify retinal detachment-induced macular degeneration. Cao et al. [34] proposed a cascade system for ocular disease screening with interpretability and scalability. By visualizing pathological and anatomical information, the system improves the accuracy and interpretability of disease prediction. Previous studies have shown that DL-based approaches are effective for fundus disease classification of UWF images. Although these networks are advantageous in extracting global features, they also tend to ignore important local information and are not suitable for representing discriminative detailed features.

B. ATTENTION MECHANISM IN DEEP LEARNING

The attention mechanism is a product of the study of human visual perception and can effectively focus on features in key regions. It enables deep learning models to focus on the target regions in order to learn more discriminative representations and to suppress unimportant features. The integration of visual attention mechanisms into deep learning architectures has made significant progress in image classification [35], [36], object localization [37]and semantic segmentation [38], [39]. Hu et al. [40] proposed a squeeze and excitation (SE) network, which adaptively adjusts the feature response of each channel while modeling the internal dependencies between channels. Wang et al. [41] proposed a self-attention-based operation that computes the interaction between any two locations to directly capture long-range dependencies and achieve effective classification performance. Cheng et al. [23] proposed a novel modular group attention block to capture the dependencies present in medical image features along two mutually independent dimensions, namely, channel and space. This approach produced promising results in medical image segmentation and classification tasks. He et al. [42] used global attention blocks and category attention blocks to compel the network to concentrate on the lesioned object thereby achieving improved DR classification. Xie et al. [43] introduced a novel cross-attention network based on the ResNet34 model, which uses scanning laser ophthalmoscopy (SLO) fundus images to complete the classification of three diseases and normal images. However, although they have achieved good results, they still need to fully consider the characteristics of different network scales, which is also very helpful in improving the performance.

III. PROPOSED METHOD

In this section, we propose a new attention-based dual-path multi-scale fusion network, as shown in Figure 2, which is mainly classified into Local CNN and Global CNN. Among them, Local CNN integrates features of different scales by Multi-scale Feature Fusion Module (MFFM), which can capture some small lesion areas and focus on local areas. In contrast, Global CNN allows a more comprehensive focus on all lesions. At the same time, to make better use of the extracted features, we also propose the Dual Attention Module (DAM), which allows the network to focus on the lesion area. These modules are described in detail in the following subsections.

A. MULTI-SCALE FEATURE FUSION MODULE

Different layers of the CNN can be used to capture different features. In particular, the lower layers are adept at capturing local features, while the deeper layers are capable of capturing global features of the same image. Feature fusion at different levels has proven useful in several computer vision tasks, such as semantic segmentation, classification, and detection [44]. Common fusion operations, such as addition or concatenation, are performed at the pixel level. However, due to the lack of semantic information of low-level features, the performance gain is limited. Therefore, for UWF images, to solve the salience problem and highlight the differentiated regions of different scales in fundus images, we introduce a MFFM by encoding the multi-scale features attention [45], [46]. Next, we will introduce the details of the MFFM shown in Figure 3. Specifically, We employ four ResBlocks [20]to generate four sets of feature maps X_i and X_j from the original fundus images at different resolutions, where i(i = 1,2,3) and j(j=4) denote the level of resolution. In order to retain more spatial information, we modify the down sampling step of X_j to 1. The content of the appeal considers two sources of information: a high-level feature map $X_j \in R^{C_j \times H_j \times W_j}$ and a low-level feature map $X_i \in R^{C_i \times H_i \times W_i}$. Then X_i and X_j are transformed into compact embeddings $X_i^c \in R^{C \times N_h}$ and $X_j^c \in R^{C \times S}$ using 1 × 1 convolution and spatial pyramid pooling, where $N_h = H_n \times W_n$ and S denotes pyramid pooling pixels, as shown in Figure 4. Next, X_j^c is matrix multiplied by X_i, and then softmax is applied to compute the affinity matrix M $\in R^{S \times N_h}$, and finally the fusion output $X_f \in R^{C_n \times N_h}$ is computed by the matrix product of M and X_i^{cT}, with the following equation:

$$X_f = MFFM(X_i, X_j) = \phi(\mathcal{F}_s(X_j^c X_i^{cT})X_i^c, X_j^c) \tag{1}$$

$$X = \phi(X_f^1, X_f^2, X_f^3) \tag{2}$$

where ϕ denotes a 1×1 convolution that reduces these features to a compact embedding, $\mathcal{F}_s$ denotes the *softmax*

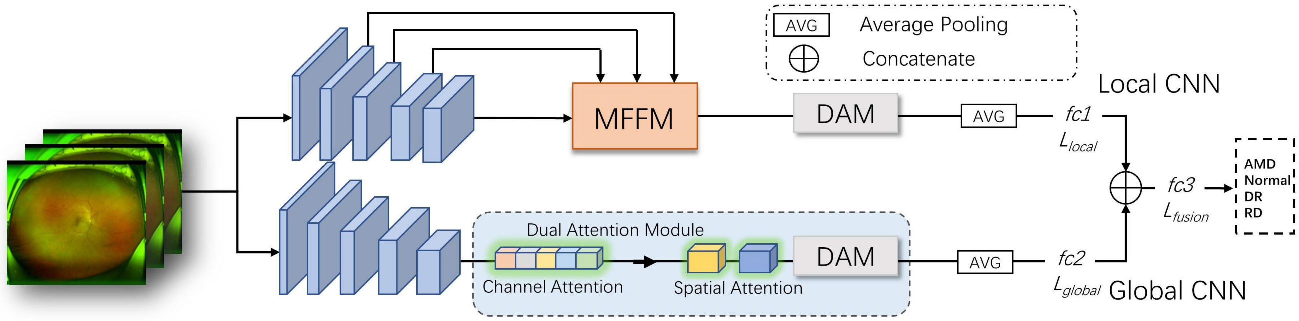


FIGURE 2. Illustration of the architecture of our proposed model which is mainly divided into Local CNN and Global CNN. Local CNN is composed of MFFM and DAM. Due to the integration of different scale features, the network can capture some small lesion areas. Global CNN is composed of DAM, which enables the network to focus on the target area and capture global information through channel and spatial attention mechanism.

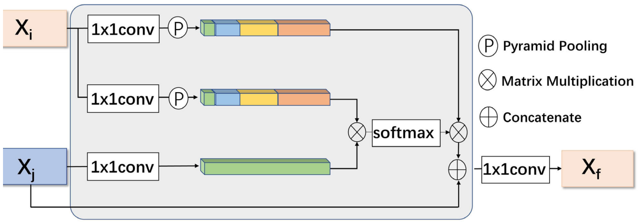


FIGURE 3. The detailed architecture of the Multi-scale Feature Fusion Module (MFFM).

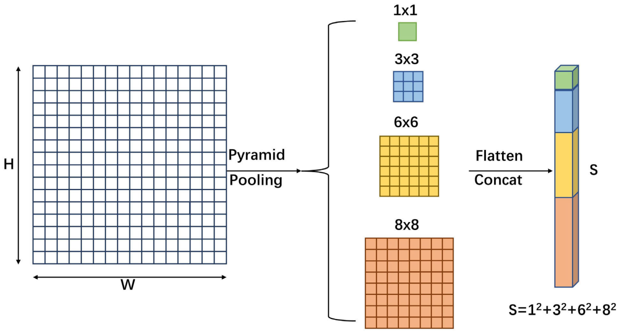


FIGURE 4. The detailed architecture of the pyramid pooling.

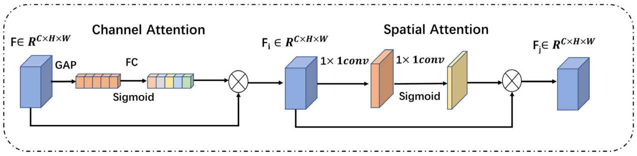

F∈ $R^{C\times H\times W}$
Channel Attention
GAP
FC
Sigmoid
F_i ∈ $R^{C\times H\times W}$
Spatial Attention
1× 1conv
1× 1conv
Sigmoid
F_j∈ $R^{C\times H\times W}$

FIGURE 5. The detailed architecture of the Dual Attention Module (DAM).

activation function. X is expressed as the last multi-scale fusion feature.

B. DUAL ATTENTION MODULE

As shown in Figure 5, the DAM includes both channel and spatial attention mechanisms. To identify and diagnose specific feature channels, this subsection relies on the SE attention mechanism [40] as the channel attention module, which learns a channel attention weight. We define the feature map as $F \in R^{C\times H\times W}$. The channel attention feature map $F_i \in R^{C^i\times H\times W}$ is computed by the following equation:

$$F_i = \sigma(W(GAP(F))) \times F \tag{3}$$

where the sigmoid function is denoted by σ, GAP denotes global average pooling, and W represents the two fully connected (FC) layers activated with Relu [47]. In the first FC layer, the number of neurons is set to rC, where r denotes the dimensionality reduction rate, which is typically set empirically to 1/16. In the subsequent FC layers, the number of neurons is set to C. Spatial attention is a mechanism that determines the importance of individual spatial locations by training spatial attention weights that work in conjunction with channel attention. Given the previously derived feature map F_i, the spatial attention feature map $F_j \in R^{C^j\times H\times W}$ is computed using the following equation:

$$F_j = \sigma(Conv(F_i)) \times F_i \tag{4}$$

where σ denotes the sigmoid function and Conv denotes a method of combining the spatial information of each channel into a single channel that consists of two 1×1 convolution operations. The first convolution operation outputs c/r channels, which are then processed using the Relu function, where c is the number of input channels and r is the reduction rate. The second 1×1 convolution operation outputs 1 channel.

IV. EXPERIMENTS AND DISCUSSIONS

A. DATASET AND EVALUATION METRICS

The UWF images used in this study are acquired by Ningbo Eye Hospital using Optos. All images are obtained in accordance with regulatory approval and patient consent. The dataset is classified by two experienced ophthalmologists, resulting in a total of 416 images of AMD, 477 images of DR, 228 images of RD, and 444 images of Normal. Therefore, we collect a total of 1565 UWF images and randomly divide them into training set and test set at a ratio of 4:1. The distribution of the data is shown in Table 1, where 1250 images are assigned to training and 315 images are assigned to testing. In addition, the performance of the proposed method is evaluated using various evaluation metrics, including Accuracy, Precision, Recall, and F1 score.

TABLE 1. The training and test set data distribution.

Class	AMD	DR	RD	Normal
Train	333	381	182	354
Test	83	96	46	90

TABLE 2. Experimental configuration.

	lr	Batch size	Epoch	Optimizer
Setting	0.0005	16	100	Adam

B. IMPLEMENTATION DETAILS

We feed the UWF images in random order to the proposed framework during the training process. To reduce overfitting, we perform data enhancement strategies including random vertical flip, horizontal flip, and random rotation. In addition, dropout is also used to mitigate overfitting. Since the pixel size of the original UWF image is 3900×3072, which is too large for the network, we reduce the size of the original image to 448×448. The Adam optimizer is used to train the network, with an initial learning rate of 0.0005 and a cosine annealing learning rate decay strategy, as shown in Table 2. The training batch size for all models is set to 16, and the number of training epochs is 100. As shown in the training curve in Figure 6, the network reaches convergence within 100 epochs. The training is performed using the PyTorch framework in Python, and all experiments are run on an Ubuntu 16.04 operating system with an NVIDIA GeForce 1080 Ti GPU. The classification loss function selected for this study is the cross-entropy function. We define three loss functions as L_{local}, L_{global}, and L_{fusion}, respectively. The final formula is as follows:

$$L_{total} = L_{local} + L_{global} + L_{fusion}, \tag{5}$$

While all of L_{local}, L_{global}, and L_{fusion} are used in the loss calculation during training, we predict the model based on L_{fusion} at test time, since L_{fusion} already contains information from L_{local} and L_{global}.

C. CLASSIFICATION PERFORMANCE COMPARISON

The experimental results of different methods are shown in Table 3. To demonstrate the superiority of our model, we use transfer learning to train the well-known deep structure, such as InceptionV3 [48], VGG16 [19], ResNet34 and ResNet50 [20] models. These networks contain the characteristics of deep networks, with InceptionV3 having the characteristics of a multi-branch architecture and the ability to perform extensive feature extraction. ResNet34 and ResNet50 can combine shallow information and deep information through residual connections. Therefore, these models are a good baseline for the study. We modify the last fully connected layer to four nodes in each model and train the model under the same settings. Based on the results in rows 1 to 4 of the table, ResNet34 outperforms the other baseline models and becomes the preferred backbone of the proposed framework. To further demonstrate the effectiveness of the proposed approach, this paper also compares some other attention-based as well as multiscale approaches (the backbone networks are all based on Resnet34). CANet [49] captures long-range correlations by embedding location information into channel attention with less computational overhead. CABNet [42] uesd the global attention block and the category attention block to improve model classification by focusing the network on lesion objects. CAMB [43] used a multi-branch network based on UWF images and combines multi-scale and attention mechanisms to improve classification of multiple diseases. Rows 5 to 8 of the table show that our proposed method achieves the most advanced results in all evaluation metrics, with an accuracy of 92.33%. For the best baseline model (Resnet34), the accuracy metric improved by 3.51% and by 1.91% over the CABNet model, which ranked second in the comparison methods. The above experimental results demonstrate the remarkable classification performance of our proposed method when applied to the diagnosis of fundus diseases in UWF images has a good application prospect.

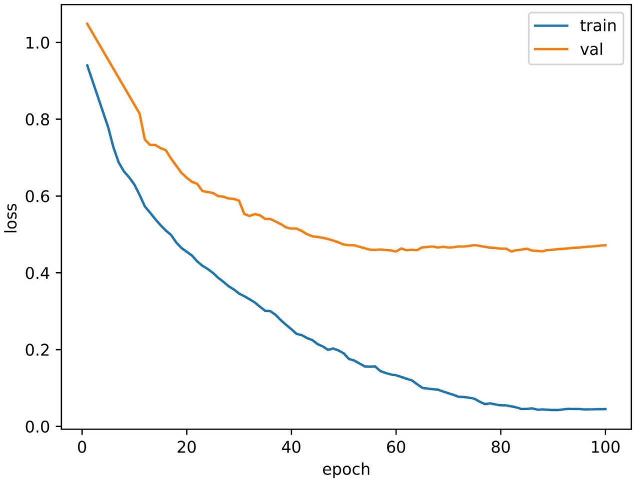


FIGURE 6. The training curves.

TABLE 3. The classification results of different methods (%).

Model	Accuracy	Precision	Recall	F1
VGG16	85.62	86.23	85.015	85.50
ResNet34	88.82	89.41	89.57	89.47
ResNet50	88.50	88.86	90.22	89.31
InceptionV3	87.86	87.88	87.97	87.77
CANet	89.46	89.74	90.46	90.14
CAMB	89.78	89.81	91.00	90.30
CABNet	90.42	90.35	91.50	90.85
Proposed	92.33	92.55	93.36	92.91

We plot receiver operating characteristic (ROC) curves and compare eight models on the test dataset, including AMD, DR, RD and Normal UWF images, as illustrated in

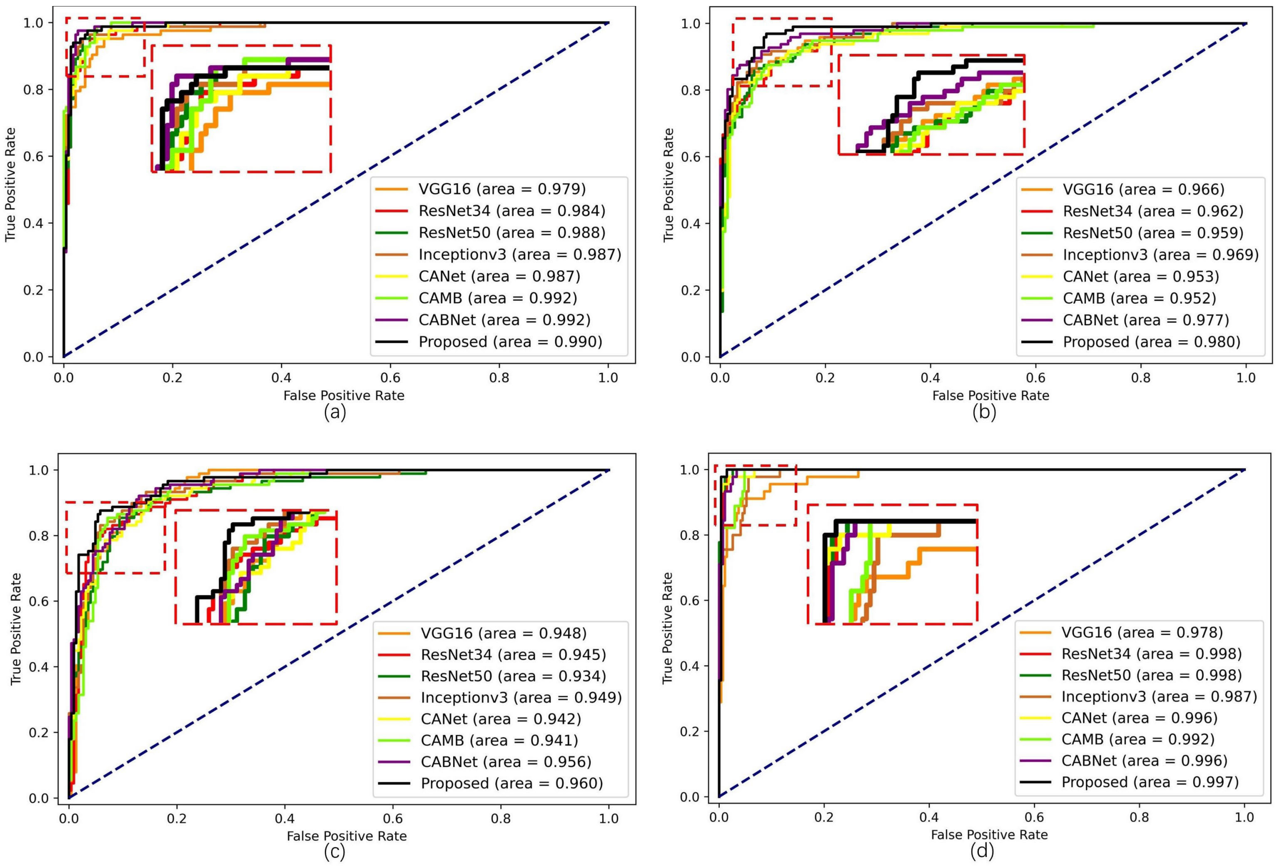


FIGURE 7. The receiver operating characteristic curve on each disease of comparison of different methods. (a)-(d) indicate the results of the AMD, DR, Normal and RD UWF images, respectively.

Figure 7. The performance of a network's classification ability is positively correlated with the size of the area covered by the ROC curve. The black curve represents our proposed method, which achieved excellent classification performance in all four prediction categories. Moreover, we can see from Figure 7 that our method significantly outperforms other comparative methods for DR fundus diseases. That's because many UWF images of DR have similar features to Normal categories. For example, the hemorrhages are not obvious in some of the DR images, while there is a lot of noise in Normal images, and these two categories can be easily misclassified. However, our two-branch multi-scale attention network can guide the network to extract DR disease-related features and capture subtle lesion regions by focusing on key target regions.

D. ABLATION STUDY

As mentioned above, we propose a dual-path multi-scale attention network consisting of Local CNN and Global CNN. Specifically, we compare the following network structures: ResNet34 network alone, Local CNN with MFFM and DAM, and Global CNN with DAM. Note that our proposed models all use ResNet34 as the backbone. The comparative ablation experiments are shown in Table 4. Compared with ResNet34, Local CNN significantly improved, mainly due to the fusion of information of different scales by MFFM. Because the shallow network possesses higher resolution and the deep network possesses richer semantic information. The network can focus more on the tiny local lesions by fusing the shallow network with the deep network. In addition, the results of Global CNN are also better than ResNet34, mainly because DAM can make the network more focused on the lesion area and capture global information. Our proposed method combines Local CNN and Global CNN, and the classification result of the fused network is much better than that of the network without fusion. The accuracy of the former network is 1.60% higher than that of the latter network. These results show that our proposed network effectively improves the classification of UWF images by fusing local information with global information discriminate lesion regions of different sizes.

To further analyze the results, we plot the confusion matrix of the contrastive ablation network, as shown in Figure 8. Based on the confusion matrix, it can be observed that the designed network can greatly improve the prediction performance of RD and DR fundus diseases, achieving an accuracy of 100% in accurately identifying RD, which is highly desirable for practical applications. In addition, we use t-SNE [50]

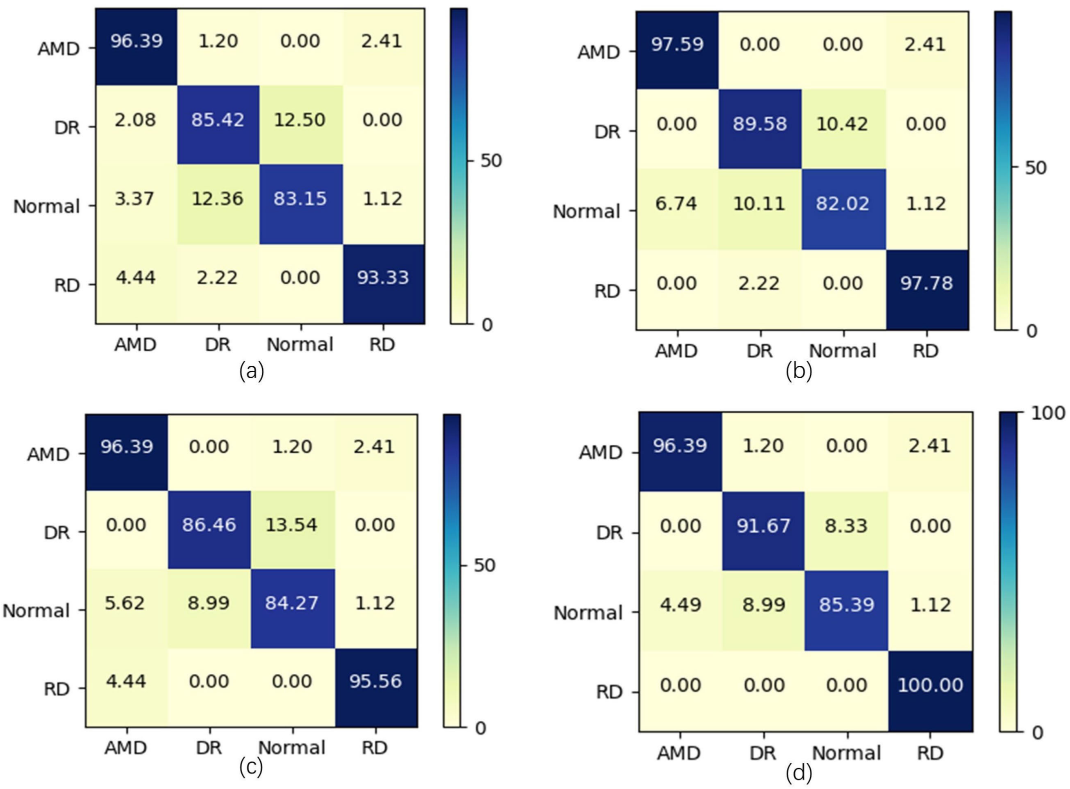


FIGURE 8. The confusion matrix of four classes classification: AMD, DR, Normal, and RD. (a)–(d) represent the results of ResNet34, Local CNN, Global CNN, and proposed methods, respectively.

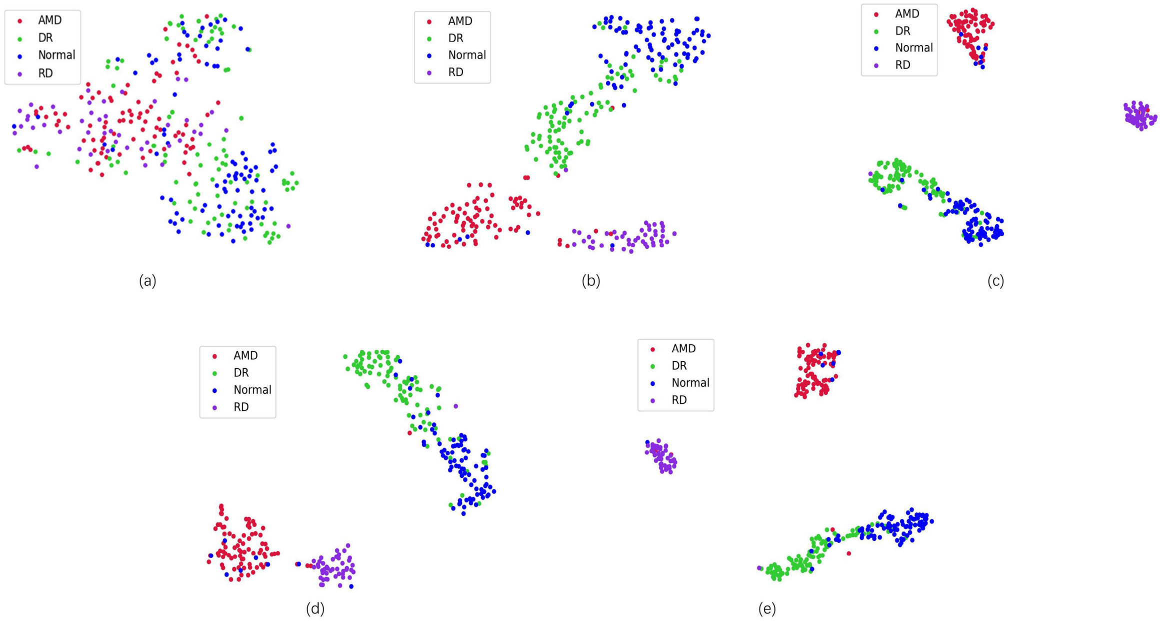


FIGURE 9. Visualization of t-SNE for different comparison networks on the test set. (a)–(e) represent the original test dataset, ResNet34, Local CNN, Global CNN and the proposed method, respectively.

for the visualization and interpretation of the high dimensional features that are acquired by the network. As shown in Figure 9, we can intuitively see that the features of different types of UWF images are mixed, so it is difficult to determine the cluster center. After the UWF image is processed by ResNet34, the clustering becomes clearer. However, its

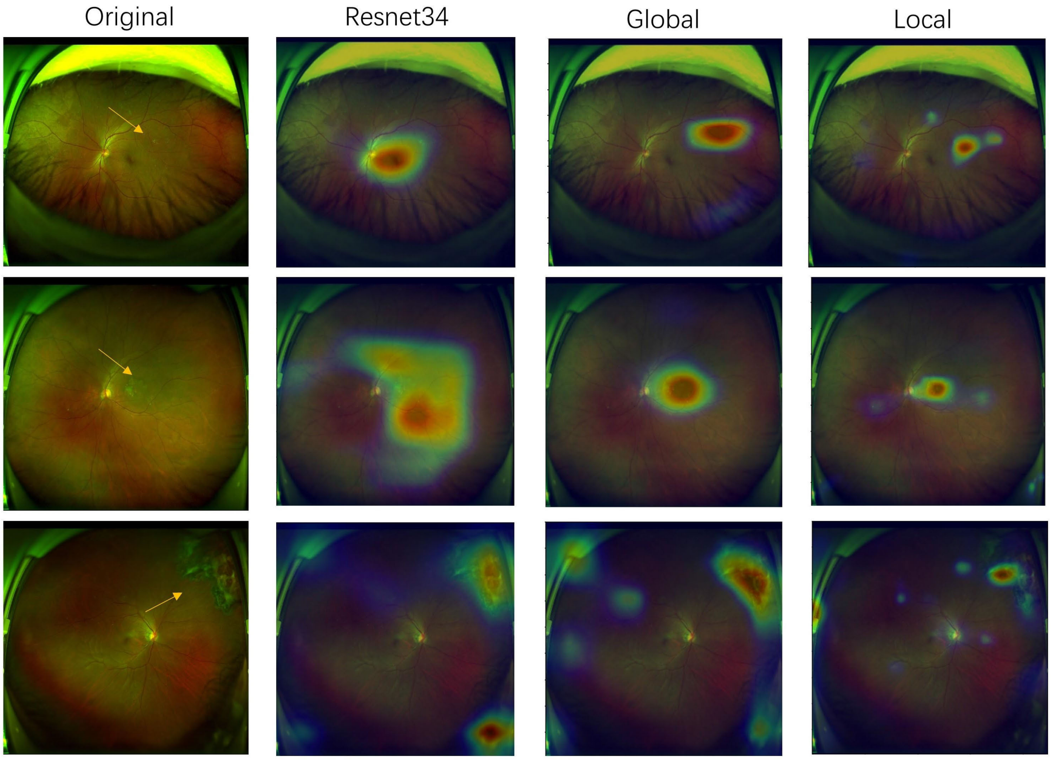


FIGURE 10. Performance Comparison of different models. We show three types of UWF images (from top to bottom, i.e. DR, AMD and RD, respectively). The first column provides the original image, where the yellow arrow indicates the lesion area. The second column provides the heatmaps without attention, the third column provides the heatmaps of Global CNN, and the four column shows the heatmaps refined by Local CNN.

TABLE 4. The test dataset classification results of the contrasting ablation network models(%).

Model	Accuracy	Precision	Recall	F1
ResNet34	88.82	89.41	89.57	89.47
Local	90.73	91.06	91.74	91.35
Global	89.50	90.17	90.94	90.52
Proposed	92.33	92.55	93.36	92.91

clustering boundary is still challenging to divide. Compared with ResNet34, our proposed network separation effect is the best, which is consistent with Table 4.

To assess the interpretability of the model and gain a better understanding of the impact of Local CNN and Global CNN, Grad CAM [51] is employed to visualize the results. As shown in Figure 10, we provide three UWF images corresponding to DR, AMD, and RD, and the yellow arrows in the first column indicate the approximate lesion areas. The third and fourth columns display the results of the Global CNN and Local CNN, respectively, both of which outperform the model without the attention mechanism in the second column. This indicates that the extracted features are effective in distinguishing between prediction categories. However, in models lacking attentional mechanisms, it may appear that irrelevant regions are highlighted, such as shown in the DR and AMD heatmaps in the second column, while for the lesioned regions of RD, they may not be covered by the heatmaps. For challenging cases, such as DR shown in the first row, the lesion heatmap generated by the Local CNN can cover the lesion region even if the lesion is small, indicating that the Local CNN is able to capture smaller lesions. As shown in the heatmap in the third column, the Global CNN can focus on the lesion region more comprehensively and avoid information redundancy through the global attention map. Thus, by combining Local CNN with Global CNN, the network can adaptively identify lesion regions of different sizes and focus on disease-specific features. This approach effectively improves the classification performance of the proposed network and provides good interpretability, which is useful for clinical diagnosis.

We compare our proposed DAM with the attention modules SENet [40] and CBAM [52], which are widely used for classification tasks. We apply these attention modules in the last feature layer of the dual-path network and present the experimental results in Table 5. SENet only uses channel attention and ignores spatial attention. Thus, it may not capture the details of local lesions. In contrast, DAM is a combination of channel and spatial attention learning, with better performance. CBAM also considers channel and spatial information, but as shown in Table 5, the classification

TABLE 5. The classification results of different attention blocks on the testing dataset(%).

Model	Accuracy	Precision	Recall	F1
W/o DAM	91.05	91.69	92.38	92.00
SENet	91.37	91.86	92.64	92.20
CBAM	91.70	92.15	92.90	92.49
DAM	92.33	92.55	93.36	92.91

TABLE 6. Comparison with state-of-the-art results on IDRiD dataset(%). * indicates that the result is obtained from [53].

Rank	Model	Accuracy Rank
1	LzyUNCC*	74.76
2	Our model	69.90
3	SKD [54]	67.96
4	SUNet [55]	65.06
5	VRT*	59.22
6	Mammoth*	55.34
7	HarangiM1*	55.34
8	AVSASVA*	54.37
9	HarangiM2*	47.57

performance of DAM is better than that of CBAM. The possible reason is that CBAM uses maximum pooling, which causes the network to be affected by background noise during classification.

E. VALIDATION

Table 6 shows the evaluation results of our model and other models on the IDRiD challenge. Our model is trained using only the data from Sub-challenge2 [53], which consists of 413 training images and 103 test images. We use the same experimental setup for the training of different models, and the input size of the images was 512 × 512. Notably, this competitive dataset covers both diabetic retinopathy and diabetic macular edema, while the references [54], [55] provide only independent diabetic retinopathy results. Moreover, it is noteworthy that all the methods compared have used external datasets to strengthen the effectiveness of their models, thus increasing the final classification accuracy of Sub-challenge2. In contrast, our proposed approach uses only the provided dataset to train the model parameters for classification. As shown in the third column of Table 6, our method outperforms other methods for smaller input sizes with an accuracy of 69.90%, which is only surpassed by the Lzyuncc [53] scheme that uses an input size of 896 × 896. Experimental results show that our proposed method retains good classification performance on other modal data, which demonstrates the robustness and generalizability of our method.

V. CONCLUSION AND FUTURE WORK

In this article, we present a novel dual-path and multi-scale enhanced attention network for retinal multi-disease classification using UWF fundus images. The network introduces a dual-path attention mechanism, which enables the model to effectively focus on small lesion features and global pathological distribution features through a multi-scale feature fusion module and a dual-attention module. The multi-scale feature fusion module enables the network to explore and fuse semantic information at different levels, highlighting different scales of difference regions in fundus images and thus identifying small-scale lesions. Furthermore, the dual-attention module with a global attention map is added after the last feature layer of the dual-path network to better explore the obtained features and focus on lesion regions while avoiding information redundancy. The experimental results show that the proposed framework can effectively classify three different types of fundus diseases as well as normal images, and achieves a relatively good classification performance. The ablation experiments confirm that the developed module can effectively improve the classification performance of the network. In the future, we will focus on improving the classification performance by strengthening the data collection and pre-processing of UWF images. In addition, since the conventional color fundus datasets are relatively large while the available UWF data is very limited, we thus may also consider to utilize transfer learning techniques to improve the classification performance on UWF images by taking advantage of the well-labeled color fundus data.

ACKNOWLEDGMENT

(Fangsheng Chen and Shaodong Ma are co-first authors.)

REFERENCES

[1] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. R. Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 128, pp. 271–281, Apr. 2018.

[2] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," *Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[3] None, "The repair of rhegmatogenous retinal detachments," *Ophthalmology*, vol. 103, no. 8, pp. 1313–1324, 1996.

[4] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, 2010.

[5] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Prog. Retinal Eye Res.*, vol. 67, pp. 1–29, Nov. 2018.

[6] Y. Tan, K.-F. Yang, S.-X. Zhao, and Y.-J. Li, "Retinal vessel segmentation with skeletal prior and contrastive loss," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2238–2251, Sep. 2022.

[7] A. Khandouzi, A. Ariafar, Z. Mashayekhpour, M. Pazira, and Y. Baleghi, "Retinal vessel segmentation, a review of classic and deep methods," *Ann. Biomed. Eng.*, vol. 50, no. 10, pp. 1292–1314, 2022.

[8] H. Xiong, S. Liu, R. V. Sharan, E. Coiera, and S. Berkovsky, "Weak label based Bayesian U-Net for optic disc segmentation in fundus images," *Artif. Intell. Med.*, vol. 126, Apr. 2022, Art. no. 102261.

[9] L. Wang, H. Liu, Y. Lu, H. Chen, J. Zhang, and J. Pu, "A coarse-to-fine deep learning framework for optic disc segmentation in fundus images," *Biomed. Signal Process. Control*, vol. 51, pp. 82–89, May 2019.

[10] M. R. K. Mookiah, S. Hogg, T. J. MacGillivray, V. Prathiba, R. Pradeepa, V. Mohan, R. M. Anjana, A. S. Doney, C. N. A. Palmer, and E. Trucco, "A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101905.

[11] V. Deepa, C. S. Kumar, and T. Cherian, "Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6255–6265, Sep. 2022.

[12] P. K. Chaudhary and R. B. Pachori, "Automatic diagnosis of different grades of diabetic retinopathy and diabetic macular edema using 2-D-FBSE-FAWT," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[13] P. Porwal, "IDRiD: Diabetic retinopathy–segmentation and grading challenge," *Med. image Anal.*, vol. 59, Jan. 2020, Art. no. 101561.

[14] W. Wang, Z. Xu, W. Yu, J. Zhao, and X. Li, *Two-Stream CNN With Loose Pair Training for Multi-Modal AMD Categorization*. Cham, Switzerland: Springer, 2019.

[15] X. Wang, L. Ju, X. Zhao, and Z. Ge, "Retinal abnormalities recognition using regional multitask learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*. Springer, 2019, pp. 30–38.

[16] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-widefield fundus images," *Sci. Rep.*, vol. 11, no. 1, p. 1897, Jan. 2021.

[17] W. Zhang, X. Zhao, Y. Chen, J. Zhong, and Z. Yi, "DeepUWF: An automated ultra-wide-field fundus screening system via deep learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 2988–2996, Aug. 2021.

[18] Z. Li, C. Guo, D. Nie, D. Lin, T. Cui, Y. Zhu, C. Chen, L. Zhao, X. Zhang, M. Dongye, D. Wang, F. Xu, C. Jin, P. Zhang, Y. Han, P. Yan, and H. Lin, "Automated detection of retinal exudates and drusen in ultra-widefield fundus images based on deep learning," *Eye*, vol. 36, no. 8, pp. 1681–1686, Aug. 2022.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] X. Wang, Y. Yuan, D. Guo, X. Huang, Y. Cui, M. Xia, Z. Wang, C. Bai, and S. Chen, "SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102459.

[22] H. Ríos et al., "A deep learning model for classification of diabetic retinopathy in eye fundus images based on retinal lesion detection," in *Proc. 17th Int. Symp. Med. Inf. Process. Anal.*, vol. 12088. Bellingham, WA, USA: SPIE, 2021, pp. 253–260.

[23] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, and H. Lu, "ResGANet: Residual group attention network for medical image classification and segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102313.

[24] Z. Fu, J. Li, and Z. Hua, "MSA-Net: Multiscale spatial attention network for medical image segmentation," *Alexandria Eng. J.*, vol. 70, pp. 453–473, May 2023.

[25] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[26] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102444.

[27] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.

[28] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016.

[29] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Med.*, vol. 25, no. 1, pp. 30–36, Jan. 2019.

[30] H. Masumoto, H. Tabuchi, S. Nakakura, N. Ishitobi, M. Miki, and H. Enno, "Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity," *J. Glaucoma*, vol. 27, no. 7, pp. 647–652, 2018.

[31] T. Nagasawa, H. Tabuchi, H. Masumoto, H. Enno, M. Niki, Z. Ohara, Y. Yoshizumi, H. Ohsugi, and Y. Mitamura, "Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatment-naïve proliferative diabetic retinopathy," *Int. Ophthalmol.*, vol. 39, no. 10, pp. 2153–2159, Oct. 2019.

[32] L. Ju, X. Wang, X. Zhao, P. Bonnington, T. Drummond, and Z. Ge, "Leveraging regular fundus images for training UWF fundus diagnosis models via adversarial learning and pseudo-labeling," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2911–2925, Oct. 2021.

[33] Z. Li, C. Guo, D. Nie, D. Lin, Y. Zhu, C. Chen, X. Wu, F. Xu, C. Jin, X. Zhang, H. Xiao, K. Zhang, L. Zhao, P. Yan, W. Lai, J. Li, W. Feng, Y. Li, D. S. Wei Ting, and H. Lin, "Deep learning for detecting retinal detachment and discerning macular status using ultra-widefield fundus images," *Commun. Biol.*, vol. 3, no. 1, p. 15, Jan. 2020.

[34] J. Cao, K. You, J. Zhou, M. Xu, P. Xu, L. Wen, S. Wang, K. Jin, L. Lou, Y. Wang, and J. Ye, "A cascade eye diseases screening system with interpretability and expandability in ultra-wide field fundus images: A multicentre diagnostic accuracy study," *eClinicalMedicine*, vol. 53, Nov. 2022, Art. no. 101633.

[35] Y. Cui, F. Liu, X. Liu, L. Li, and X. Qian, "TCSPANet: Two-staged contrastive learning and sub-patch attention based network for PolSAR image classification," *Remote Sens.*, vol. 14, no. 10, p. 2451, May 2022.

[36] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, May 2020.

[37] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2219–2228.

[38] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, and X. Wen, "CANet: Co-attention network for RGB-D semantic segmentation," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108468.

[39] Y. Su, W. Liu, M. Cheng, Z. Yuan, and C. Wang, "Local fusion attention network for semantic segmentation of building facade point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[42] A. He, T. Li, N. Li, K. Wang, and H. Fu, "CABNet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 143–153, Jan. 2021.

[43] H. Xie, X. Zeng, H. Lei, J. Du, J. Wang, G. Zhang, J. Cao, T. Wang, and B. Lei, "Cross-attention multi-branch network for fundus diseases classification using SLO images," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102031.

[44] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2169–2178.

[45] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3300–3310.

[46] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 593–602.

[47] G. E. Hinton, A. Krizhevsky, and I. Sutskever, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, nos. 1106–1114, p. 1.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[49] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[50] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[52] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018.

[53] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, and F. Mériaudeau, "IDRiD: Diabetic retinopathy—Segmentation and grading challenge," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101561.

[54] L. Luo, D. Xue, and X. Feng, "Automatic diabetic retinopathy grading via self-knowledge distillation," *Electronics*, vol. 9, no. 9, p. 1337, Aug. 2020.

[55] Z. Tu, S. Gao, K. Zhou, X. Chen, H. Fu, Z. Gu, J. Cheng, Z. Yu, and J. Liu, "SUNet: A lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1378–1382.

FANGSHENG CHEN is currently pursuing the master's degree with the School of Mechanical Engineering, Zhejiang University of Technology. His research interests include medical image processing and artificial intelligence.

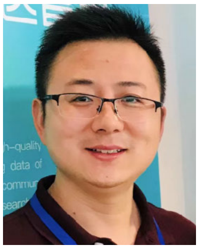
SHAODONG MA is currently an Assistant Professor with the Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China. His research interests include medical image processing and computer-aided diagnosis.

JINKUI HAO is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, University of Chinese Academy of Sciences, Ningbo, China. His research interests include medical image processing and deep learning.

MENGTING LIU received the Ph.D. degree in biomedical engineering from Louisiana Tech University. He is currently an Associate Professor with the School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China. His research interests include neuroimaging analysis, brain disease diagnosis, cognitive neuroscience, and deep learning.

YUANYUAN GU is currently an Engineer with the Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China. His research interest includes medical image processing and analysis.

QUANYONG YI received the M.D. degree in ophthalmology from Soochow University, Soochow, China. He has extensive clinical experience in complex retinal detachment, macular disease, diabetic retinopathy, complex ocular trauma surgery, cataract ultrasound-emulsification combined with IOL implantation, minimally invasive combined anterior, and posterior segment surgery. He has published more than 40 papers.



JIONG ZHANG received the Ph.D. degree from the IMAG/e Group, Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands, in 2017. After, he moved to the U.S. and continued his Postdoctoral Research with the Laboratory of NeuroImaging (LONI), Keck School of Medicine, University of Southern California (USC), Los Angeles, USA. He is currently an Associate Professor with the Group of Intelligent Medical Imaging (iMED), Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences (CAS). In the last few years, he has published over 70 papers in leading medical imaging journals and conferences. His research interests include AI-based ophthalmic image processing, ophthalmic microvascular and cerebrovascular analysis, shape analysis, and neuroimage analysis. He was actively involved in many academic services, including the Publicity Chair of the ICCCV, in 2022; and a Program Committee Member of AAAI (2022), CVPR Workshop on CVPM (2019–2023), ICCV Workshop on CVPM (2019), MICCAI Workshop on OMIA (since 2017), and VSIP (2021).

YITIAN ZHAO (Member, IEEE) received the Ph.D. degree in 3-D image analysis from the Department of Computer Science, Aberystwyth University, in 2013. He is currently the Director and a Professor of the Laboratory of Intelligent Medical Imaging (iMED), Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences. His research interests include ophthalmic medical image processing, vessel structure analysis, and eye and brain joint computing. He was the Area Chair of MICCAI (2021 and 2022); a Committee Associate Member of IEEE BISP (2021–2022); the Program Co-Chair of ICIMH (2021) and VSIP (2021); and a Program Committee Member of AAAI (2021), MICCAI Workshop on OMIA (since 2017), MIUA (2019), and MICAD (2021). He has served as an Associate Editor for IEEE Transactions on Medical Imaging and the editorial board member (specialty with ophthalmology) for *Scientific Reports* (since 2017).