Medical Image Analysis 71 (2021) 102031

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Cross-attention multi-branch network for fundus diseases classification using SLO images

Hai Xie[a], Xianlu Zeng[b], Haijun Lei[c], Jie Du[a], Jiantao Wang[b], Guoming Zhang[b,*], Jiuwen Cao[d], Tianfu Wang[a], Baiying Lei[a,*]

[a] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
[b] Shenzhen Eye Hospital, Shenzhen Key Ophthalmic Laboratory, Health Science Center, Shenzhen University, The Second Affiliated Hospital of Jinan University, Shenzhen, China
[c] Guangdong Province Key Laboratory of Popular High-performance Computers, School of Computer and Software Engineering, Shenzhen University, Shenzhen, China
[d] Key Lab for IOT and Information Fusion Technology of Zhejiang, Artificial Intelligence Institute, Hangzhou Dianzi University, Hangzhou, China

ARTICLE INFO

Article history:
Received 13 October 2020
Revised 24 January 2021
Accepted 3 March 2021
Available online 10 March 2021

Keywords:
Fundus diseases classification
SLO
Multi-branch network
ASPP
Depth-wise attention
Cross-attention

ABSTRACT

Fundus diseases classification is vital for the health of human beings. However, most of existing methods detect diseases by means of single angle fundus images, which lead to the lack of pathological information. To address this limitation, this paper proposes a novel deep learning method to complete different fundus diseases classification tasks using ultra-wide field scanning laser ophthalmoscopy (SLO) images, which have an ultra-wide field view of 180–200°. The proposed deep model consists of multi-branch network, atrous spatial pyramid pooling module (ASPP), cross-attention and depth-wise attention module. Specifically, the multi-branch network employs the ResNet-34 model as the backbone to extract feature information, where the ResNet-34 model with two-branch is followed by the ASPP module to extract multi-scale spatial contextual features by setting different dilated rates. The depth-wise attention module can provide the global attention map from the multi-branch network, which enables the network to focus on the salient targets of interest. The cross-attention module adopts the cross-fusion mode to fuse the channel and spatial attention maps from the ResNet-34 model with two-branch, which can enhance the representation ability of the disease-specific features. The extensive experiments on our collected SLO images and two publicly available datasets demonstrate that the proposed method can outperform the state-of-the-art methods and achieve quite promising classification performance of the fundus diseases.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The eyes are called the windows of the heart and represent the primary sense of the human-beings. Hence, fundus diseases affect the lives of people greatly. The three fundus diseases(i.e., Coats, retinitis pigmentosa (RP), diabetic retinopathy (DR)) have a great impact on the vision as the analysis objects (Aaberg and Shevchenko, 2019; Cohen and Gardner, 2016; Hamel, 2006). For example, the Coats disease mostly occurs in adolescence (Aaberg et al., 2019), if not found and treated in time, it can easily cause retinal detachment or even blindness over time. Therefore, many researchers are devoted to studying the methods of detecting fundus diseases, which can further assist the doctors to make precise decisions. However, most existing methods detected the fundus diseases using the RetCam fundus images (Coyner et al., 2019; Zhao et al., 2019), which only provides single angle scan that is difficult to detect the fundus diseases being located in the regions beyond 45°. To address this situation, some researchers selected the ultra-wide field scanning laser ophthalmoscopy (SLO) images (Webb and Hughes, 1981) as the studying objects (Ashok and Murugesan, 2017; Haleem et al., 2014). The main reason of using the SLO images is that, they are with an ultra-wide field view of 180–200°, where more fundus regions can be included via single scans image than the fundus images from RetCam. Thus, the cost and computing resources can be greatly reduced. For instance, Pellegrini et al. proposed a fully automated artery/vein classification technique for the ultra-wide field view fundus images (Pellegrini et al., 2017).

* Corresponding authors.
E-mail addresses: 13823509060@163.com (G. Zhang), leiby@szu.edu.cn (B. Lei).

https://doi.org/10.1016/j.media.2021.102031
1361-8415/© 2021 Elsevier B.V. All rights reserved.

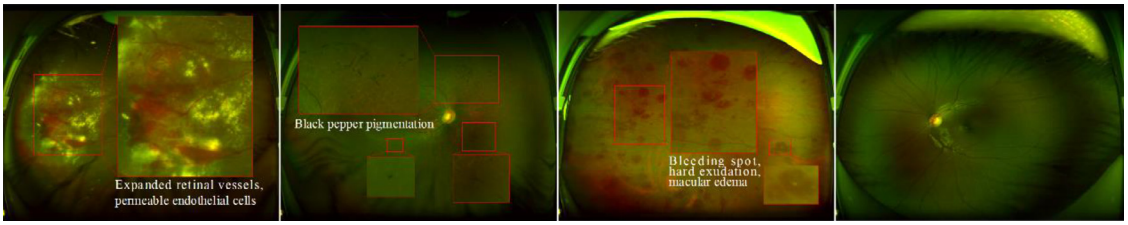


Fig. 1. Illustration of the selected diseased and normal SLO images. From left to right: Coats, RP, DR and Normal image, where the lesion regions are enlarged and the characteristics of different diseases are also illustrated.

Chen et al. realized automatic dewarping of retinal images to improve image quality by using adaptive optics confocal SLO images (Chen et al., 2019). Nevertheless, these methods utilized the traditional hand-crafted feature information to complete certain task, which are limited to improve the performance.

In clinical practice, the Coats is characterized by expanded retinal vessels with more permeable endothelial cells, RP appears with the black pepper pigmentation and the blood vessels are thin, while the pathological changes of DR are expressed with bleeding spot, hard exudation, vein beading, and macular oedema. An example to illustrate the characteristics of the three selected fundus diseases using SLO images is shown in Fig. 1. From Fig. 1, we can see that the contrast of the lesion and other normal areas is low and there are some similarities in the shape of lesions for different diseases. Such these problems put great challenges to computer vision analysis tasks for SLO images.

In addition, some lesion information appears to be very small compared to the global image for SLO images, which causes the background regions to be more obvious than the target areas. In order to extract discriminative features of the objects, many researchers tried to use the deep learning models to extract the high-level features (Badar et al., 2020; Matsuba et al., 2019). Nagasato et al. employed the VGG-16 model (Simonyan and Zisserman, 2014) to extract the feature information for detecting central retinal vein occlusion using ultra-wide field SLO images (Nagasawa et al., 2019). Due to the lack of leveraging the feature information from the shallow layers, VGG-16 model is limited to extract the discriminative features. In order to solve this problem, some studies with relative complex deep neural network model were proposed to detect the fundus diseases using SLO images (Cunefare et al., 2018; Hamwood et al., 2019; Meyer et al., 2017). For example, Hamwood et al. proposed a fully convolutional network (FCN) based on U-Net (Ronneberger et al., 2015) architecture to determine the locations of cone photoreceptors within adaptive SLO images (Hamwood et al., 2019), where the feature information is fully used to improve the performance. However, these network models tend to extract global features and ignore some important local contextual features, which is not conducive toexpressing semantic features of the salient objects. For the reason that the local contextual features can express texture information of the specific target, which is vital for extracting discriminative features. Besides, the attention mechanism can enforce the network to focus on the salient target area to extract the features related to the specific disease.

To address the above-mentioned problems and improve the detection performance of fundus diseases for SLO images, we propose a novel deep network model to extract high-level feature information and focus on the lesion objects. The proposed model consists of the following modules: multi-branch ResNet-34 model (He et al., 2016), atrous spatial pyramid pooling (ASPP) module (Chen et al., 2017), depth-wise attention, and cross-attention modules. In detail, the ResNet-34 model is selected as the backbone to extract features, where the residual connection enables the model to learn the global and local feature information so that the extracted features are finer than those extracted using straightforward networks. Due to the low contrast and the similar appearance amongst different diseases, a network with the capacity of extracting multiple deep features from different subnetworks can be needed. Hence, in order to obtain more deep features, we design a multi-branch network by copying the path layer3 and layer4 of ResNet-34 model (called Path 1 and Path 2). In addition, an original ResNet-34 model is added to the whole framework to extract deep features that are used as the input of the depth-wise attention to retain the original deep feature from ResNet-34 (i.e., Path 3). The obtained deep high-level features are with small size, which contain insufficient feature information. Thus, the ASPP module is used to extract high-level contextual information by increasing the receptive field with different dilated rates after layer3 and layer4. Because the Path 3 uses the pre-trained parameters on ImageNet dataset, which can guide the network to learn the general shallow information quickly. Therefore, the devised depth-wise attention module between multiple subnetworks can help the network to pay attention to the salient objects. For Path 1 and Path 2, the trained parameters of the last two feature extraction layers are different, which can cause the situation that the learned features can be redundant and complementary. To make full use of the features from the two paths, we also design a novel attention mechanism (i.e., cross-attention), which is conducive to enabling the network to focus on the lesion regions and appearance by learning the differentiation and relationship amongst different diseases to avoid redundant information and realize information complementation by cross-connection fusion mode.

In summary, the contributions of this paper are listed as follows:

- We develop a multi-branch network based on ResNet-34 model, which can extract more deep level features than single network. In addition, an ASPP module is leveraged to refine the extracted deep level features by expanding the receptive field.
- A depth-wise attention module is devised to fuse the extracted high-level features, which is used to fuse the features from ASPP module and the last feature extraction layer of the original ResNet-34 model that enforces the network to focus on the salient targets of interest.
- We also devise a cross-attention module that learns the distinction and relationship amongst different diseases by channel and spatial attention strategy and integrates the extracted attention map via cross-fusion mode to gain the relevant features of specific-disease.

The rest of this paper is organized as below. Section 2 describes the related work about multi-branch network and attention mechanism in the medical image applications. Section 3 illustrates in details the principle of the proposed method. The experimental results are given in the Section 4. Section 5 presents the discussions of this paper. Finally, Section 6 gives the conclusion of this paper.

2. Related work

2.1. Multi-branch network

To improve the representation ability of the extracted features, many researchers designed multiple networks to capture more features (Chen et al., 2018; Fu et al., 2018b; Ke et al., 2018; Lai et al., 2019; Li et al., 2018; Sun et al., 2019; Yu et al., 2018). For example, Sun et al. utilized separate parallel reconstruction networks to extract features and a deep feature sharing network was adopted to reduce the number of network parameters to complete the multi-contrast MRI data reconstruction (Sun et al., 2019). To improve the brain MR imaging segmentation performance, (Wei et al., 2019) proposed a multi-model, multi-size and multi-view (called M3Net) deep neural network, where three identical modules to segment transaxial, coronal, and sagittal MR slices, respectively, and each module consists of multi-size U-Nets and multi-size back propagation neural networks. Fu et al. devised a novel disc-aware ensemble network to handle automatic glaucoma screening, which integrated the deep hierarchical context information of the global fundus image and the local optic disc region (Fu et al., 2018a). Yang et al. developed an ensemble of multi-scale convolution networks to classify hematoxylin–eosin stained breast histopathological microscopy images (Yang et al., 2019), which converted each image to multiple scales, and then used the training patches cropped and augmented at each scale to fine-tune the pre-trained three backbone networks, respectively.

2.2. Attention mechanism

To enable the network to focus the object areas and extract the discriminative features, the attention mechanisms attracted great interests of researchers (Mou et al., 2019; Nie et al., 2019; Shaikh et al., 2019; Zhang et al., 2019a). For instance, Schlemper et al. developed a novel attention gate model for medical image analysis, which could learn to suppress the irrelevant regions in an input image while highlighting salient features for a special task (Schlemper et al., 2019). In order to eliminate influence of the irrelevant noisy areas, Guan et al. proposed to integrate the global and local cues into a three-branch attention guided convolution neural network to identify thorax diseases and achieved promising performance (Guan et al., 2020). Zhang et al. proposed an attention residual learning convolutional neural network model for skin lesion classification in dermoscopy images to address the issue with the lack ability to focus on semantically meaningful lesion parts (Zhang et al., 2019a). In order to learn rich structural information, Zhang et al. designed an attention guided network to preserve the structural information and guide the expanding operation, where the attention block is leveraged to exclude the noise and reduce the negative influence of background (Zhang et al., 2019b).

3. Methodology

This paper proposes a novel multi-branch network based on ResNet-34 model to extract high-level features, where two attention mechanisms are devised to fuse the extracted features to obtain the discriminative features that are utilized to improve the classification performance of the different diseases for SLO images. The proposed framework is illustrated in Fig. 2. From Fig. 2, we can observe that our proposed network architecture consists of four parts: the multi-branch network, ASPP, depth-wise attention, and cross-attention module. We will explain these modules in details in the following subsections.

3.1. Multi-branch network and ASPP module

With the improvement of image resolution and complexity, more and more multi-branch networks were developed to extract rich feature information (Ke et al., 2019; Lin et al., 2019; Liu et al., 2019). Inspired by (Li et al., 2019), we propose a multi-branch network based on ResNet-34 model to accomplish the specific disease classification task for SLO images. As shown in Fig. 2, we can see that our multi-branch network consists of three paths: Path 1, Path 2, and Path 3. Alike (Wu et al., 2019), we build another branch network after layer2 by copying the layer3 and layer4 of ResNet-34 model, thus, the parameters can be shared before the layer3 so that the total parameters of the whole network can be reduced. The Path 3 is the original ResNet-34 model that takes the original SLO images as input to extract deep features. With the proposed multi-branch network, we can extract affluent deep features for the subsequent network modules.

The atrous convolution theory is widely applied to maintain the feature resolution unchanged by increasing the receptive field in computer vision tasks (Chen et al., 2014). Let $f_{out}[i]$ represent the output feature and $f_{in}[i]$ denote input feature, the atrous convolution can be defined as follows:

$$f_{out}[i] = \sum_{k=1}^{K} f_{in}[i + d*k]*\omega[k], \tag{1}$$

where d is the dilated rate, $\omega[k]$ denotes the k-th parameter of filter, and K is the filter size. From Eq. (1), we can infer that we can obtain different receptive fields with different dilated rates.

For SLO images, the salient object areas are small in contrast to the background, therefore, a feature extractor with multi-scale is demanded to expand the reception field. Inspired by (Chen et al., 2018; Yang et al., 2018), we recommend ASPP to accomplish the operation, where several atrous convolution layers with different dilation rates and these atrous convolution layers take the last feature extraction layer of ResNet-34 model as the input. With the obtained multiple scale features with different reception fields, we concatenate these features together to gain the high-level contextual features. We define the $F_{K,d}(f_{in})$ as an atrous convolution and f_{in} denotes the input feature. The mathematical definition of ASPP is represented as follows:

$$F_A = F_{3,1}(f_{in}) + F_{3,6}(f_{in}) + F_{3,12}(f_{in}) + F_{3,18}(f_{in}), \tag{2}$$

The structure of ASPP is shown in Fig. 3. From Fig. 3, we can see that we add a point convolution layer with the kernel size of 1×1 after every atrous convolution layer and integrate these multi-scale features by concatenation operation. Thus, we can acquire multi-scale contextual feature information with different receptive fields, further improve the classification performance for SLO images.

3.2. Depth-wise attention module

For SLO images, there exists much interference and the contrast is very low between the objects and background. Therefore, in order to fuse the features from multi-branch networks and enforce the network to focus on the salient target areas, inspired by (Schlemper et al., 2019), we devise a novel strategy, called depth-wise attention module. The depth-wise attention module takes the features from the original ResNet-34 model and ASPP modules as the input and output a fusion attention map F_{out}, as shown in Fig. 4. We define the features from the original ResNet-34 model and ASPP modules as F_R and F_A, respectively. To reduce the parameters of network model, we select the pointwise convolution to extract features, which is represented by P_{conv}. Thus, the depth-wise attention module can be expressed as follows:

$$F_{att} = \sigma_2(p_{conv}(\sigma_1(p_{conv}(F_R) + p_{conv}(F_A)))), \tag{3}$$

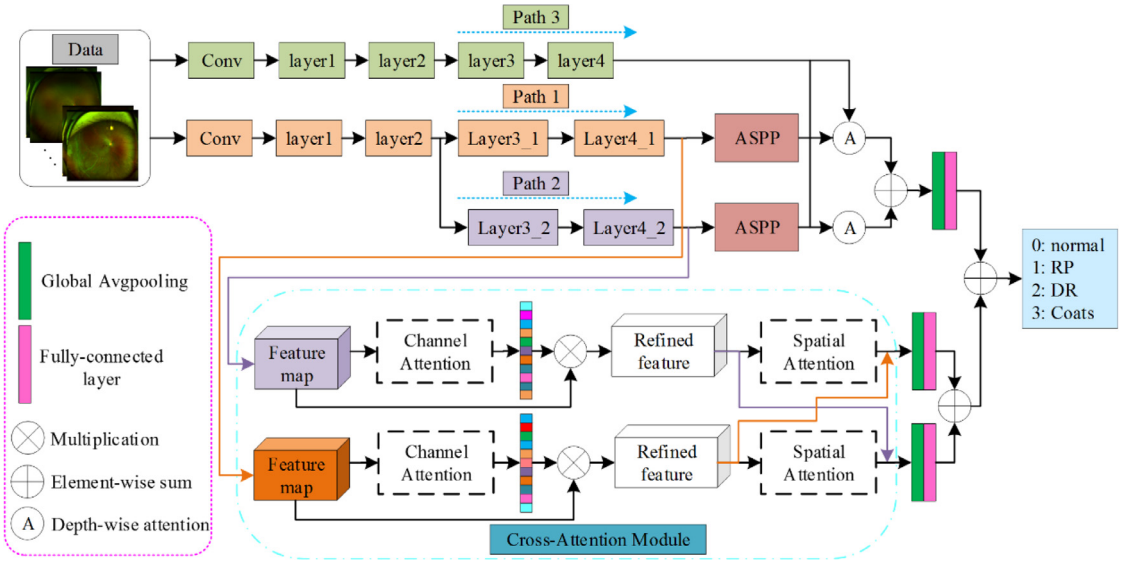


Fig. 2. Architecture of the proposed model. Path 1, Path 2, and Path 3 are employed to extract multiple features that are used as the inputs of the subsequent modules. ASPP module is utilized to express multi-scale features by different dilated rates so that the abstract feature information can be kept in the processing of encoding. The depth-wise attention module is designed to fuse the extracted features from the layer4 of Path 3 and the ASPP modules of Path 1 and Path 2. The cross-attention strategy is proposed to integrate the deep features from the Path 1 and Path 2 by channel and spatial attention mechanism via cross-fusion mode.

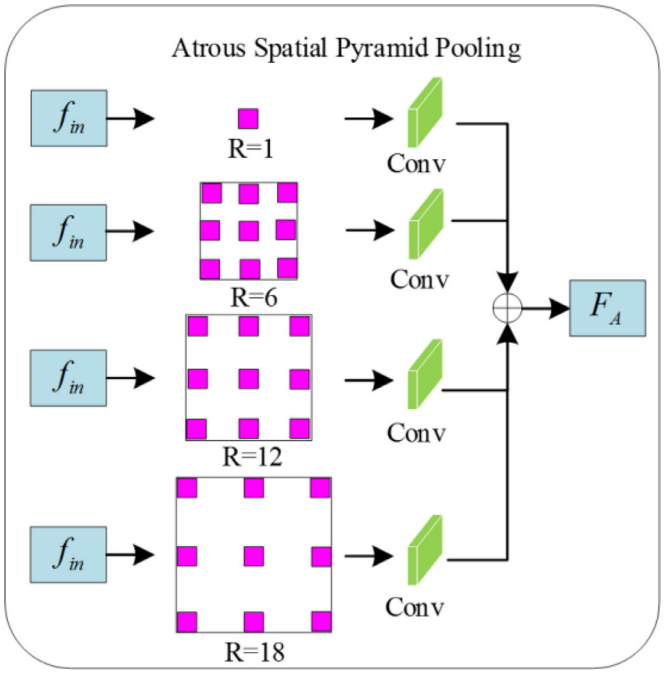


Fig. 3. Structure of the proposed ASPP module. We take the last feature extraction layer of ResNet-34 model as the input of atrous convolution with different dilation rates. After pointwise convolution, the features are concatenated to get the ASPP feature F_A.

$$F_{out} = F_{att} \times F_A, \tag{4}$$

where F_{att} denotes the obtained attention maps. $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ indicate the ReLU and Sigmoid activation function, respectively. F_{out} represents the final attention feature maps, which are followed by the global adaptive average pooling layer and fully connection layer that output the prediction probability of the specific disease. Consequently, we can utilize the depth-wise attention module to gain the discriminative high-level features by the way of deep supervision.

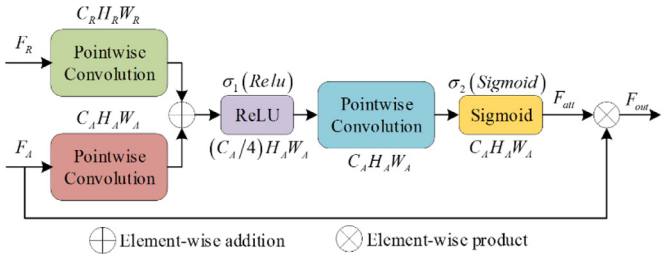


Fig. 4. Structure of the proposed depth-wise attention module. This module takes the features from layer4 of the original ResNet-34 model and two ASPP modules as inputs. These features are added after pointwise convolution, where the channel reduction rate is set to 4. After ReLU, pointwise convolution and Sigmoid, we get the attention map with the same size as the original input features. Finally, the extracted attention is multiplied by the features from ASPP module to obtain the final attention map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Cross-attention module

There are different characteristics for different diseases. To learn the distinction and relationship of each disease for SLO images, an attention mechanism that focuses on the salient features of each disease is needed. In (He et al., 2020; Li et al., 2019; Woo et al., 2018), the authors proposed a channel and spatial attention to explore the discriminative feature in view of channel and spatial. Following this view, we devise a cross-attention module to enforce the network to pay attention to the feature of specific disease by jointly using the channel and spatial attention strategy. As shown in Fig. 5, the channel attention mechanism is similar to SENet (Hu et al., 2018). However, the SENet only uses a global pooling operation, which is limited to express the high-level feature. Hence, we use the channel attention module with two pooling operations (i.e. average and max) to express more high-level features. For the reason that the features from max pooling can reflect the importance of most salient objects, which can compensate the features from average pooling operation that encode global features information. To consider the importance of the different locations, the spatial attention module is utilized to focus on the different places of the same channel so that the proposed network can express the

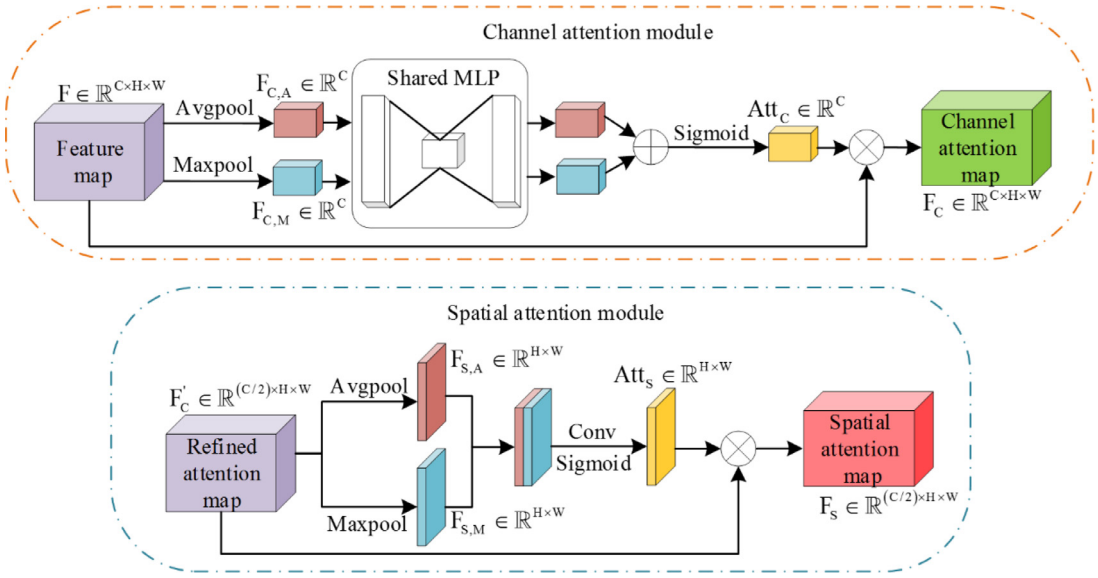


Fig. 5. Structure of the channel and spatial attention module. The feature map from the last feature extraction layer of Path 1 and Path 2 are chosen as the input of channel attention module. After average and max pooling layers, there are two branches to generate channel attention maps, which are fed into the shared Multi-Layer Perception (MLP) and then added, followed by a Sigmoid function to produce the attention weights that are multiplied by the original feature map to get the channel attention map. The gained channel attention maps are processed by a fully connection layer to generate the refined attention map with the half channel of the input channel attention map, which is the input of the spatial attention module. Afterward, we employ a pointwise convolution and sigmoid function to produce the spatial attention weights, which are multiplied by the refined channel attention map to get spatial attention map. $\oplus$ and $\otimes$ represent the element-wise addition and multiplication operations, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

attention maps with channel and spatial information, further enhance the representation ability of the key features.

Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, we leverage the channel and spatial attention modules to highlight the target areas and extract rich contextual information that can distinguish the categories of the different diseases. When the input feature is fed into the channel attention, we adopt the global average pooling and max pooling layers to obtain two features: $F_{C,A} \in \mathbb{R}^C$, $F_{C,M} \in \mathbb{R}^C$. Subsequently, a shared multi-layer perception (MLP) and Sigmoid activation function are used to generate the channel attention map $Att_C \in \mathbb{R}^C$. The channel attention mechanism can be described as follows:

$$Att_C = \sigma\left(M_1 ReLU\left(M_0 F_{C,A}\right) + M_1 ReLU\left(M_0 F_{C,M}\right)\right), \tag{5}$$

where $\sigma(\cdot)$ is the Sigmoid activation function. M_0 and M_1 are the weights of the MLP. After obtaining the attention weights Att_C, we combine it with the input feature map by multiplication operation to get the channel attention map F_C.

$$F_C = F \times Att_c, \tag{6}$$

To restrain the interference of irrelevant areas, we further utilize the spatial attention module to enable the network to extract the key features with the spatial position discrimination that can boost the feature representation for different diseases. We apply the flatten and fully connection layers to the extracted channel attention maps F_C to obtain the refined attention map $F' \in \mathbb{R}^{(C/2)\times H\times W}$, which is the input of the spatial attention module. As shown in Fig. 5, the spatial attention mechanism can be expressed as below:

$$Att_s = \sigma\left(P_{conv}(F_{S,A},\ F_{S,M})\right), \tag{7}$$

where P_{conv} and $\sigma(\cdot)$ are the pointwise convolution and Sigmoid activation function, respectively. Similar to the calculation of the channel attention map, we obtain the spatial attention map by using the spatial attention weights Att_s to multiply the input feature F_C':

$$F_S = F_C^{'} \times Att_s, \tag{8}$$

In order to enhance the contribution of the attention module for the prediction of specific diseases, we adopt cross-connection mode to fuse the attention features from the Path 1 and Path 2 of the proposed multi-branch network. We define the channel and spatial attention map for Path 1 and Path 2 as F_C^{path1}, F_S^{path1}, F_C^{path2}, F_S^{path2}, respectively. Thus, the proposed cross-attention F_{CA} mechanism can be described as follows:

$$F_{CA} = stack\left(F_C^{path1},\ F_S^{path2}\right) + stack\left(F_S^{path1}, F_C^{path2}\right), \tag{9}$$

where the $stack(\cdot)$ operation is the element composition in the view of the last dimension. Thus, we can obtain rich features related to the specific disease and suppress the features being irrelevant to the targets by focusing on the appearance of the specific disease and abandoning the learned redundant information. Compared with the SENet, the advantage is that the cross-attention not only utilize the channel attention information, but also makes full use of the spatial feature information. The difference from channel and spatial attention module is that the cross-attention module fuses the channel attention map and spatial attention map from two different subnetworks via cross-connection mode. Hence, the extracted attention maps are more robust and discriminative by crossing connection mode from the two-branch networks than that of the single path.

4. Experiments

4.1. Implementation details

4.1.1. Database and evaluation metrics

We choose Coats, RP, DR, and Normal SLO images as the categories of classification task, for this reason that the Coats and RP are the common fundus diseases amongst teenagers and DR is a common fundus disease in elderly. Developing an automatic classification system by analysing these diseases can enable the proposed network model to adapt to different age groups in clinical. Our SLO images are collected by Shenzhen Eye Hospital, which are captured via Optos from January 2016 to November 2019. After

Table 1
The data distribution of The Training and Testing sets.

	Coats	RP	DR	Normal
Training	188	379	318	437
Testing	47	94	80	110

screening and labelling from two experts, we get 235 Coats images, 473 RP images, 398 DR images, and 547 Normal images. Thus, we obtain 1653 SLO images in total, which are randomly divided into training and testing data according to the proportion of 4:1, thus, we get 1322 images for training and 331images for testing and the data distribution of training and testing sets are displayed in Table 1. Since the size of the original SLO images is 2600 $\times$ 2048, which is too large for deep neural networks, we resize the original images to 448 $\times$ 448 as the input of the proposed network model after weighing the pixel loss caused by compressing images. In addition, we conduct colour augmentation by randomly changing the brightness, contrast and hue of the training images. Also, the data augmentation is done by randomly horizontal and vertical flipping to enrich the diversity of the training images. We employ Accuracy, Precision, Recall and F1-score to evaluate the classification performance of fundus diseases for SLO images.

4.1.2. Experimental setup

We implement the proposed framework and perform experiments using Pytorch library. Only one NVIDA TITAN XP GPU is used to accelerate the speed in the processing of training and testing. The batch size of the training and testing are set to 16 and 1, respectively. The maximum epoch is set to 120 and the learning rate is initialized to 0.0001 that is multiplied by 0.1 every 30 epochs. We choose the Adam with the betas = (0.9, 0.997) as the optimizer for the stage of training. The whole framework is trained in an end-to-end manner. To meet the clinical requirement, we use the pointwise convolution to extract features as much as possible. Hence, the whole training process takes 1h40min and the testing time for per SLO image is approximately 0.021 s. The cross-entropy function is selected as the classification loss function.

4.2. Experimental results

The classification results of different methods in terms of the evaluation metrics are displayed in Table 2. In Table 2, the ResNet-18, ResNet-34 and ResNet-50 are developed by (He et al., 2016). These models are with residual connection that can leverage the shallow layers feature information, so that the network can preserve the global information while extracting deep features. Therefore, the ResNet models are chosen as the main studying targeted and compared networks. Besides, the Inceptionv3 model (Szegedy et al., 2016) is also selected as the compared network, for the reason that the Inceptionv3 consists of multiple branches that has great ability of extracting features and is similar to our devised network architecture. SE represents the proposed SENet by (Hu et al., 2018), which can be considered as a channel attention model. So, the SENet is also opted as the compared network. The SE-ResNet-18, SE-ResNet-34, SE-ResNet-50, and SE-Inceptionv3 indicate the ResNet-18, ResNet-34, ResNet-50 and Inceptionv3 are used as the backbone of SENet, respectively. From line 1 to line 4 of the Table 2, we can observe that the ResNet-34 model achieves the best results amongst these compared backbone models. From line 5 to line 8 of the Table 2, we can see that the SE-ResNet-34 acquires the best performance amongst these compared backbone models with squeeze-and-excitation mechanism. The results demonstrate that, the deeper the network, the classification performance seems not better for SLO images. For the reason that the deeper network easily occurs over-fitting than shallower network for a small dataset. Hence, we employ the ResNet-34 model as the backbone of the proposed framework. Besides, we devise a multiple stream networks based on ResNet-34 model so that the proposed architecture can not only extract rich deep-level features, but also effectively avoid overfitting. To enable the network extract multi-scale feature information and further improve the classification, the ASPP module is added to the last feature extraction layer of Path 1 and Path 2 of the proposed framework through setting different dilation rates. The depth-wise and cross-attention modules are designed to enforce the network focus on the areas of interest and enhance the feature extraction ability of specific disease, further boost the classification performance. We can also see that our proposed method obtains the best classification performance in all the evaluation metrics and achieves the highest accuracy of 95.47%, which is 3.02% higher than the second-best model (e.g., SE-ResNet-34) amongst the compared methods. These experimental results show that our proposed method is promising in the fundus diseases classification for SLO images. And the training curves are shown in Fig. 6, we can observe that the network tends to converge at the 80th epoch and the overfitting issue does not occur, even if the data is relatively small. For the reason that we use multiple strategies to alleviate this issue, such as data augmentation, adding dropout layer and using pre-trained parameters on ImageNet for Path 3.

We plot the receiver operating characteristic (ROC) curves of the compared nine methods on the test dataset that includes RP, DR, Coats fundus diseases and Normal SLO images, which are shown

Table 2
The classification performances of different methods (%).

Method	Accuracy	Precision	Recall	F1-score
ResNet-18	88.82	88.55	88.24	88.29
ResNet-34	90.33	90.04	89.85	89.93
ResNet-50	85.50	86.27	83.17	84.32
Inceptionv3	87.31	87.67	86.19	86.78
SE-ResNet-18	91.24	91.84	90.21	90.92
SE-ResNet-34	92.45	91.83	92.12	91.91
SE-ResNet-50	92.15	91.29	91.48	91.39
SE-Inceptionv3	91.54	90.53	91.62	90.89
Proposed	**95.47**	**95.41**	**94.22**	**94.75**

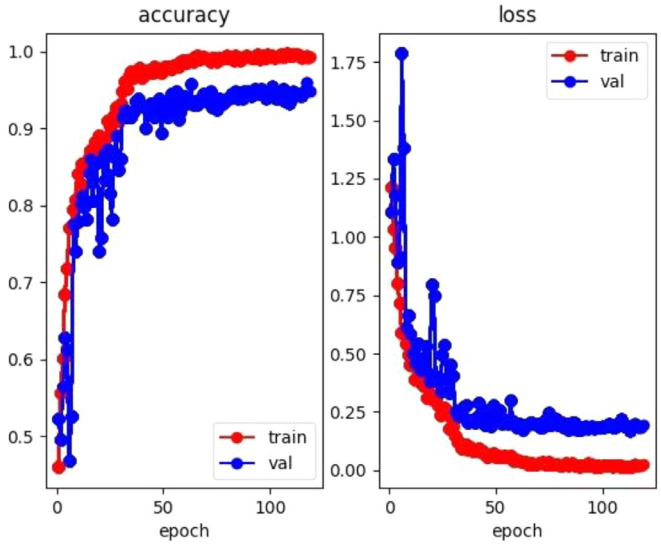


Fig. 6. The training curves. The left and right charts represent the accuracy and loss respectively.

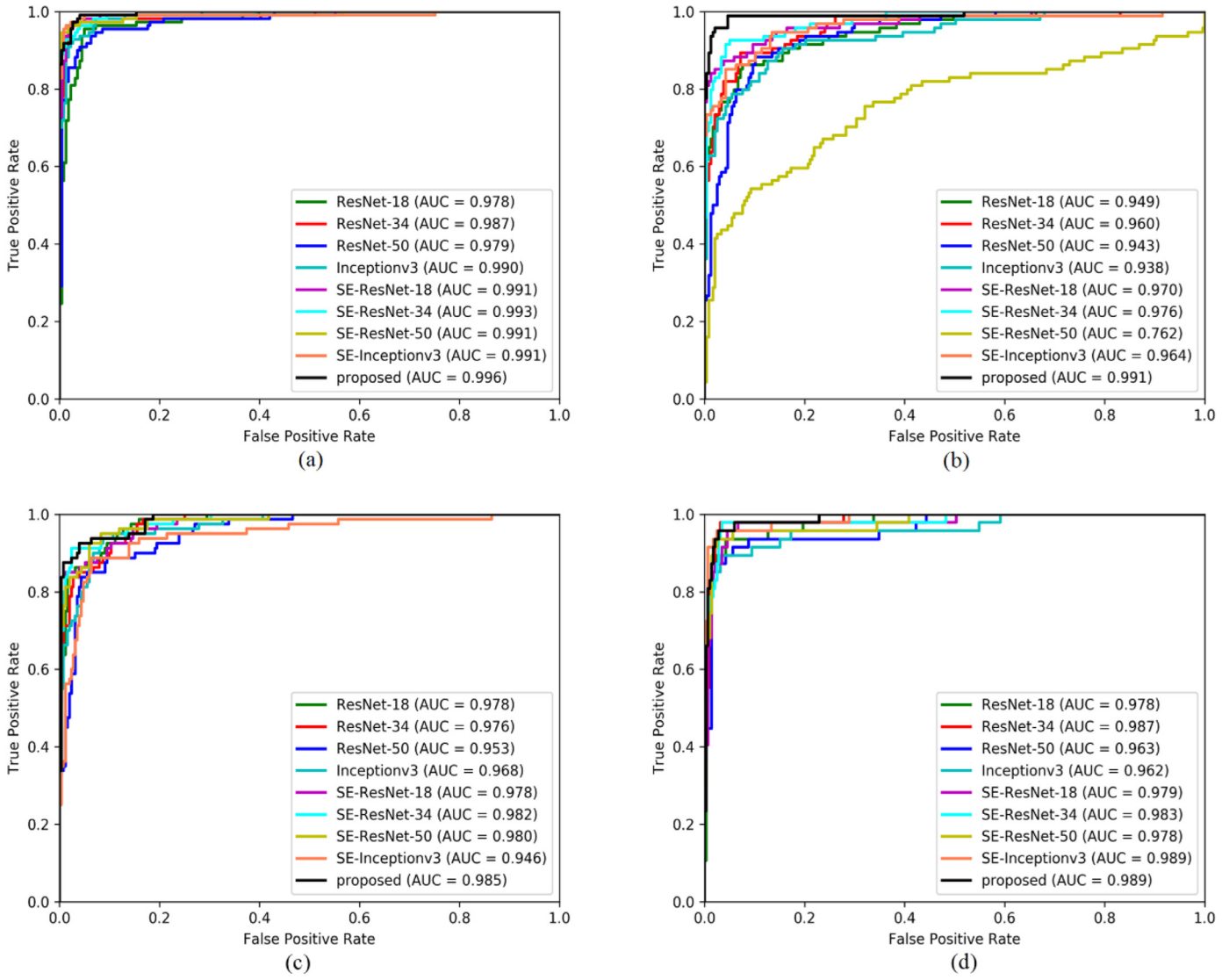


Fig. 7. ROC curves on each disease of the compared models and proposed method. (a)-(d) indicate the results of the RP, DR, Coats fundus disease and Normal SLO images, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Fig. 7. We can see that the black line represents the proposed method and achieves the best classification performance on all the predicted four classes. In addition, we can also observe that our method obviously outperforms the other compared methods in terms of DR fundus disease from Fig. 7(b). The reason is that the characteristics of many SLO images of DR are similar to the other classes, for example, the hemorrhagic spots of some DR images are not obvious so that these images are easily predicted as RP images. However, our devised two attention modules can enforce the network to concentrate on the key target areas of interest, which guide our network to extract the related feature on DR disease and suppress the irrelevant features for DR disease. In addition, we also observe that our method acquires the highest area under curve (AUC) value compared with the other methods in all the predicted classes.

In order to intuitively show the superiority of our proposed method, we also plot the radar chart using evaluation metrics according to the predicted four categories, which is displayed in Fig. 8. We can observe that our method achieves the best performance in terms of accuracy and F1-score, which proves that the devised multi-branch network, depth-wise and cross-attention modules can boost the network to extract discriminative feature for classifying the three diseases and normal SLO images. Thus, our method can potentially assist the doctors to make a decision on the disease that the patients suffer from. From the first two figures in Fig. 8, we can observe that the proposed network gains the best classification performance on the all evaluation metrics, which indicates that our method is quite promising to distinguish the characteristics of the DR and RP diseases.

To illustrate the number of the predicted classes for the comparative methods, we count the number of the predicted category that is correctly predicted as itself and is wrongly predicted as other categories, and plot the qualitative visualization histogram for the result, as shown in Fig. 9. We can observe that the number of the all classes that are predicted correctly by adding the SE module is greater than that of the model without the SE module, which demonstrate that the channel attention mechanism in SE module can enhance the module to focus on the specific nidus areas. In particular, comparing the chart of the proposed and (f), we can observe that based on ResNet-34 model, the devised modules are integrated to the proposed framework can bring great gains for

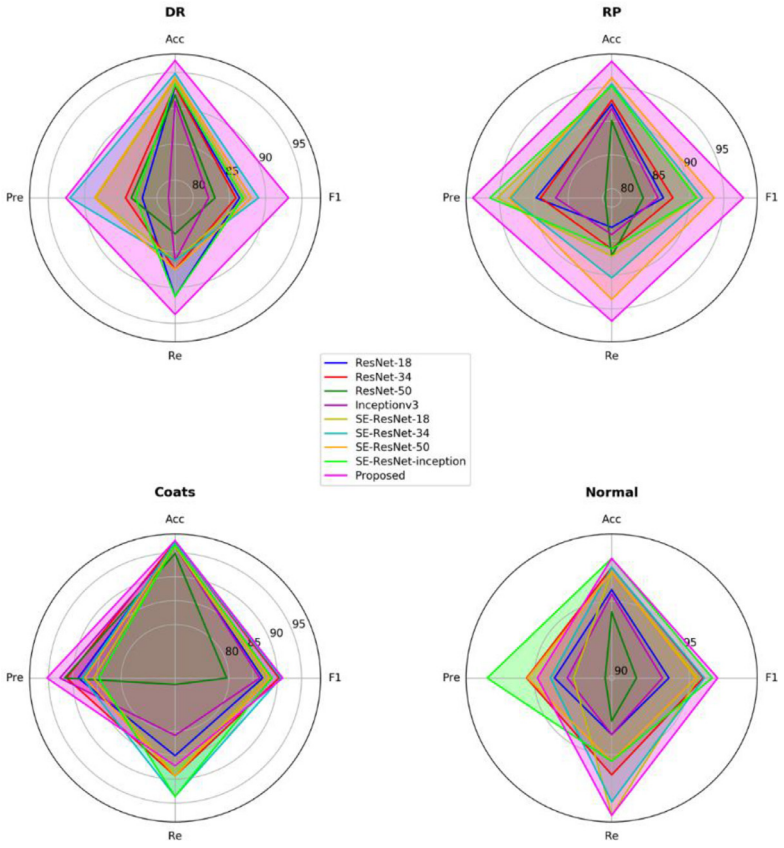


Fig. 8. Radar chart on all evaluation metrics of different methods in terms of the all predicted classes. Acc, Pre, Re, F1 indicates the accuracy, precision, recall and F1-score, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predicting the specific category, especially for the category of Normal, RP and DR.

5. Discussion

5.1. Ablation study

As mentioned above, our proposed network architecture consists of multi-branch network, ASPP, depth-wise and cross-attention modules. Therefore, we conduct ablation experiments with certain modules, which verifies that the devised module is effective to improve the classification performance. In detail, we choose several network structures as the compared networks as follows: the single ResNet-34 model (called Res34), multi-branch network (called MB), multi-branch network with ASPP (called MB-ASPP), multi-branch network with ASPP and depth-wise attention module (called MB-ASPP-DA), multi-branch network with ASPP, depth-wise, and cross-attention modules (called proposed). Note that, our proposed architecture is based on ResNet-34 model. Hence, the ResNet-34 model is also selected as the backbone of the contrasted ablation network. The experimental results of the contrasted ablation network models are displayed in Table 3.

From Table 3, we can see that the results of the multi-branch network based on ResNet-34 model is superior to that of the single ResNet-34 model, which benefits from the case that the multi-branch can extract more deep-level features to express the characteristics of the specific disease. Following the last feature extraction layers of the Path 1 and Path 2, the ASPP module is added to Path 1 and Path 2, which causes that the classification results have been further improved. For the reason that, the ASPP enable the multi-branch network to extract multi-scale features with different dilation rates so that the spatial contextual feature information can be used to enhance the ability of extracting features. To effectively fuse the features extracted by the multi-branch network, the depth-wise attention is devised to keep the global information while enforcing the network to focus on the objects of interest. From the rows 3 and 4 of Table 3, we can see that the accuracy of the network with the depth-wise attention module is 0.91% higher than that of the network without the depth-wise attention module, which indicates that the depth-wise attention module is effective and helpful to boost the classification performance on SLO images.

From rows 4 and 5 of Table 3, we can observe that the classification results of the network with cross-attention module is greatly superior to that of the network without cross-attention module, and the accuracy of the former is 2.72% higher that of the latter. These results demonstrate that the cross-attention module can handle the deep-level features from Path 1 and Path 2, which enable the network to focus on the characteristics of specific disease and repress the irrelevant features in a cross-fusion manner, further improve the classification performance on the three diseases and normal SLO images.

Table 3
The classification results of the contrasted ablation network models on the testing dataset (%).

Method	Accuracy	Precision	Recall	F1-score
Res34	90.33	90.04	89.85	89.93
MB	91.24	91.09	90.82	90.95
MB-ASPP	91.84	91.65	91.14	91.37
MB-ASPP-DA	92.75	92.37	92.47	92.40
Proposed	**95.47**	**95.41**	**94.22**	**94.75**

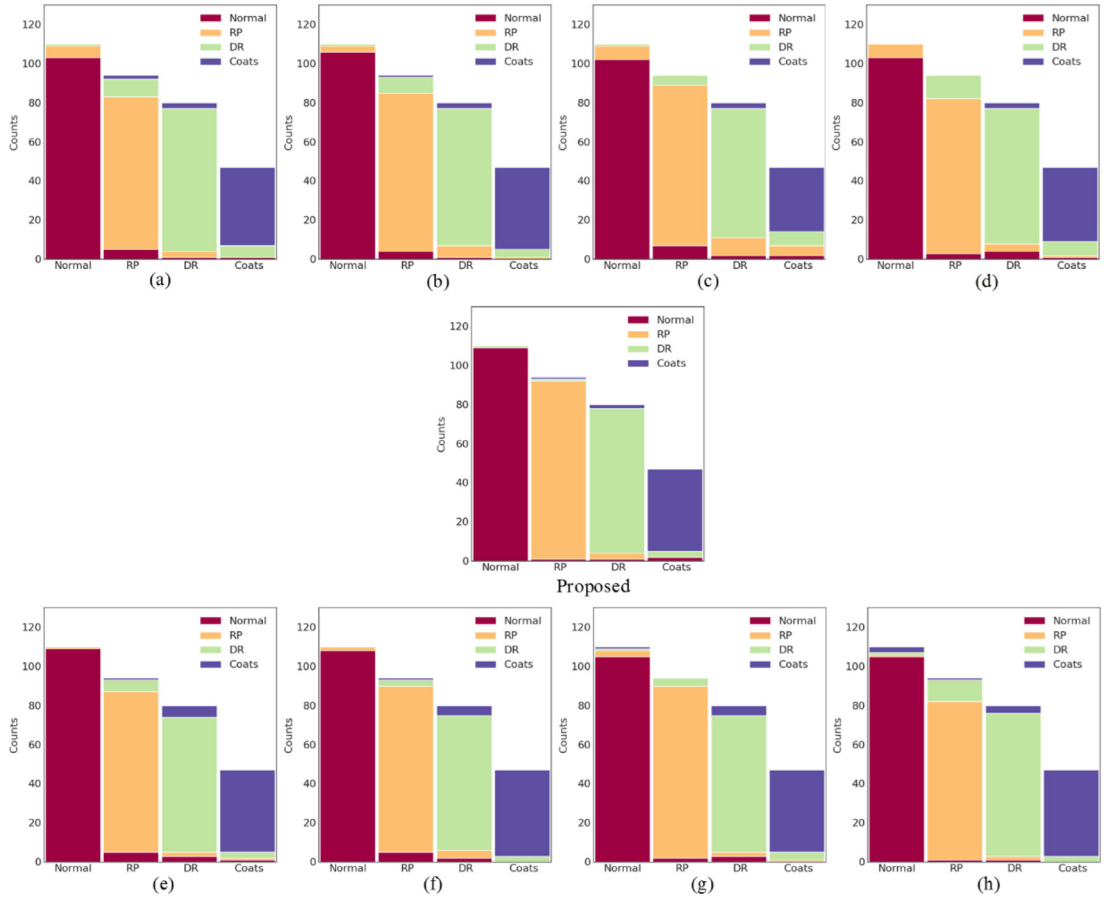


Fig. 9. The qualitative visualization of the number of the predicted classes for different methods. (a)-(d) represent the predicted results of ResNet-18, ResNet-34, ResNet-50 and Inceptionv3 respectively. (e)-(h) indicate the predicted results of SE-ResNet-18, SE-ResNet-34, SE-ResNet-50 and SE-Inceptionv3 respectively. The middle chart is the predicted results by our proposed model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We plot the confusion matrices of the contrasted ablation networks to indicate the devised modules can improve the classification intuitively, which is shown in Fig. 10. We can observe that our designed modules can greatly improve the predicted performance in terms of RP and DR fundus diseases. In addition, we also implement the t-SNE visualization (Maaten and Hinton, 2008) of the compared networks on the testing dataset, where the features learned by the last feature extraction layer of the compared networks are visualized, which is displayed in Fig. 11. We can see that, every time a module is added to the proposed framework, the separation effect will be better, which is consistent with Table 3.

To illustrate the good ability of extracting features for the devised modules, we also conduct the experiments with the class activation maps (CAMs) by using the output features of the designed modules, which can verify that the extracted features are effective to distinguish the predicted class. The Grad-CAMs visualization is shown in Fig. 12, where the white close boundaries in the column one represent the approximate area of the lesions. As observed in the row one, all models focus on the location of optic cup, for the reason that the first row represents the normal SLO image, where there is no lesion so that all models can only learn the salient features from optic cup in contrast to the whole images. For the diseased SLO images, each module is integrated to the devised framework, which can enforce the framework to focus on the specific disease class. In particular, the proposed model shows strong superiority with respect to extracting the disease-related feature for the category of Coats. Hence, the proposed modules are effective and can well learn the features of the salient areas and focus on the features of the specific disease so that the target class can be predicted accurately.

Table 4
The comparison results of different attention strategies (%).

Method	Accuracy	Precision	Recall	F1-score
CAM	93.96	93.01	93.37	93.16
SAM	93.35	92.15	93.35	92.63
CSAM	94.56	94.00	93.49	93.73
Proposed	**95.47**	**95.41**	**94.22**	**94.75**

To show the superiority of the cross-attention module, we conduct the comparison experiments in terms of only using channel attention module (CAM) or spatial attention module (SAM), and utilizing channel attention before spatial attention module without cross fusion (CSAM). The experimental results are reported in Table 4. We can see that the channel and spatial attention modules being fused via cross-connection mode achieves the best classification performance, which indicates the cross fusion mode for attention mechanism is effective.

To further show the effectiveness of the designed cross-attention module, we conduct the Grad-CAM visualization of the learned features in terms of different attention strategies, and the results are shown in Fig. 13. We can observe that the redundant feature information can be obtained by the SAM, CAM and CSAM and the cross-attention can better recognize the regions that are related to the specific disease.

We also consider the influence on the results from the pooling mode in the attention module. To verify the effectiveness of the

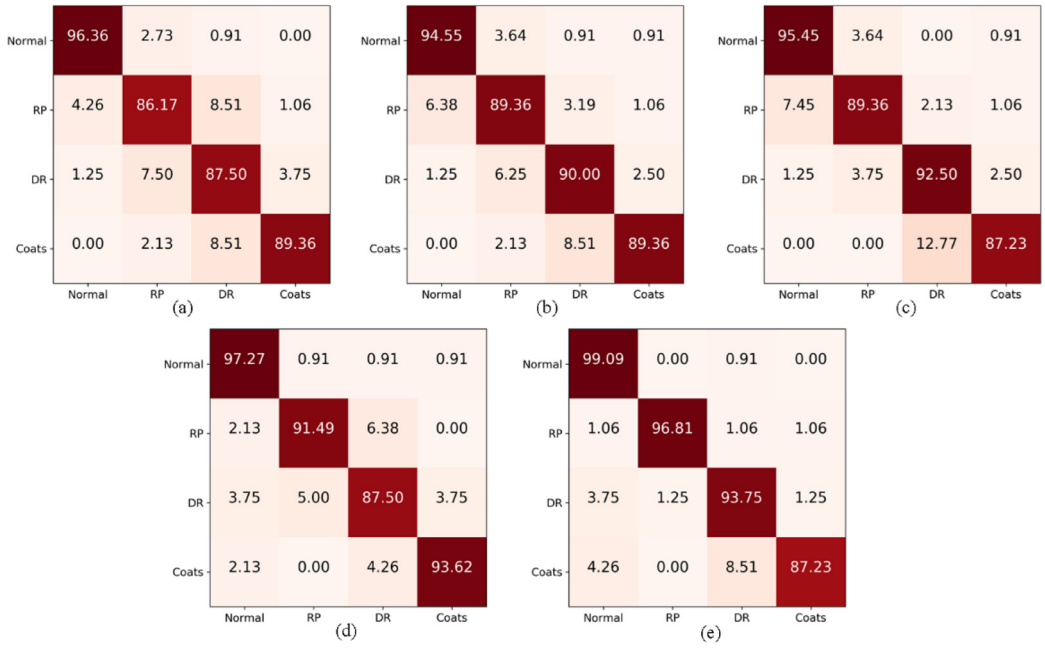


Fig. 10. Confusion matrices of the compared network models on the testing dataset. The numbers in the confusion matrices denote the percentage of the predicted class. (a) – (e) represent the results of Res34, MB, MB-ASPP, MB-ASPP-DA and proposed methods, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

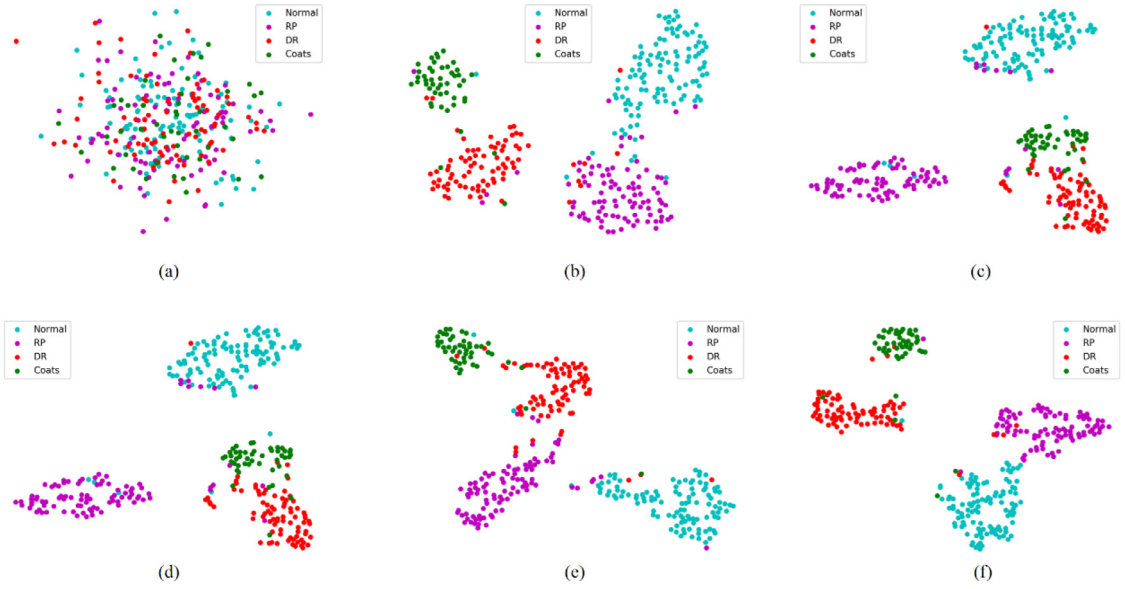


Fig. 11. The t-SNE visualization of the compared networks on the testing dataset. (a) – (f) represent the distribution of the original testing dataset, Res34, MB, MB-ASPP, MB-ASPP-DA and proposed methods on the testing dataset, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
The comparative results of different pooling strategies (%).

Pooling strategies	Accuracy	Precision	Recall	F1-score
Max-pooling	94.86	94.59	93.50	93.98
Average pooling	95.17	94.65	94.16	94.35
Proposed	**95.47**	**95.41**	**94.22**	**94.75**

joint use of max-pooling and average pooling, we conduct a series of ablation experiments, and the results are reported in Table 5. From Table 5, we can see that the proposed model achieves the best performance with both max-pooling and average pooling being used. For the reason that the both two pooling strategies can provide more feature information for the subsequent attention operation than single pooling operation.

5.2. Validation

To demonstrate that the proposed method is effective for classifying the fundus diseases using SLO images, we conduct a group of comparative experiments on a public dataset that is collected by Tsukazaki Hospital [50]. This dataset contains 150 SLO images of RP and 223 SLO images of Normal fundus, which is divided by a ratio of 4:1 into training and testing dataset. The experimental setting is the same as the Section 4.1. The experimental results are

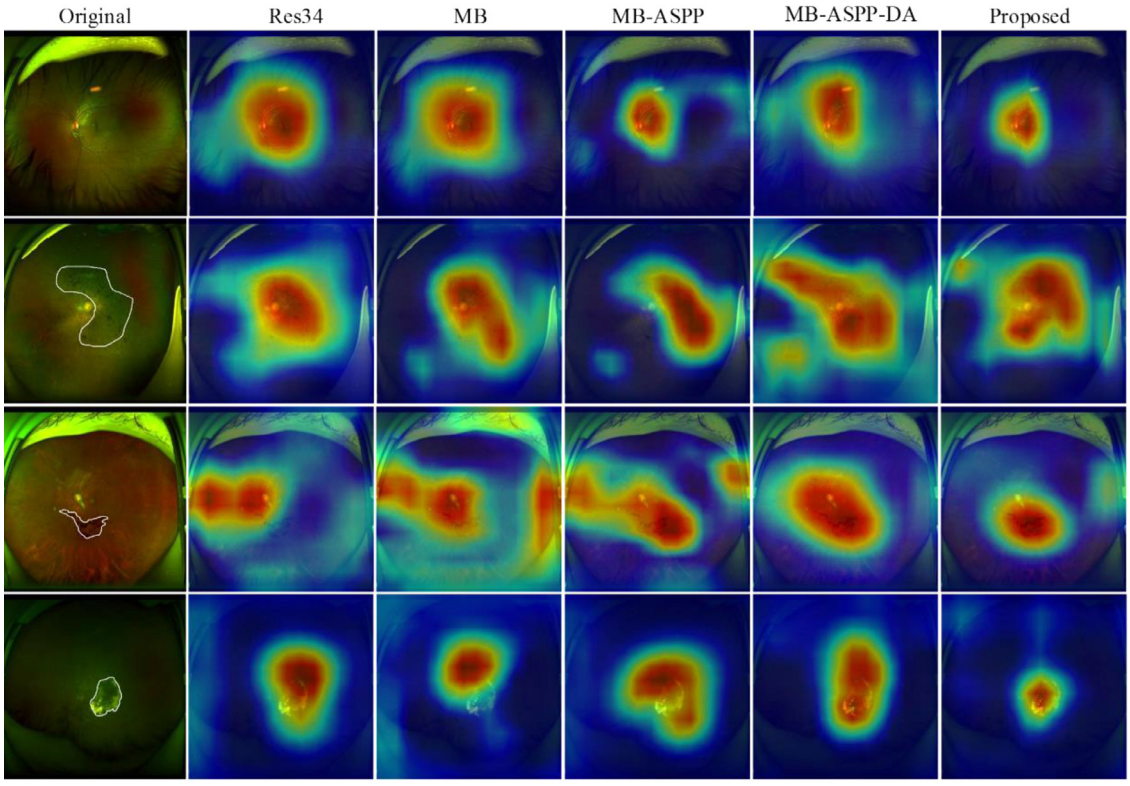


Fig. 12. The visualization of Grad-CAM of the different devised modules. Row 1 to row 4 indicate the category of Normal, RP, DR and Coats, respectively.

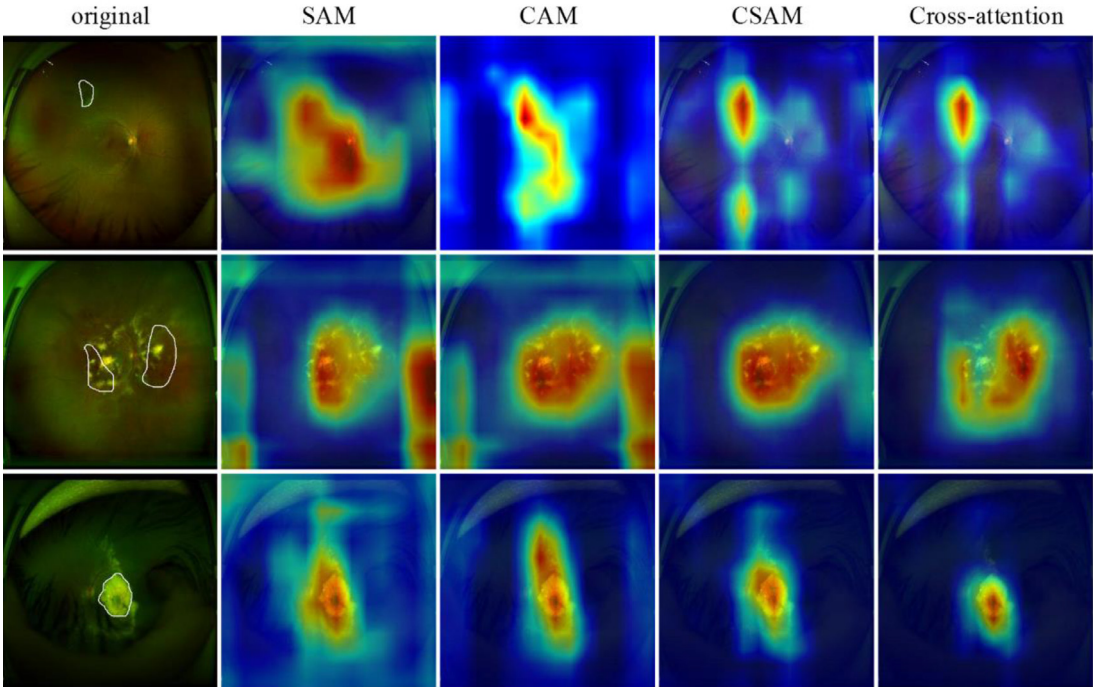


Fig. 13. The Grad-CAM visualization of the learned features with respect to different attention strategies. Row 1 to row 3 indicate the disease category of RP, DR and Coats.

reported in Table 6, where the method abbreviations stand for the same meaning as those in Table 3.

From Table 6, we can see that the proposed method achieves the best performance by integrating these mentioned modules in Section 3. Each module can improve the performance on detecting RP using SLO images. In addition, we plot the ROC curves of the contrasted ablation networks to illustrate the effectiveness of the proposed modules and framework, which is shown in Fig. 14. We can observe that our proposed method obtains the maximum AUC value for the RP and normal SLO images, respectively.

To show the generalization ability of the proposed model on other modality of fundus images, such as colour fundus images.

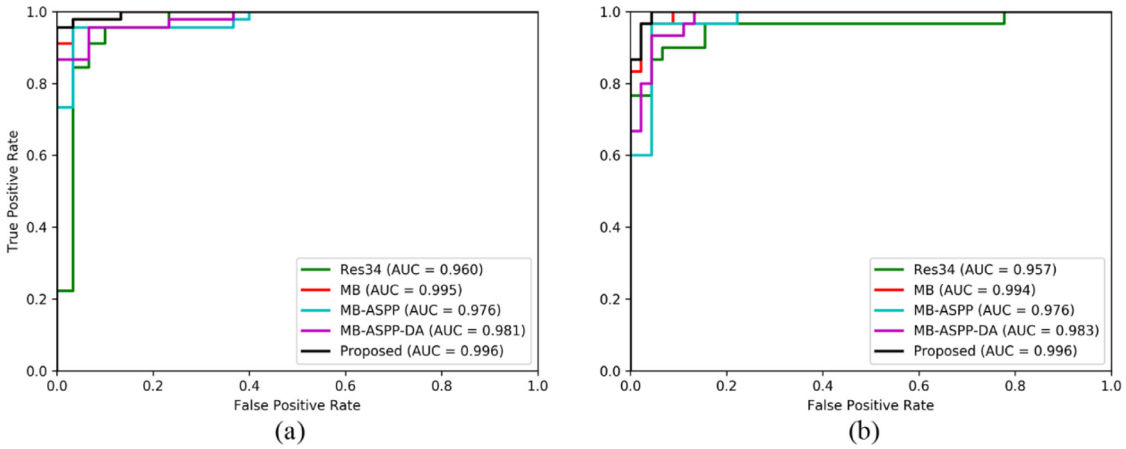


Fig. 14. ROC curves on the public dataset from the Tsukazaki Hospital of the contrasted ablation networks. (a) and (b) indicate the results on RP and Normal SLO images, respectively.

Table 6
The comparison results of the used different modules on the public dataset (%).

Method	Accuracy	Precision	Recall	F1-score
Res34	92.00	91.67	91.67	91.67
MB	93.33	92.75	93.89	93.16
MB-ASPP	94.67	94.44	94.44	94.44
MB-ASPP-DA	96.00	95.64	96.11	95.86
Proposed	**97.33**	**97.22**	**97.22**	**97.22**

Table 7
The comparison results with the reported results on the IDRID2018 dataset.

Method	Joint accuracy (%)	Rank
Our model	63.11	1
LzyUNCC	63.11	1
VRT	55.34	2
Mammoth	51.46	3
HarangiM1	47.57	4
AVSASVA	47.57	4
HarangiM2	40.78	5

We conduct the experiments on the Indian Diabetic Retinopathy Image Dataset (IDRID2018). We mainly do experiments for the subchallenge-2, in which the purpose of this task is to evaluate the performance of the designed networks for jointly recognizing the grading of diabetic retinopathy and diabetic macular oedema. The dataset for the subchallenge-2 consists of 413 training images and 103 testing images. We use the same experimental settings to train and test this dataset except that the size of the training images is set to 512 $\times$ 512. The comparison results with the competitors are shown in Table 7.

The joint recognition accuracy of the grading of diabetic retinopathy and diabetic macular oedema is 63.11%, which is the same as the result of rank one that is reported on the official website of the competition (https://idrid.grand-challenge.org/Leaderboard/). As reported in (Porwal et al., 2020), we can see that the methods of all competitors used external dataset to train their models, further enhanced the final classification performance for subchallenge-2. However, our method only uses the provided data on the site and the parameters of out proposed model to train the network and obtain the same ranking as rank one, which also demonstrates that our proposed architecture is with good generality, even for different modal data.

5.3. Limitation and future work

Although the proposed framework achieves the best performance for classifying the three fundus diseases and normal SLO images in contrast to the compared network models, there exist some limitations in this paper. These limitations are summarized as follows. Firstly, the number of the SLO dataset is insufficient for deep neural network to learn deep-level and discriminative features. To keep the dataset pure, we just conduct colour enhancement in the aspects of brightness, contrast and hue, which lacks of multiple transformations so that the network only learns limited features. Secondly, the studied types of fundus disease species only have RP, DR and Coats due to the difficulty of collecting data, which causes the issue that the proposed algorithm cannot be widely applied in clinical.

To address the mentioned limitations above, we need to strengthen the collecting and processing intensity of the SLO images in future work. For instance, we can extend the dataset to address the problem of small dataset through various manners of enhancing data, such as cropping, rotation, and flipping. We can also collect SLO images from the multiple centers to address the problem of the less data, meanwhile, the types of the studied disease species can be increased with the data of multiple centers. Besides, exploring the data of multiple centers also enables the developed model to robustly adapt to the data from different conditions, such as different machine models and medical staffs.

In addition, the research of the subtypes of fundus diseases is a quite significant work for the diagnosis in clinical. For the reason that the subtypes of fundus diseases can reflect the changed condition of the disease the patients suffer from, which can assist the doctors to control the deterioration of the disease in time by scientific treatment.

6. Conclusions

In this paper, we propose a novel cross-attention multi-branch network based on ResNet-34 model to accomplish the classification task of the selected three diseases and normal images using SLO fundus images. Specifically, the proposed framework mainly includes four parts: multi-branch network, ASPP, depth-wise attention and cross-attention modules. The multi-branch employs the ResNet-34 model as the backbone to expand the network branch to extract more deep-level features. Following the last feature extraction layer of the multi-branch network, the ASPP is utilized to extract the multi-scale features via setting different dilation rates to gain the rich spatial contextual feature information. The depth-

wise attention module is devised to fuse the features from the ASPP module and another branch network with original ResNet-34 model so that the whole framework can preserve the global information while focusing on the target areas of interest. To make full use of the extracted features from the multi-branch network and enforce the network to highlight the characteristics of the specific disease and suppress the irrelevant areas, we design a cross-attention module to fuse the channel attention maps and spatial attention maps from the Path 1 and Path 2 through the crisscross connection manner. The extensive experimental results demonstrate that our proposed framework achieves quite promising classification performance on the three fundus diseases and normal images. Through the ablation studies, we can draw the view that, the developed modules can effectively boost the classification performance.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Hai Xie: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Xianlu Zeng:** Validation, Investigation, Resources. **Haijun Lei:** Conceptualization, Supervision, Project administration. **Jie Du:** Conceptualization, Writing – review & editing. **Jiantao Wang:** Validation, Resources, Data curation. **Guoming Zhang:** Conceptualization, Validation, Resources, Data curation, Project administration. **Jiuwen Cao:** Conceptualization, Writing – review & editing. **Tianfu Wang:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Baiying Lei:** Conceptualization, Methodology, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China (Nos. U1909209, 61871274, 61801305 and 81571758), Guangdong Pearl River Talents Plan (2016ZT06S220), National Natural Science Foundation of Guangdong Province (No. 2017A030313377), Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515110605), Shenzhen Key Basic Research Project (Nos. JCYJ20180507184647636, GJHZ 20180418190529516, GJHZ20190822095414576, JCYJ20190808145011259 and JSGG 20180507183215520), and Shenzhen Peacock Plan (Nos. KQTD2016053112051497 and KQTD20150330161049 26).

References

Aaberg, T.M., Shevchenko, L., 2019. Coats' Disease. Clinical Ophthalmic Oncology 5–17 Springer.

Ashok, V., Murugesan, G., 2017. Detection of retinal area from scanning laser ophthalmoscope images (SLO) using deep neural network. Int, J. Biomed. Eng. Tech. 23 (2–4), 303–314.

Badar, M., Haris, M., Fatima, A., 2020. Application of deep learning for retinal image analysis: a review. Comput. Sci. Review 35, 100203.

Chen, H., He, Y., Wei, L., Li, X., Zhang, Y., 2019. Automatic Dewarping of Retina Images in Adaptive Optics Confocal Scanning Laser Ophthalmoscope. IEEE Access 7, 59585–59599.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K., & Yuille, A.L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern. Anal. Mach. Intell. 40 (4), 834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. IEEE ECCV, pp. 801–818.

Cohen, S.R., Gardner, T.W., 2016. Diabetic retinopathy and diabetic macular edema. Retinal Pharmacotherapeutics 55, 137–146 Karger Publishers.

Coyner, A.S., Swan, R., Campbell, J.P., Ostmo, S., Brown, J.M., Kalpathy-Cramer, J., Kim, S.J., Jonas, K.E., Chan, R.P., Chiang, M.F., 2019. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. Ophthalmo. Retina 3 (5), 444–450.

Cunefare, D., Langlo, C.S., Patterson, E.J., Blau, S., Dubra, A., Carroll, J., Farsiu, S., 2018. Deep learning based detection of cone photoreceptors with multimodal adaptive optics scanning light ophthalmoscope images of achromatopsia. Biomed. Opt. Exp. 9 (8), 3740–3756.

Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018a. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imag. 37 (7), 1597–1605.

Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D.W.K., Liu, J., Cao, X., 2018b. Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans. Med. Imag. 37 (11), 2493–2501.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y., 2020. Thorax disease classification with attention guided convolutional neural network. Pattern Recognit. Lett. 131, 38–45.

Haleem, M.S., Han, L., van Hemert, J., Li, B., Fleming, A., 2014. Retinal area detector from Scanning Laser Ophthalmoscope (SLO) images for diagnosing retinal diseases. IEEE J. Biomed. Health Inf. 19 (4), 1472–1482.

Hamel, C., 2006. Retinitis pigmentosa. Orphanet.J. Rare. Dis. 1 (1), 40.

Hamwood, J., Alonso-Caneiro, D., Sampson, D.M., Collins, M.J., Chen, F.K., 2019. Automatic Detection of Cone Photoreceptors With Fully Convolutional Networks. Trans. Vis. Sci. Tech. 8 (6) 10-10.

He, A., Li, T., Li, N., Wang, K., Fu, H., 2020. CABNet: category Attention Block for Imbalanced Diabetic Retinopathy Grading. IEEE Trans. Med. Imag. 40 (1), 143–153.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE CVPR, pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proc. IEEE CVPR, pp. 7132–7141.

Ke, L., Chang, M.-C., Qi, H., Lyu, S., 2018. Multi-scale structure-aware network for human pose estimation. In: Proc. IEEE ECCV, pp. 713–728.

Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W., 2019. Dual Student: breaking the Limits of the Teacher in Semi-supervised Learning. In: Proc. IEEE ICCV, pp. 6728–6736.

Lai, R., Li, Y., Guan, J., Xiong, A., 2019. Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. IEEE Access 7, 114385–114399.

Li, H., He, X., Zhou, F., Yu, Z., Ni, D., Chen, S., Wang, T., Lei, B., 2018. Dense deconvolutional network for skin lesion segmentation. IEEE J. Biomed. Health Inf. 23 (2), 527–537.

Li, X., Hu, X., Yu, L., Zhu, L., Fu, C.-W., Heng, P.-A., 2019. CANet: cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading. IEEE Trans. Med. Imag. 39 (5), 1483–1493.

Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., Lei, B., 2019. Multi-task learning for quality assessment of fetal head ultrasound images. Med. Image Anal. 58, 101548.

Liu, Y.; Wang, Y.; Wang, S.; Liang, T.; Zhao, Q.; Tang, Z. and Ling, H. (2019). CBNet: a Novel Composite Backbone Network Architecture for Object Detection. arXiv preprint arXiv:1909.03625.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Matsuba, S., Tabuchi, H., Ohsugi, H., Enno, H., Ishitobi, N., Masumoto, H., Kiuchi, Y., 2019. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. Int. Ophthalmo 39 (6), 1269–1275.

Meyer, M.I., Costa, P., Galdran, A., Mendonça, A.M., Campilho, A., 2017. A deep neural network for vessel segmentation of scanning laser ophthalmoscopy images. In: Int. Conf. Image Anal. Recongni, pp. 507–515.

Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., Qi, H., Zheng, Y., Frangi, A., Liu, J., 2019. CS-Net: channel and Spatial Attention Network for Curvilinear Structure Segmentation. In: Proc. Int. Conf. MICCAI, pp. 721–730.

Nagasawa, T., Tabuchi, H., Masumoto, H., Enno, H., Niki, M., Ohara, Z., Yoshizumi, Y., Ohsugi, H., Mitamura, Y., 2019. Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatment-naïve proliferative diabetic retinopathy. Int. Ophthalmo. 39 (10), 2153–2159.

Nie, D., Wang, L., Xiang, L., Zhou, S., Adeli, E., Shen, D., 2019. Difficulty-Aware Attention Network with Confidence Learning for Medical Image Segmentation. In: Proc. AAAI, 33, pp. 1085–1092.

Pellegrini, E., Robertson, G., MacGillivray, T., van Hemert, J., Houston, G., Trucco, E., 2017. A graph cut approach to artery/vein classification in ultra-widefield scanning laser ophthalmoscopy. IEEE Trans. Med. Imag. 37 (2), 516–526.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., 2020. Idrid: diabetic retinopathy–segmentation and grading challenge. Med. Image Anal. 59, 101561.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. MICCAI, pp. 234–241.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.

Shaikh, M., Kollerathu, V.A., Krishnamurthi, G., 2019. Recurrent Attention Mechanism Networks for Enhanced Classification of Biomedical Images. In: Proc. IEEE ISBI, pp. 1260–1264.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sun, L., Fan, Z., Fu, X., Huang, Y., Ding, X., Paisley, J., 2019. A deep information sharing network for multi-contrast compressed sensing MRI reconstruction. IEEE Trans. Image Process. 28 (12), 6141–6153.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.
Webb, R.H., Hughes, G.W., 1981. Scanning laser ophthalmoscope. IEEE Trans. Biomed. Eng. 7, 488–492.
Wei, J., Xia, Y., Zhang, Y., 2019. M3Net: a multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. Pattern Recognit 91, 366–378.
Woo, S., Park, J., Lee, J.-Y., So Kweon, I., 2018. Cbam: convolutional block attention module. In: Proc. IEEE ECCV, pp. 3–19.
Wu, Z., Su, L., Huang, Q., 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In: Proc. IEEE CVPR, pp. 3907–3916.
Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes. In: Proc. IEEE CVPR, pp. 3684–3692.
Yang, Z., Ran, L., Zhang, S., Xia, Y., Zhang, Y., 2019. EMS-Net: ensemble of Multiscale Convolutional Neural Networks for Classification of Breast Cancer Histology Images. Neurocomputing 366, 46–53.
Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L., 2018. Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. In: Proc. IEEE CVPR, pp. 8280–8289.
Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019a. Attention residual learning for skin lesion classification. IEEE Trans. Med. Imag. 38 (9), 2092–2103.
Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., Tan, M., Xu, Y., 2019b. Attention guided network for retinal image segmentation. In: Proc. Int. Conf. MICCAI, pp. 797–805.
Zhao, J., Lei, B., Wu, Z., Zhang, Y., Li, Y., Wang, L., Tian, R., Chen, Y., Ma, D., Wang, J., 2019. A deep learning framework for identifying zone I in RetCam images. IEEE Access 7, 103530–103537.