PLOS ONE

RESEARCH ARTICLE

# MHANet: A hybrid attention mechanism for retinal diseases classification

**Lianghui Xu, Liejun Wang*, Shuli Cheng, Yongming Li**

College of Information Science and Engineering, Xinjiang University, Urumqi, China

* wljxju@xju.edu.cn





**Citation:** Xu L, Wang L, Cheng S, Li Y (2021) MHANet: A hybrid attention mechanism for retinal diseases classification. PLoS ONE 16(12): e0261285. https://doi.org/10.1371/journal.pone.0261285

**Editor:** Jie Zhang, Newcastle University, UNITED KINGDOM

**Received:** September 6, 2021

**Accepted:** November 26, 2021

**Published:** December 16, 2021



**Data Availability Statement:** Considering the influence of the number of datasets on the experiment, two public datasets are used in this paper. Among them, Dataset1 was provided by Rasti et al., obtained at the Noor Eye Hospital in Tehran using Spectralis SD-OCT imaging [13]. Dataset2 was provided by Kermany et al, and obtained by using Spectralis SD-OCT imaging [32]. Others can obtain the corresponding data through the following websites: Dataset1: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8166817 Dataset2: https://www.kaggle.com/paultimothymooney/kermany2018.

## Abstract

With the increase of patients with retinopathy, retinopathy recognition has become a research hotspot. In this article, we describe the etiology and symptoms of three kinds of retinal diseases, including drusen(DRUSEN), choroidal neovascularization(CNV) and diabetic macular edema(DME). In addition, we also propose a hybrid attention mechanism to classify and recognize different types of retinopathy images. In particular, the hybrid attention mechanism proposed in this paper includes parallel spatial attention mechanism and channel attention mechanism. It can extract the key features in the channel dimension and spatial dimension of retinopathy images, and reduce the negative impact of background information on classification results. The experimental results show that the hybrid attention mechanism proposed in this paper can better assist the network to focus on extracting thr fetures of the retinopathy area and enhance the adaptability to the differences of different data sets. Finally, the hybrid attention mechanism achieved 96.5% and 99.76% classification accuracy on two public OCT data sets of retinopathy, respectively.

## Introduction

Diabetic retinopathy (DR) is a common blinding disease, mostly in diabetic patients [1]. Changes in blood components of diabetic patients cause dysfunction of vascular endothelial cells, which leads to impaired retinal barrier in diabetic patients. According to a study by the International Diabetes Federation (IDF), patients with diabetes over 10 years have a 60.00% chance of developing retinopathy. Periodic examinations of diabetic patients and effective control of retinal diseases can prevent patients from transient or permanent blindness [2]. Diabetic macular edema (DME) is one of the main symptoms of diabetic retinopathy (DR) [3]. The main reason for this disease is the long-term hyperglycemia state of diabetic patients, which leads to increased vascular permeability of the retina and choroid of the eye, and the water molecules and some protein components in the blood can permeate through the damaged vascular walls and form edema in the macular area.

Another common retinopathy is age-related macular degeneration (AMD) [4, 5]. At present, the exact cause of the disease is unknown. It may be related to genetic factors, environmental effects, chronic retinal light injury, nutritional disorders, metabolic disorders and so on. However, aging and degeneration are important factors that cause age-related macular

**Funding:** This research was funded in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region grant number 2020D01C034, Tianshan Innovation Team of Xinjiang Uygur Autonomous Region grant number 2020D14044, the National Science Foundation of China under Grant U1903213, 61771416 and 62041110, the Creative Research Groups of Higher Education of Xinjiang Uygur Autonomous Region under Grant XJEDU2017T002.

**Competing interests:** No conflict of interest exits in the submission of this manuscript.

degeneration. According to clinical manifestations and pathological changes, drusen(DRUSEN) and choroidal neovascularization (CNV) are one of the main symptoms of AMD.

Among them, the early symptoms of CNV are subretinal neovascularization, but the symptoms are not obvious. In the middle stage, the blood vessels will gradually expand to a certain extent, causing leakage, rupture and bleeding, which can lead to vision loss. In the later stage, with the aggravation of vascular infiltration and bleeding, the patient's vision was severely damaged, which caused the patient's permanent visual impairment.

DRUSEN is often caused by metabolic disorders of the pigment epithelium. In the early stage, the patient's visual function may not be impaired. However, as the condition worsens, some patients have enlarged physiological blind spots in their eyes. If the situation is more serious, it may cause the patient's field of view to narrow or vision loss. AMD has become the most important blinding disease in western developed countries [4, 5]. As the macular area is the most important part of the retina and the most sensitive part of vision, 80.00% of human vision comes from the macular area. Once a lesion such as edema occurs in the macular area, the impact on vision is very serious. Therefore, regular checking for diabetic patients and effective control of the disease in early stage can prevent temporary or permanent blindness in [2].

With the rapid increase in the number of patients with retinopathy, hospitals need a lot of manpower and material resources to face complex situations. In order to reduce the burden on hospitals and doctors, the auxiliary diagnosis and analysis technology of retinopathy is an urgent technology for doctors. OCT is a new tomography technology developed in recent years [6–9]. Compared with the traditional fundus detection technology, OCT can perform non-contact and non-invasive tomography on the microstructure of living eye tissue [10]. Therefore, this paper uses the OCT retinal fundus detection image datasets for our research.

With the wide application of neural networks in many fields [11], medical image classification based on deep learning has also become a research hotspot. Many scholars have tried to use neural networks to identify retinopathy, and achieved good results. In order to better lock foreground information of the picture, image denoisingis often used in image processing [12, 13]. In addition, the attention mechanism can also lock foreground information of the image very well, so this article explores the role of the attention mechanism in the recognition of retinal diseases. Attention mechanism was originally used in machine translation [14], and now it has become an important concept in image recognition. When recognizing an image, it is similar to the human visual system. It can assist the network to pay more attention to the key information in the image. In order to better identify retinal diseases, this paper proposes a new attention mechanism, which can better improve the accuracy of retinal disease recognition.

At present, the attention mechanism in image recognition is mainly divided into channel attention mechanism and spatial attention mechanism. The channel attention mechanism is used to distinguish the relationship between the feature map channels and the spatial attention mechanism is used to distinguish the relationship between the elements in the feature map. Because the lesion area in OCT images of retinopathy has the following characteristics: (1) The lesion features are concentrated in the local area. (2) The characteristics of the lesions are not much different from those of normal tissues. Therefore, this paper proposes a multi-branch hybrid attention network (MHANet) to lock the lesion area of the picture. Specifically, the channel attention mechanism is mainly used to highlight the channel elements with the most abundant lesion features; Spatial attention mechanism is mainly used to highlight the elements of the area where the lesion features are located. Experimental results show that this algorithm has certain advantages compared with previous mainstream algorithms. In addition, the visualization results show that compared with the mainstream model using channel attention mechanism alone, the hybrid attention mechanism proposed in this paper can more accurately lock the location of the lesion area in retinal OCT images.

## Related work

With the development of machine learning and deep learning, medical image analysis based on deep learning has become a research hotspot for many scholars.

In 2014, Srinivasan et al. [15] applied machine learning methods to image analysis of retinal diseases. They used the support vector machine (SVM) as the classification tool and the HOG feature of the OCT images as the classification basis to classify and recognize the three types of retinopathy images. They constructed a multi class classifier using three linear SVM, and then classified NORMAL and AMD, NORMAL and DME, AMD and DME. Finally, AMD, DME and NORMAL achieved 100.00%, 100.00% and 86.67% classification accuracy respectively. This experimental method not only increases the complexity of the experiment but also does not fundamentally distinguish the three types of pictures at the same time. In addition, in order to promote research of ophthalmic diseases, they also created a public OCT dataset for the majority of scholars to use.

In 2016, Wang et al. [16] used same data set as Srinivasan et al. [15] to further deepen research on retinopathy. They first used two feature selection algorithms CFS (Correlation-based Feature Subset) and CSE (Classifier Subset Evaluator) to find a subset of Linear Configuration Pattern (LCP) features [17, 18], and then used the Sequence Minimization Optimization (SMO) algorithm as a classification tool to recognize and classify the selected image features. Unlike Srinivasan et al., they combine the local and global features of the picture. The experimental results show that Wang et al achieved 97.80%, 94.00% and 99.60% accuracy on AMD, DME and NORMAL respectively. Although the feature selection algorithm can reduce the redundant or useless features in the picture, some key features will be lost in the process of feature selection.

In 2017, Rasti et al. [13] used a multi-scale neural network to analyze retinopathy images. They used parallel convolution kernels of different sizes to extract the features of different regions of the feature map, and then set up a gated network to assign corresponding coefficients to the feature maps extracted by different convolution kernels to adjust the relationship between different feature maps. In addition, they also created a new OCT retinal disease dataset. For the dataset created by Srinivasan et al., the overall classification accuracy of AMD, DME and NORMAL is 99.39%. For their own datasets, the accuracy rates of AMD, DME, and NORMAL are 95.67%, 98.22% and 96.67%, respectively. Although the gated network can reconcile the relationship between the feature map and the feature map, it does not highlight the importance of the channel or region of the feature map itself.

In 2017, Karri et al. [19] used a fine-tuning pre-training model (GoogleNet) to study the OCT images of retinopathy provided by Srinivasan et al. They first changed the fully connection layer of the last layer of the network to 3 outputs, then called the parameters pre-trained on ImageNet in advance as the initialization parameters of the network, and finally used the retinopathy dataset for network training. The accuracy of the final experiment on AMD, DME and NORMAL were 89.00%, 86.00% and 99.00% respectively. Although the pre training model can make the network quickly reach the convergence state, the final classification effect needs to be improved.

In 2019, Feng et al. [20] conducted a classification and recognition study on four types of eye retina images. They first loaded the pre training parameters of VGG16 trained in Imagenet, and then changed the full connection layer of the last layer to four outputs. Finally, the overall accuracy of 97.80% was achieved on CNV, DME, DRUSEN and NORMAL. Although the migration learning method proposed in literature [19, 20] can reduce the training time of the network and alleviate the network's excessive dependence on the dataset, it reduces the generalization of the network to different datasets. For OCT image classification research,

Wang et al. [21] compared VGG16 [22], VGG19 [22], Inception-v3 [23] with CliqueNet [24], DPN92 [25], DenseNet121 [26], ResNet50 [27], ResNext101 [28]. Experiments show that the network with jump connection operation can reduce the loss of effective information in the process of feature extraction and significantly improve the classification effect.

With the development of a deep learning network, current scholars are pursuing the lightness and convenience of the network. Ding X et al. [29] proposed the RepVGG network. RepVGG adds a jump connection operation to the network based on the VGG network. On the premise of ensuring training speed and training accuracy, the accuracy of the RepVGG network on ImageNet dataset can reach more than 80%, therefore, this article uses the RepVGG network to study the retina OCT images.

In addition, the outstanding achievements of attention mechanism in the field of image have also attracted the interest of many scholars. In 2019, Jie et al. [30] proposed SENet network architecture. The core idea of the network is to assign the corresponding channel coefficients to the multi-channel feature maps, so as to highlight the relationship between the channels of the feature maps. Soon after the SENet network was proposed, its enhanced network architecture SKNet [31] was born. Compared with SEnet, SKNet studies the channel-to-channel relationship between two feature maps and has achieved better results on the ImageNet dataset.

According to the characteristics of lesion areas in OCT retinopathy images, we propose a hybrid attention mechanism, which is composed of a channel attention mechanism and a spatial attention mechanism in parallel. It mainly has the following characteristics:

1. By calculating the channel coefficients of the feature maps extracted by the convolution kernels with different expansion rates, the network can automatically identify the importance of the corresponding channel elements between different feature maps.
2. By calculating the spatial coefficients of the feature maps extracted by the convolution kernels of different sizes, the network can automatically identify the importance of regional elements in different feature maps.

## Material and methods

### Datasets

Considering the influence of the number of datasets on the experiment, two public datasets are used in this paper. Among them, Dataset1 was provided by Rasti et al., obtained at the Noor Eye Hospital in Tehran using Spectralis SD-OCT imaging [13]. Dataset2 was provided by Kermany et al, and obtained by using Spectralis SD-OCT imaging [32].

The original train dataset of Dataset2 is 83484 OCT images from 4686 patients. There are 37,205 choroidal neovascularization (CNV), 8616 drusen (DRUSEN), 11348 diabetic macular edema (DME), and 26315 normal(NORMAL). The test dataset includes 250 pictures of CNV, DRUSEN, DME and NORMAL from 633 patients. In the training stage, due to the huge difference in the number of four types of pictures, the model focuses on learning large sample data, so that small sample data is not fully learned, and generalization ability of the network is reduced. Therefore, we processed Dataset2 as follows: 1: Using 8616 DRUSEN images as the standard, 8616 images were randomly selected from CNV, DME and NORMAL. 2: The number of each picture is 8616, we divide it into train dataset and test dataset according to the ratio of 8:2.

Dataset1 has a total of 4254 images, of which 1104 are DME, 1565 are AMD, and 1585 are NORMAL. Similarly, we divide the pictures into train dataset and test dataset according to the ratio of 8:2. Table 1 shows the main statistics of the two datasets in this article.

**Table 1. Dataset statistics.**

| Dataset | | AMD | DME | NORMAL | |
|---|---|---|---|---|---|
| Dataset1 | Train | 1252 | 884 | 1268 | |
| | Test | 313 | 220 | 317 | |
| | | CNV | DRUDEN | DME | NORMAL |
| Dataset2 | Train | 6893 | 6893 | 6893 | 6893 |
| | Test | 1723 | 1723 | 1723 | 1723 |

https://doi.org/10.1371/journal.pone.0261285.t001

Since the quality and size of the two public data sets are different, we used these two data sets for training and evaluation respectively. Experimental data shows that our network model has good generalization ability. Fig 1 show examples of datasets.

## Network architecture

In this section, we will introduce our proposed Multi-branch hybrid attention network (MHA-Net) in detail. As shown in Fig 2, MHANet consists of two parts: channel attention module and spatial attention module. The main function of the channel attention module is to assign channel coefficients to different feature maps to identify the channel-to-channel correlation between the feature maps. The main function of the spatial attention module is to assign spatial coefficients to different feature maps to distinguish the importance of location information between the feature maps.

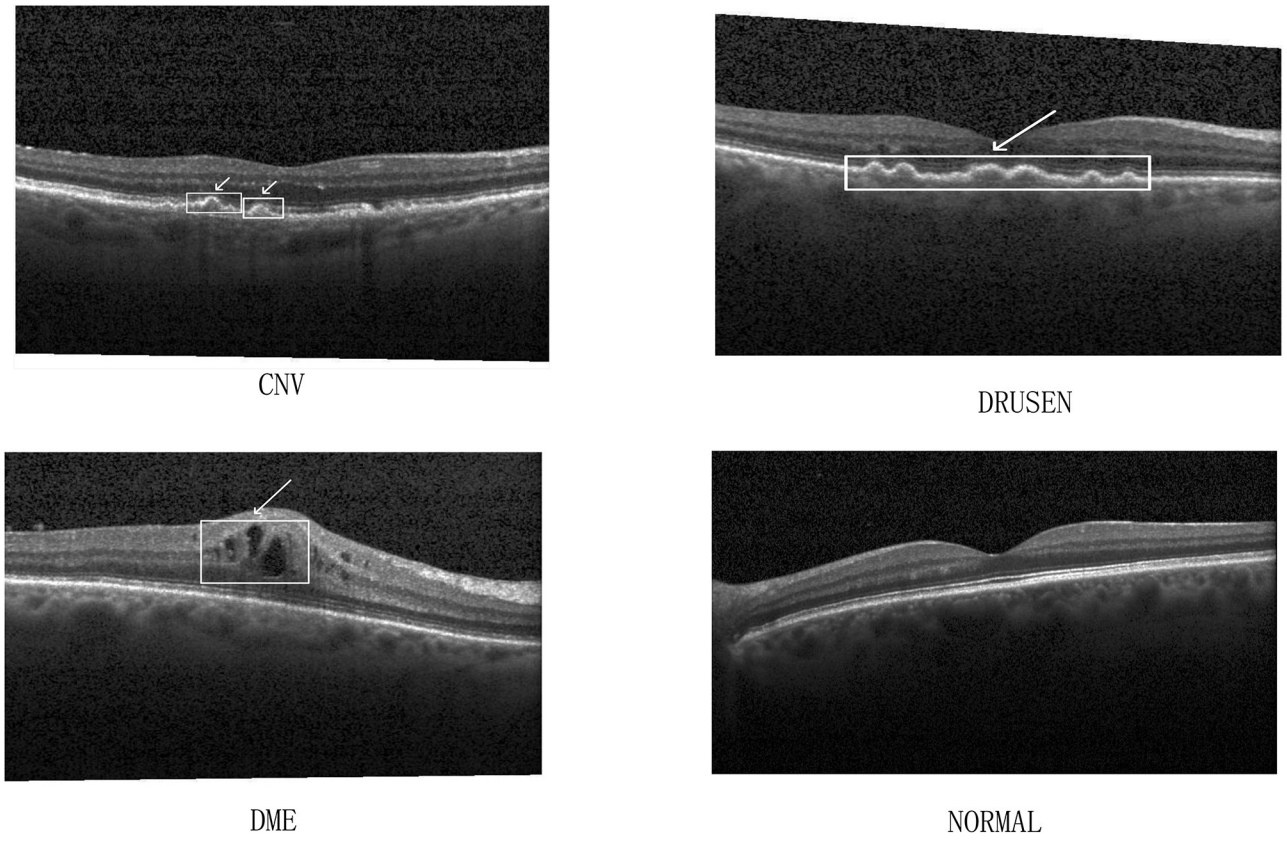


**Fig 1. Sample demo of the dataset.**

https://doi.org/10.1371/journal.pone.0261285.g001

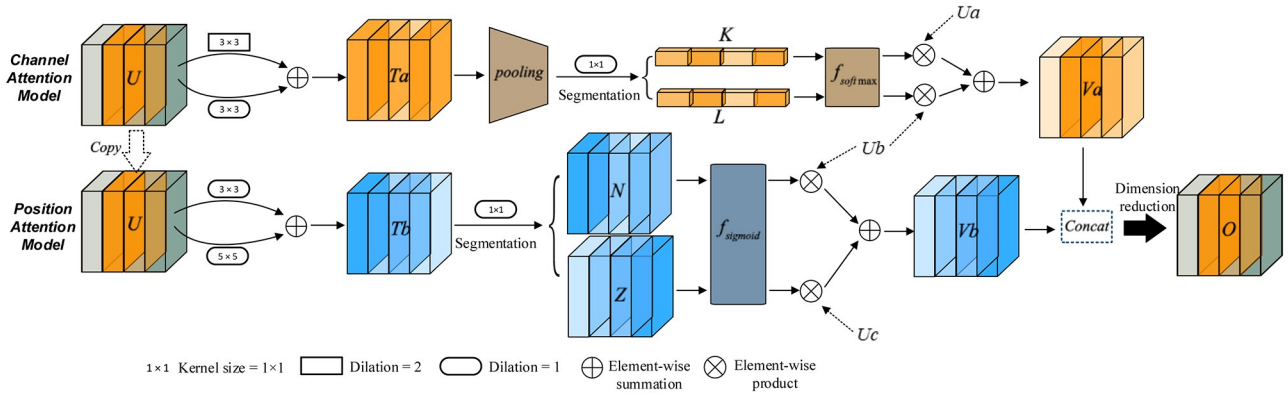


**Fig 2. Overview of the proposed structure Multi-branch hybrid attention network.**

https://doi.org/10.1371/journal.pone.0261285.g002

As shown in Fig 2, $U \in R^{C\times H\times W}$ is the feature map, C is the number of channels,W and H represent length and width of the feature map, respectively. First, we performed three convolution operations on the input feature map: $F1$: $U \rightarrow Ua \in R^{C\times H\times W}$, $F2$: $U \rightarrow Ub \in R^{C\times H\times W}$, and $F3$: $U \rightarrow Uc \in R^{C\times H\times W}$. Among them, F1 means using a convolution operation with a convolution kernel size of 3 × 3 and an expansion rate of 2. F2 represents a convolution operation with a convolution kernel size of 3 × 3 and an expansion rate of 1. F3 represents a convolution operation with a convolution kernel size of 5 × 5 and an expansion rate of 1. Secondly, the element summation operation is carried out, and the feature maps from different branches are fused to obtain $Ta \in R^{C\times H\times W}$ and $Tb \in R^{C\times H\times W}$ respectively.

**Channel attention model of MHANet.** In order to identify the channel-to-channel relationship between different feature maps, we use feature maps $Ta \in R^{C\times H\times W}$ to obtain channel attention coefficients. The specific operations are shown in the following steps:

In order to reduce redundant information in the feature map and improve fault tolerance of the model, we first perform the maximum pooling and average pooling operations on the feature map $Ta$ in the channel dimension to obtain two new feature maps $\{A, E\} \in R^{C\times 1\times 1}$. Then we add and fuse the two new feature maps so that we can get foreground information of the picture as much as possible.

$$A = \mathrm{Max}(Ta) \tag{1}$$

$$E = \mathrm{Mean}(Ta) \tag{2}$$

$$Pa_i = A_i + E_i, i \in \{0, (C-1)\} \tag{3}$$

$$Pa = [Pa_0, Pa_1, \ldots, Pa_{C-1}], Pa_i \in R^{1\times 1} \tag{4}$$

The maximum pooling and average pooling in the channel dimension are to take a maximum value and an average value from the feature map element values of each channel respectively; $A_i \in R^{1\times 1}$ and $E_i \in R^{1\times 1}$ are obtained from the i-th channel feature map of $Ta$ through maximum pooling and average pooling respectively.

In addition, in order to assign channel coefficients to the feature maps $Ua$ and $Ub$, we perform a convolution operation on the feature map $Pa$ to achieve the channel dimension upgrade to obtain the feature map $Pb \in R^{2C\times 1\times 1}$, in which the size of the convolution kernel is 1 × 1. Then the feature map $Pb$ is divided equally in the channel dimension to obtain the

feature maps $K \in R^{C\times1\times1}$ and $L \in R^{C\times1\times1}$.

$$K_i = Pb_i, i \in \{0, (C-1)\} \tag{5}$$

$$L_i = Pb_{i+C} \tag{6}$$

$$Pb = [Pb_0, Pb_1, \ldots, Pb_{C-1}, Pb_C, Pb_{C+1}, \ldots, Pb_{2C-1}] \tag{7}$$

Among them, $Pb_i \in R^{1\times1}$ and $Pb_{i+C} \in R^{1\times1}$ represent the element values in the $i$ and $i + C$ channels of the feature map $Pb$ respectively; $K_i \in R^{1\times1}$ and $L_i \in R^{1\times1}$ represent the element values in the i-th channel of the feature maps K and L respectively.

Finally, we perform a Softmax operation on K and L to obtain the channel attention coefficients $A$ tta $\in R^{C\times1\times1}$ and $A$ ttb $\in R^{C\times1\times1}$, and then we multiply $Atta$ and $Attb$ with $Ua$ and $Ub$ respectively, and the two feature maps obtained after the multiplication are added and fused to obtain the feature map $Va \in R^{C\times H\times W}$.

$$Atta_i = \frac{e^{K_i}}{e^{K_i} + e^{L_i}}, Attb_i = \frac{e^{L_i}}{e^{K_i} + e^{K_i}} \tag{8}$$

$$Va_i = Atta_i \times Ua_i + Attb_i \times Ub_i \tag{9}$$

$$Va = [Va_0, Va_1, \ldots, Va_{C-1}], \quad Va_i \in R^{H\times W} \tag{10}$$

Atta$_i \in R^{1\times1}$ and Att $b_i \in R^{1\times1}$ respectively represent the corresponding coefficients obtained from the element values $K_i$ and $L_i$, and Atta$_i$ + $Attb_i$ = 1; $Ua_i \in R^{H\times W}$ and $Ub_i \in R^{H\times W}$ respectively represent all the element values of the i-th channel of the feature maps $Ua$ and $Ub$.

**Position attention model of MHANet.** In order to enable the network to distinguish the importance of elements in different feature maps, we use the feature map $Tb \in R^{C\times H\times W}$ to calculate the spatial attention coefficient. The specific operation includes the following steps.

In order to assign spatial coefficients to the feature maps $Ua$ and $Ub$, we perform a convolution operation on the feature map $Tb$ to achieve the channel dimension upgrade to obtain the feature map $M \in R^{2C\times H\times W}$, in which the size of the convolution kernel is 1 × 1. Then the feature map M is divided equally in the channel dimension to obtain the feature maps $N \in R^{C\times H\times W}$ and $Z \in R^{C\times H\times W}$.

$$N_j = M_j, j \in \{0, (C-1)\} \tag{11}$$

$$Z_j = M_{j+C} \tag{12}$$

$$M \in [M_0, M_1, \ldots, M_{C-1}, M_C, M_{C+1}, \ldots, M_{2C-1}\} \tag{13}$$

$M_j \in R^{H\times W}$ and $M_{j+C} \in R^{H\times W}$ respectively represent the $H \times W$ element values in the $j$ and $j + C$ channels of the feature map $M$. $N_j \in R^{H\times W}$ and $Z_j \in R^{H\times W}$ represent all the element values in the j-th channel of the feature maps $N$ and $Z$ respectively.

Finally, we use the Sigmoid activation function on the feature maps $N$ and $Z$ to obtain the spatial attention coefficients $Attc \in R^{C\times H\times W}$ and $Attd \in R^{C\times H\times W}$, and then multiply the spatial attention coefficients $Attc$ and $Attd$ with the feature maps $Ub$ and $Uc$ respectively, and the two feature maps obtained after the multiplication are added and fused to obtain the feature map

$Vb \in R^{C\times H\times W}$.

$$\mathrm{Attc}_{(j,h,w)} = \frac{1}{1+e^{N_{(j,h,w)}}}, j \in \{0,(C-1)\}, h \in \{0,(H-1)\}, w \in \{0,(W-1)\} \tag{14}$$

$$\mathrm{Attd}_{(j,h,w)} = \frac{1}{1+e^{Z_{(j,h,w)}}} \tag{15}$$

$$Vb_j = Attc_j \times Ub_j + Attd_j \times Uc_j \tag{16}$$

$$Vb = [Vb_0, Vb_1, \ldots, Vb_{C-1}], Vb_j \in R^{H\times W} \tag{17}$$

$N_{(j,h,w)}$ and $Z_{(j,h,w)}$ respectively represent the element at the position $(h, w)\circ$ in the j-th channel of the feature maps N and Z; $Attc_{(j,h,w)}$ and Attd $_{(j,h,w)}$ are obtained from the corresponding $N_{(j,h,w)}$ and $Z_{(j,h,w)}$ respectively. The specific formula is shown above.

Finally, we concatenate the feature maps $V1$ and $V2$ to obtain the feature map $V \in R^{2C\times H\times W}$, and then use the convolution operation with the convolution kernel size of $1 \times 1$ to reduce the channel dimension of the feature map $V$ to obtain the feature map $O \in R^{C\times H\times W}$. It is worth noting that the final output feature map highlights the foreground information of the picture in both the channel dimension and the spatial dimension.

In addition, we provide pseudo code to introduce the overall flow of the experiment, as shown in Table 2.

## Experiment and statistical analysis method

### Experimental details

The experiments were run based on the Python 3.6, Torch 1.6.0. In the process of network training, after many experiments, it is found that the learning rate is 0.001, the batch size is set to 224, and the network can achieve the optimal effect. The loss function is 'cross-entropy loss', and the optimization method is the SGD gradient descent method.

### Statistical analysis method

Because different evaluation indexes have different evaluation significance to the experimental results, three common evaluation indexes are used to evaluate the experimental results in this paper: accuracy rate, precision rate, and recall rate. Also, to make the experiment more concise and understandable, we use the mixed matrix to show the classification results of each kind of

**Table 2. Training algortithm.**

**Tlgorithm 1** Training algortithm of **MHANet**

1:**for** number of training epochs **do**
2: **for** k steps **do**
3: Sample a batch of images $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ and corresponding labels $\mathbf{Y}_1^{true},\mathbf{Y}_2^{true},\ldots,\mathbf{Y}_n^{true}$ from the training dataset.
4: $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ are resized into a size of 224*224 and $\mathbf{X}_1^{normal},\mathbf{X}_2^{normal},\ldots,\mathbf{X}_n^{normal}$ are obtained with the same size.
5: Input $\mathbf{X}_1^{normal},\mathbf{X}_2^{normal},\ldots,\mathbf{X}_n^{normal}$ into the hybrid attention mechanism network **MHANet**, and obtain output value $\mathbf{Y}_1^{predict},\mathbf{Y}_2^{predict},\ldots,\mathbf{Y}_n^{predict}$.
6: Use the cross-entropy loss function to calculated true lable $\mathbf{Y}_1^{true},\mathbf{Y}_2^{true},\ldots,\mathbf{Y}_n^{true}$ and the predicted label $\mathbf{Y}_1^{predict},\mathbf{Y}_2^{predict},\ldots,\mathbf{Y}_n^{predict}$, and then use the SGD gradient descent method to update the network parameters.
7: end for
8: end for

https://doi.org/10.1371/journal.pone.0261285.t002

**Table 3. Accuracy of Dataset1.**

| Dataset | Model | ACC(%) |
|---|---|---|
| Dataset1 | VGG16 | 96.47 |
| | RepVGG | 90.82 |
| | ResNet50 | 98.00 |
| | Res2Net50 | 97.05 |
| | SENet | 97.41 |
| | SKNet | 99.05 |
| | MHANet(our) | **99.76** |

https://doi.org/10.1371/journal.pone.0261285.t003

data. Accuracy, precision, and recall are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{Rec } all = \frac{TP}{TP + FN} \tag{20}$$

$$F1 = 2\frac{\text{Pre cision} \times \text{Rec all}}{\text{Pre cision} + \text{Rec all}} \tag{21}$$

**Table 4. Precision, Recall, F1 of Dataset1.**

| Dataset | Model | Precision(%) | Recall(%) | F1((%) |
|---|---|---|---|---|
| AMD | VGG16 | 97.74 | 97.12 | 97.43 |
| | RepVGG | 91.16 | 92.33 | 91.74 |
| | ResNet50 | 99.35 | 97.76 | 98.55 |
| | Res2Net50 | 98.37 | 96.80 | 97.58 |
| | SENet | 99.01 | 96.48 | 97.73 |
| | SKNet | 99.67 | 98.72 | 99.19 |
| | MHANet(our) | **1.00** | **99.36** | **99.67** |
| DME | VGG16 | 95.92 | 96.36 | 96.14 |
| | RepVGG | 93.17 | 86.81 | 89.88 |
| | ResNet50 | 96.00 | 98.18 | 97.07 |
| | Res2Net50 | 94.32 | 98.18 | 96.21 |
| | SENet | 94.71 | 97.72 | 96.19 |
| | SKNet | 96.90 | 99.54 | 98.20 |
| | MHANet(our) | **99.09** | **1.00** | **99.54** |
| NORMAL | VGG16 | 95.59 | 95.89 | 95.74 |
| | RepVGG | 89.00 | 92.11 | 90.54 |
| | ResNet50 | 98.10 | 98.106 | 98.10 |
| | Res2Net50 | 97.76 | 96.52 | 97.14 |
| | SENet | 97.79 | 98.10 | 97.95 |
| | SKNet | **1.00** | 99.05 | 99.52 |
| | MHANet(our) | **1.00** | **1.00** | **1.00** |

https://doi.org/10.1371/journal.pone.0261285.t004

**Table 5. Accuracy of Dataset2.**

| Dataset | Model | ACC(%) |
|---|---|---|
| Dataset1 | VGG16 | 92.19 |
| | RepVGG | 94.45 |
| | ResNet50 | 95.31 |
| | Res2Net50 | 95.47 |
| | SENet | 95.24 |
| | SKNet | 95.40 |
| | MHANet(our) | **96.51** |

https://doi.org/10.1371/journal.pone.0261285.t005

TP represents the number of samples that are actually positive and predicted to be positive; TN represents the number of samples that are actually negative and predicted to be negative; FP represents the number of samples that are actually negative and predicted to be positive; FN represents the number of samples that are actually positive and predicted to be negative.

**Table 6. Precision, Recall, F1 of Dataset2.**

| Dataset | Model | Precision(%) | Recall(%) | F1((%) |
|---|---|---|---|---|
| CNV | VGG16 | 94.94 | 92.62 | 93.77 |
| | RepVGG | 94.21 | 94.60 | 94.41 |
| | ResNet50 | 95.38 | 94.77 | 95.08 |
| | Res2Net50 | 95.72 | 94.77 | 95.24 |
| | SENet | 95.64 | 94.25 | 94.94 |
| | SKNet | 95.95 | 95.06 | 95.51 |
| | MHANet(our) | **96.44** | **95.99** | **96.21** |
| DRUSEN | VGG16 | 93.41 | 92.28 | 92.84 |
| | RepVGG | 96.25 | 94.08 | 95.15 |
| | ResNet50 | 97.04 | 95.41 | 96.22 |
| | Res2Net50 | 97.29 | 96.11 | 96.70 |
| | SENet | 96.81 | 95.41 | 96.11 |
| | SKNet | 96.55 | 95.87 | 96.21 |
| | MHANet(our) | **97.60** | **96.92** | **97.26** |
| DME | VGG16 | 90.43 | 89.43 | 89.93 |
| | RepVGG | 94.20 | 92.45 | 93.32 |
| | ResNet50 | 93.62 | 94.66 | 94.14 |
| | Res2Net50 | 93.91 | 94.08 | 93.99 |
| | SENet | 93.66 | 94.37 | 94.01 |
| | SKNet | 94.29 | 93.96 | 94.12 |
| | MHANet(our) | **95.25** | **95.58** | **95.42** |
| NORMAL | VGG16 | 90.13 | 94.42 | 92.23 |
| | RepVGG | 93.22 | 96.69 | 94.92 |
| | ResNet50 | 95.24 | 96.40 | 95.81 |
| | Res2Net50 | 94.99 | 96.92 | 95.94 |
| | SENet | 94.88 | 96.92 | 95.89 |
| | SKNet | 94.82 | 96.69 | 95.74 |
| | MHANet(our) | **96.77** | **97.56** | **97.16** |

https://doi.org/10.1371/journal.pone.0261285.t006

## Experimental results

In order to better illustrate the positive effect of the hybrid attention mechanism on the recognition of retinopathy, we give the classification accuracy of the entire dataset and the classification index score of each type of data.

From the accuracy score of each model on Dataset1, it can be seen that the attention mechanism plays an active role in the three types of image recognition of AMD, DME, and NORMAL. In particular, our proposed MHANet network achieved the best results in retinal image classification, and the overall accuracy reached 99.76%, as shown in Table 3.

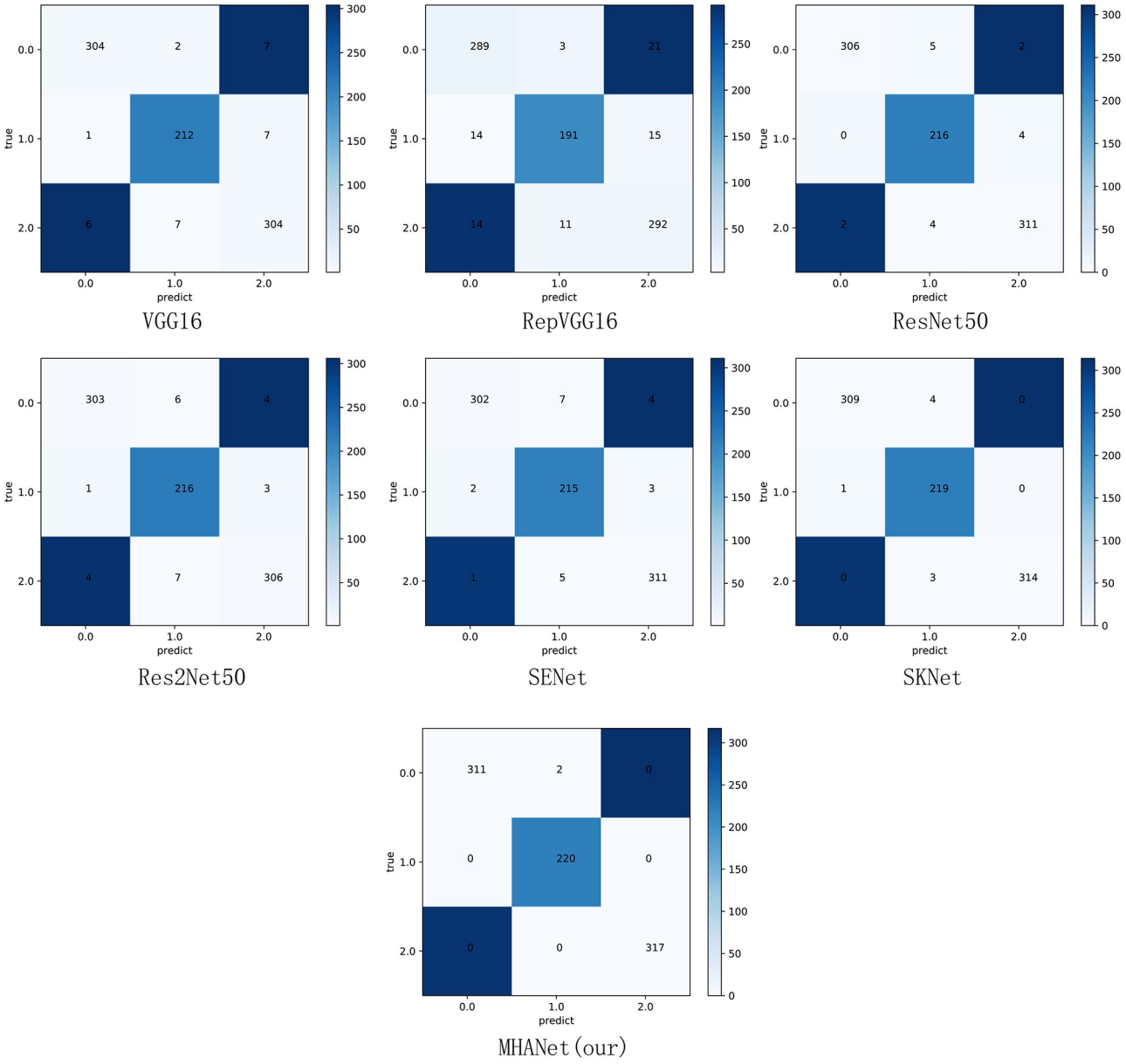


**Fig 3. Confusion matrix of the Dataset1.**

https://doi.org/10.1371/journal.pone.0261285.g003

Table 4 shows the Precision, Recall, and F1 of each type of picture. It can be seen from the table that the F1 values obtained by ResNet50, Res2Net50, SENet and SKNet on AMD and NORMAL are significantly better than their F1 values obtained on DME, which indicates that generalization ability of the above models has certain defects, and hybrid attention proposed in this article effectively alleviates this problem.

For the Dataset2, the four types of pictures are divided according to the ratio of 1:1:1:1, so the network will not have the over fitting problem caused by the large difference in the number of training pictures. The MHANet proposed in this paper adopts both spatial attention mechanism and channel domain attention mechanism, and achieves the best results. See Table 5 for specific experimental data.

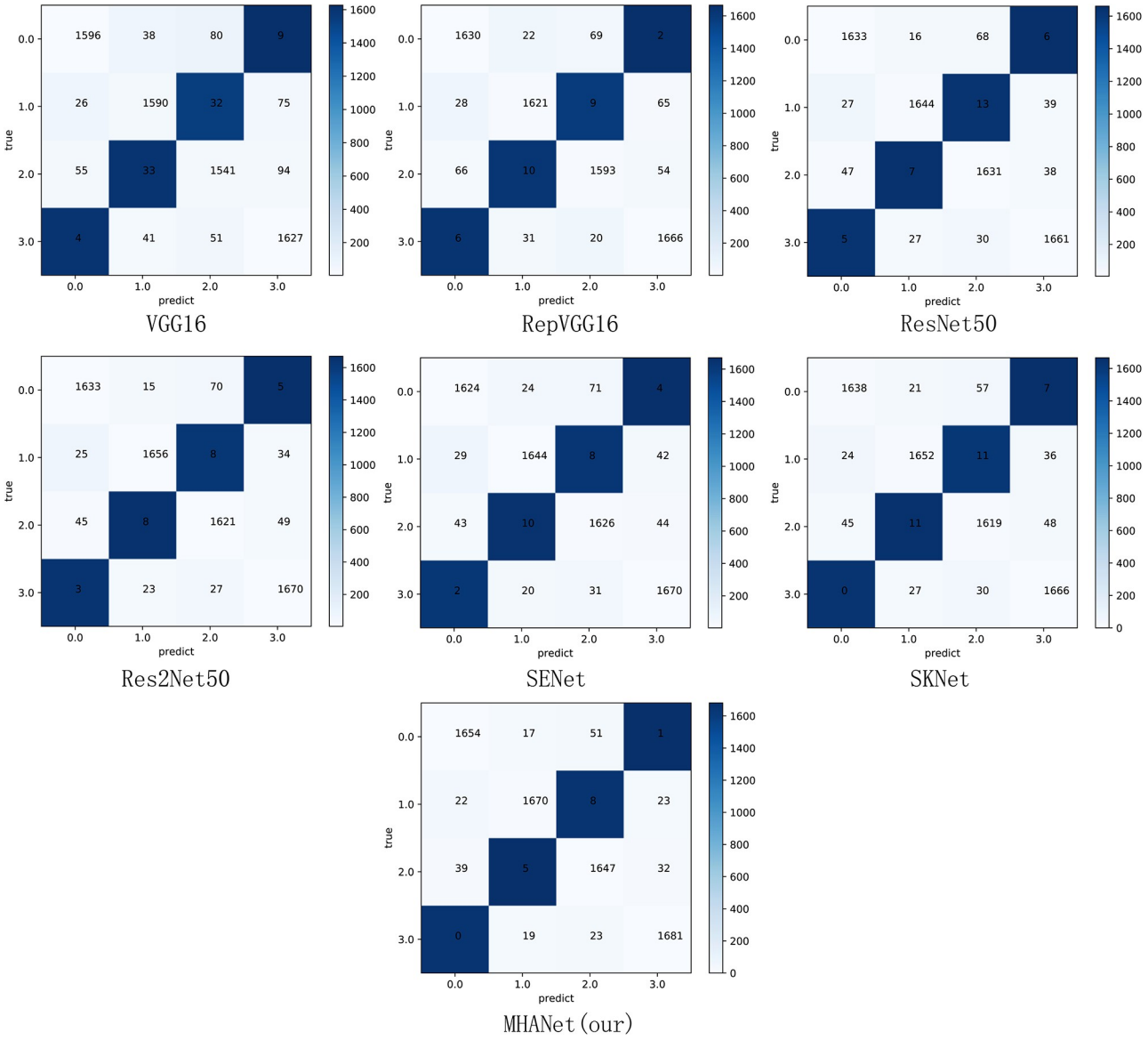


**Fig 4. Confusion matrix of the Dataset2.**

https://doi.org/10.1371/journal.pone.0261285.g004

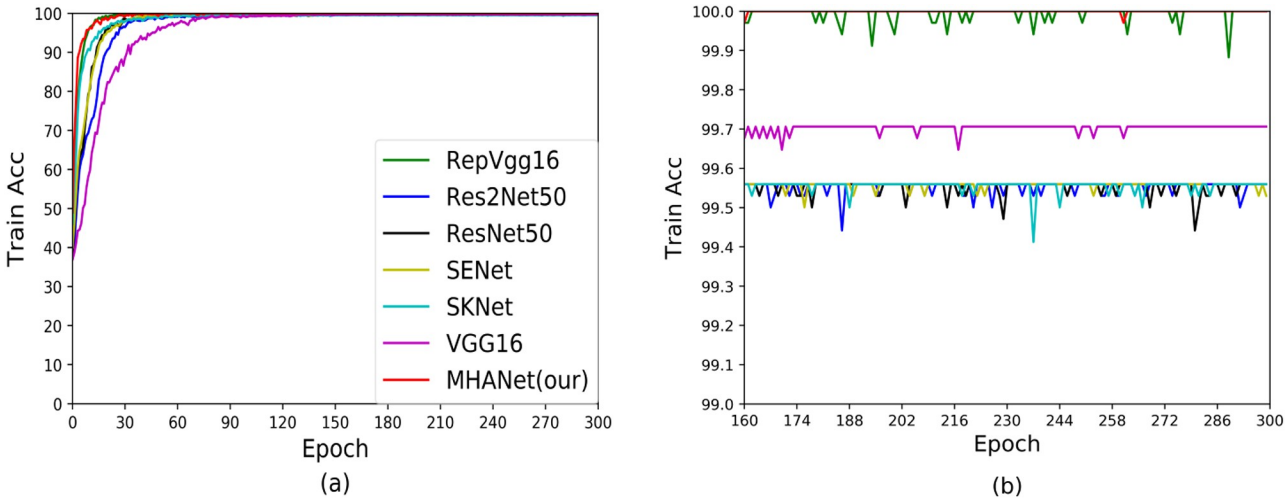


**Fig 5. Training accuracy curves of Dataset1.**

https://doi.org/10.1371/journal.pone.0261285.g005

As shown in Table 6, the MHANet network proposed in this paper not only achieved the best results in accuracy, but also has certain improvements in Precision, Recall, and F1 compared to other networks. This is because the MHANet network can not only automatically highlight the importance of channels, but also distinguish the importance of element values in the entire feature map. In order to better prove the feasibility of the innovation in this paper, this paper further demonstrates the experimental results of Dataset1 and Dataset2 through a confusion matrix. The results are shown in Figs 3 and 4.

As shown in Fig 3, we give the confusion matrix obtained by each model based on Dataset1, where 0.0 represents AMD, 1.0 represents DME and 2.0 represents NORMAL. From each confusion matrix, we can know that other models will more or less misclassify these three types of pictures. MHANet proposed in this paper can completely and correctly classify DME and NORMAL, and minimize the error of AMD as DME.

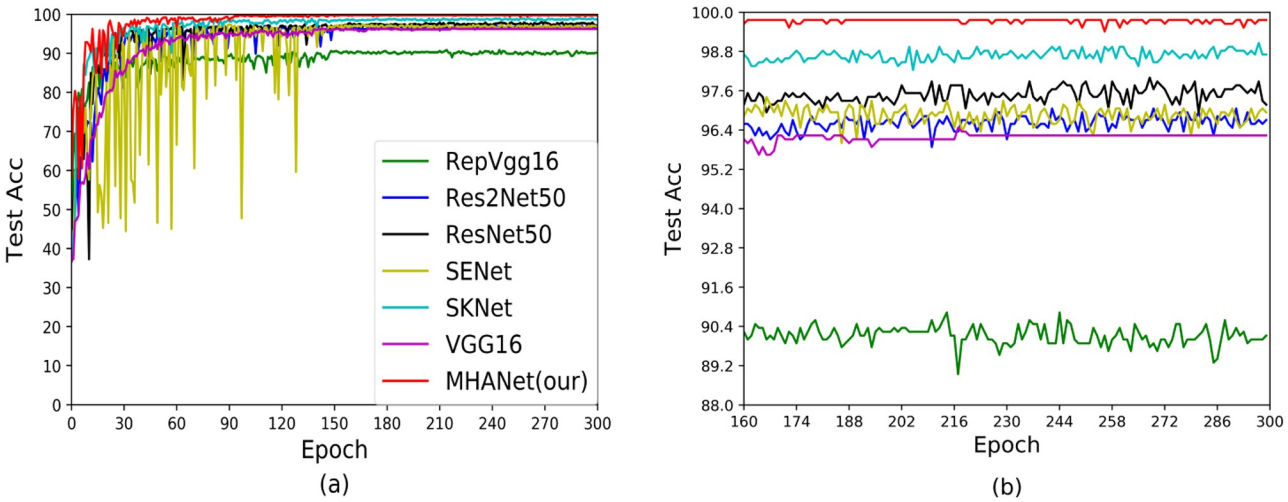


**Fig 6. Test accuracy curves of Dataset1.**

https://doi.org/10.1371/journal.pone.0261285.g006

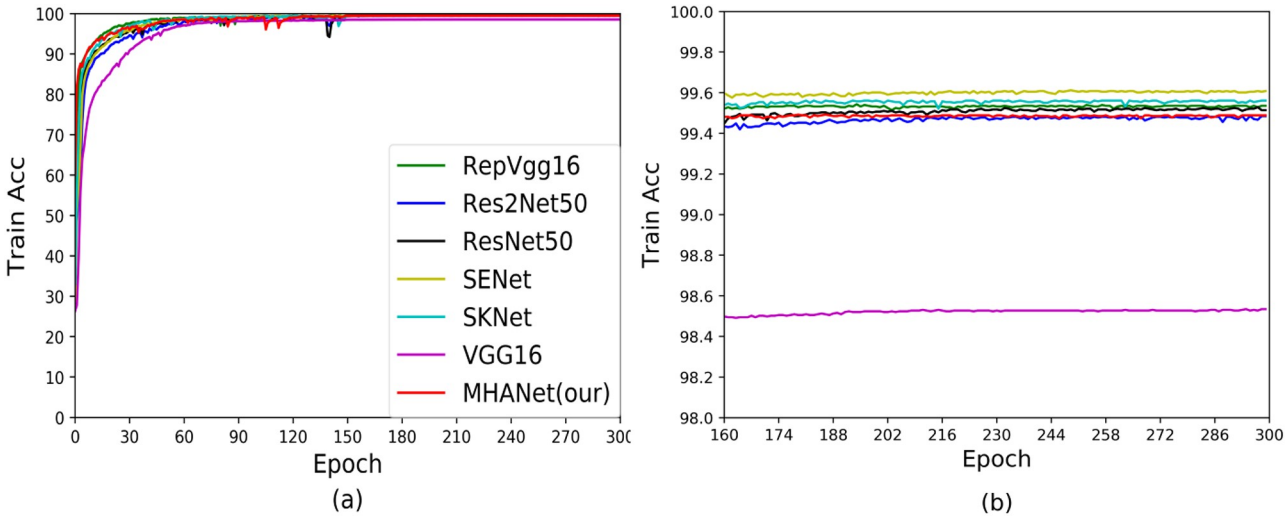


**Fig 7. Training accuracy curves of Dataset2.**

https://doi.org/10.1371/journal.pone.0261285.g007

As shown in Fig 4, we give the confusion matrix obtained by each model based on Dataset2. Among them, 0.0 represents CNV, 1.0 represents DRUSEN, 2.0 represents DME, and 3.0 represents NORMAL. From the perspective of the model, the MHANet model has the largest number of correct image classifications among the five models. Among them, MHANet has 298 correctly classified pictures more than VGG16, 142 correctly classified pictures more than RepVGG, 80 correctly classified pictures more than ResNet50, 72 correctly classified pictures more than Res2Net50, 88 correctly classified pictures more than SENet and 77 correctly classified pictures more than SKNet. This fully proves that this model has achieved the best effect in the task of retinal disease recognition. In addition, other models are prone to mispredict sample 2 as sample 0 and mispredict sample 3 as sample 2. For these two kinds of images that are easy to distinguish errors, MHANet has the best classification effect.

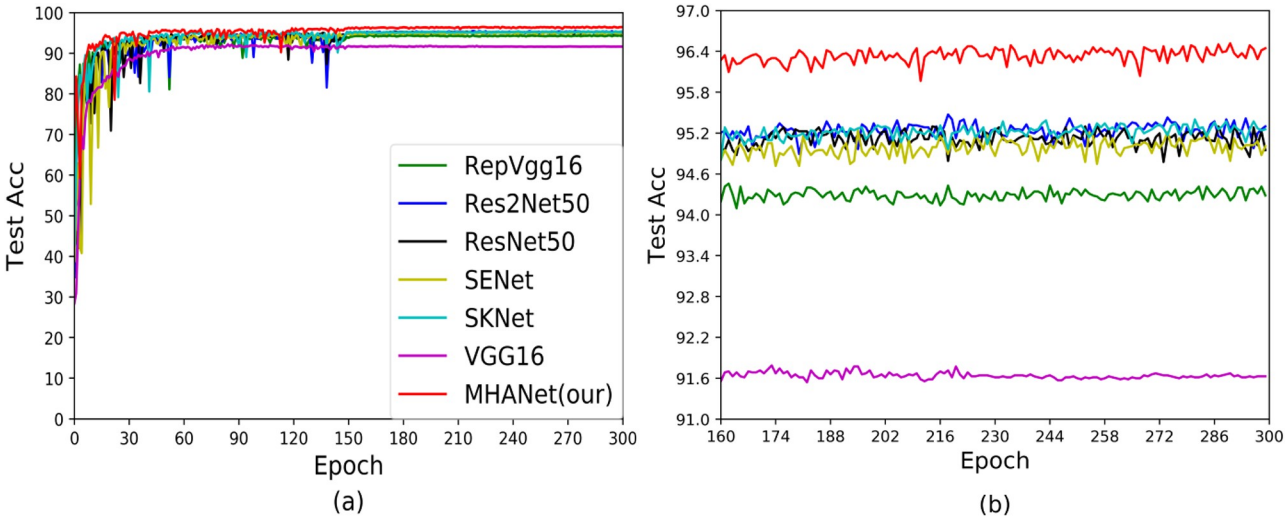


**Fig 8. Test accuracy curves of Dataset2.**

https://doi.org/10.1371/journal.pone.0261285.g008

In order to analyze the generalization ability of the model, we give the accuracy curve during model training and the accuracy curve during testing, as shown in Figs 5–8. Among them, Figs 5 and 6 are the accuracy curves during training and testing based on Dataset1. It can be seen from the figure that VGG16 and RepVGG can get very high scores during training, but the test results are not satisfactory, indicating that generalization ability of VGG16 and RepVGG on Dataset1 is not strong. ResNet50 and Res2Net50 have achieved good results in both the train stage and the test stage, indicating that ResNet50 and Res2Net50 have better generalization capabilities on Dataset1, but ResNet50 is better than Res2nNet50 on the test set. This is because Res2Net50 has more parameters than ResNet50, which makes it difficult for the network to reach the best state. Similarly, it can be seen from the figure that SENet and SKNet have good generalization ability on Dataset1, but the test results have not been greatly

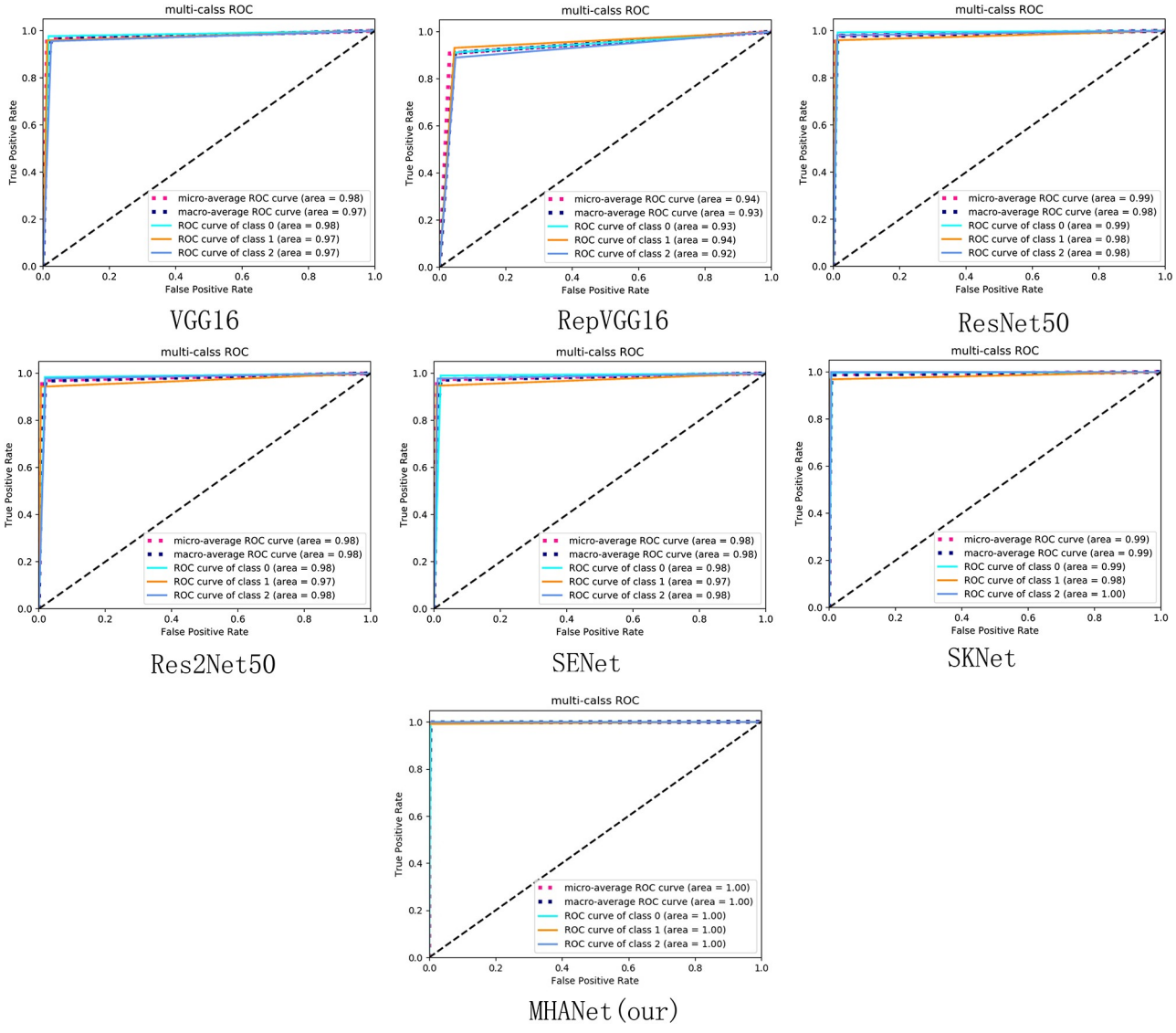


**Fig 9. The micro-average ROC curve is obtained by the micro method in the sklearn.metrics.roc-auc-score function.** The macro-average ROC curve is obtained by the macro method in the sklearn.metrics.roc-auc-score function. Class 0, class 1, and class 2 in the figure represent AMD, DME, and NORMAL, respectively.

https://doi.org/10.1371/journal.pone.0261285.g009

improved, so the network has certain limitations. Finally, we can see that the MHANet network can get the best results in the train stage and the test stage, which proved that MHANet has good convergence and has made breakthrough.

Figs 7 and 8 show the accuracy during training and testing based on Dataset2. It can be seen from the figure that compared with Dataset1, RepVGG performed better on Dataset2. In addition, ResNet50, Res2Net50, SENet, SKNet and MHANet achieved convergence in the train and test stages, but MHANet achieved the best experimental results in the testing phase. This once again proved that the MHANet network can get good experimental results on small datasets, and good results can be obtained on large datasets.

As shown in Fig 9, The AUC values obtained by SENet, SKNet and MHANet are all above 0.96 on Dataset 1 and the average AUC values of three types are all above 0.97,which shows

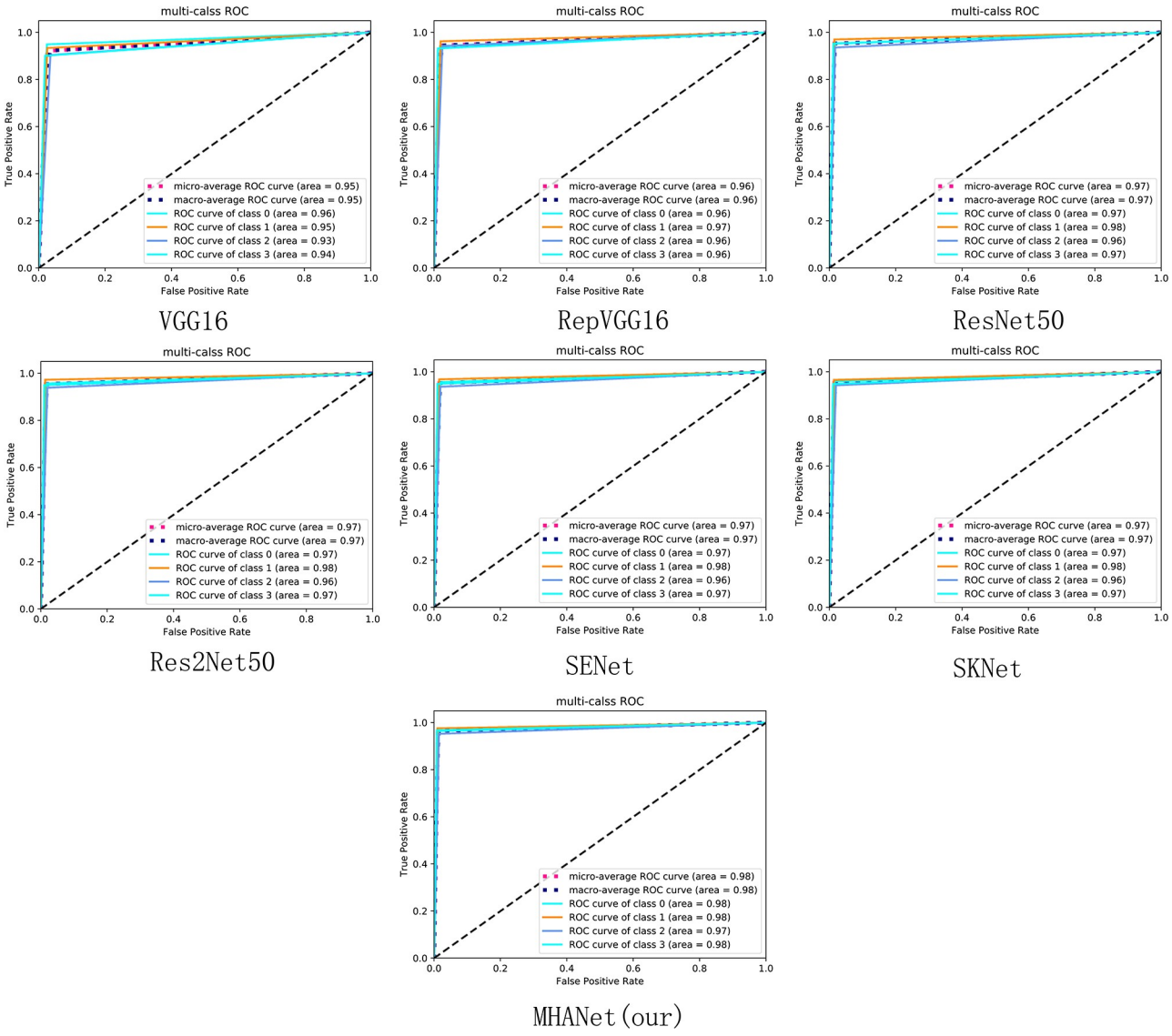


**Fig 10. The micro-average ROC curve is obtained by the micro method in the sklearn.metrics.roc-auc-score function.** The macro-average ROC curve is obtained by the macro method in the sklearn.metrics.roc-auc-score function. Class 0, class 1, and class 2 in the figure represent CNV, DURSEN, DME, and NORMAL, respectively.

https://doi.org/10.1371/journal.pone.0261285.g010

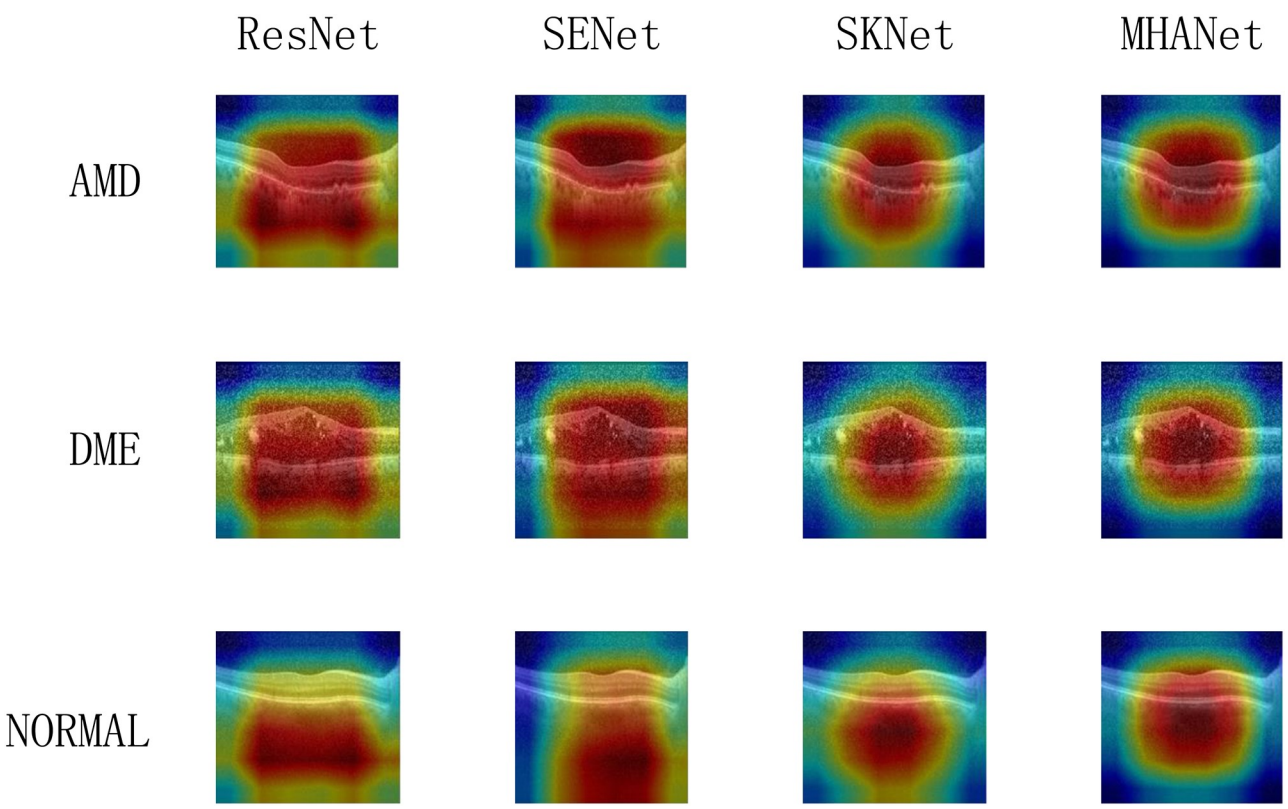


**Fig 11. Heat map of Dataset1.**

https://doi.org/10.1371/journal.pone.0261285.g011

that attention mechanism model we trained on Dataset 1 can well balance the precision and recall indicators. Among them, the AUC values of the MHANet model proposed in this paper all reached 1, which also shows that the hybrid attention mechanism can have a better effect in retinal disease classification experiments than using only the channel attention mechanism.

As shown in Fig 10, The AUC values obtained by ResNet, Res2Net, SENet, and SKNet on Dataset 2 and the average AUC values of the three types are not much different, which shows that it is difficult to improve the discrimination ability of the network on Dataset2 using only the channel attention mechanism. The AUC value obtained on Dataset2 and the average AUC value of the three types of MHANet, which is composed of the parallel channel attention mechanism and the spatial attention mechanism, are the highest.

As shown in Figs 11 and 12, we show the focusing ability of different networks. Among them, ResNet50 without attention mechanism has completely different focusing ability on two different data sets. On the Dataset1, the focus range of ResNet50 is wide and the focus center does not accurately fall on the foreground information part. On the Dataset2, ResNet50 shrinks the focus range but the focus center still does not accurately fall on the foreground information part. SENet and SKNet adopt different channel attention mechanisms that cause their focusing abilities to be different. On the two datasets, the focusing ability of SKNet is obviously better than that of SENet. However, the focus centers of SENet and SKNet are still not accurately gathered in foreground information. The MHANet network we proposed has a small focus range on two different datasets and the focus center falls on the foreground information part.

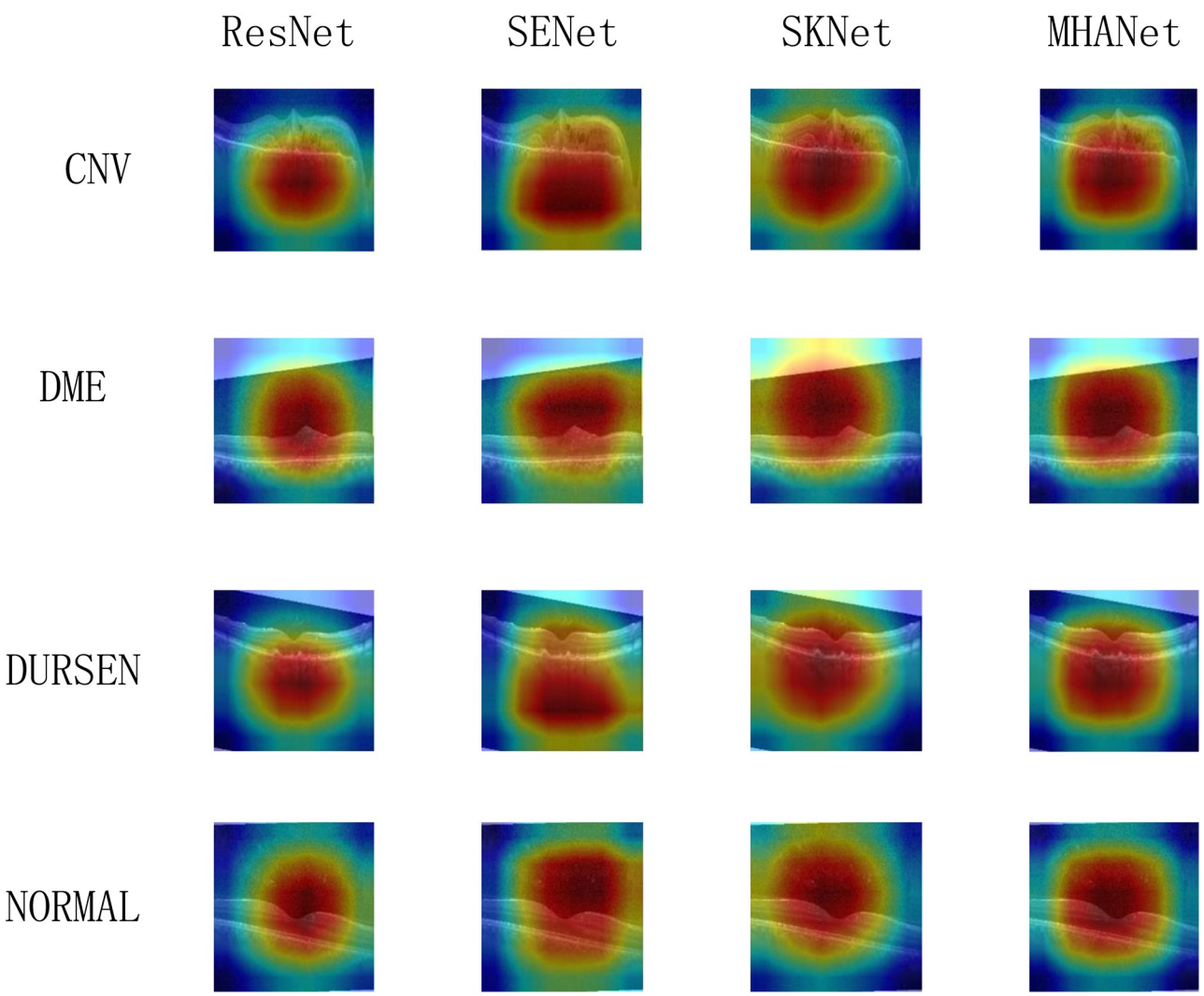


**Fig 12. Heat map of Dataset2.**

https://doi.org/10.1371/journal.pone.0261285.g012

## Conclusion

According to the distribution of lesion features in retinopathy images, we propose a hybrid attention mechanism to help the network lock foreground information in the image more accurately. The MHANet network module is composed of a parallel channel attention mechanism and a spatial attention mechanism. The channel attention mechanism assigns different channel coefficients to each channel feature map in the channel dimension, so as to highlight the channel feature map with the most abundant lesion features. The spatial attention mechanism assigns corresponding position coefficient to each element in the spatial dimension, so as to highlight the importance of elements in the lesion area. This hybrid attention mechanism can help the network focus on the lesion area in both the spatial dimension and the channel dimension. In this paper, the feasibility of the MHANet module is verified on two public retinopathy datasets. The experimental results show that the MHANet module can not only improve the accuracy of network classification and recognition, but also improve the generalization ability of the network.

## Author Contributions

**Conceptualization:** Lianghui Xu.

**Formal analysis:** Liejun Wang, Shuli Cheng.

**Funding acquisition:** Liejun Wang.

**Methodology:** Yongming Li.

**Resources:** Liejun Wang.

**Software:** Lianghui Xu.

**Supervision:** Liejun Wang, Yongming Li.

**Validation:** Shuli Cheng, Yongming Li.

**Visualization:** Shuli Cheng.

**Writing – original draft:** Lianghui Xu.

**Writing – review & editing:** Liejun Wang.

## References

1. Romero-Aroca Pedro. Current status in diabetic macular edema treatments. World journal of diabetes. 2013; 4(5):165. https://doi.org/10.4239/wjd.v4.i5.165 PMID: 24147200
2. Wong Tien Y and Sun Jennifer and Kawasaki Ryo and Ruamviboonsuk Paisan and Gupta Neeru and Lansingh Van Charles et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018; 125(10):1608–1622. https://doi.org/10.1016/j.ophtha.2018.04.007 PMID: 29776671
3. Mandić Krešimir and Vukojević Nenad and Jukić Tomislav and Katušić Damir and Mandić Jelena Juri. Changes of drusen number and central retinal thickness in age-related macular degeneration patients over two years. Acta Clin Croat. 2016; 55(3):354–9. PMID: 29045093
4. Friedman David S and O'Colmain Benita J and Munoz Beatriz and Tomany Sandra C and McCarty Cathy and De Jong PT et al. Prevalence of age-related macular degeneration in the United States. Arch ophthalmol. 2004; 122(4):564–572. https://doi.org/10.1001/archopht.122.4.564 PMID: 15078675
5. Bressler Neil M. Early detection and treatment of neovascular age-related macular degeneration. The Journal of the American Board of Family Practice. 2002; 15(2):142–152. PMID: 12002198
6. Javitt Jonathan C and Aiello Lloyd Paul and Chiang Yenpin and Ferris Frederick L and Canner Joseph K and Greenfield Sheldon. Preventive eye care in people with diabetes is cost-saving to the federal government: implications for health-care reform. Diabetes care. 1994; 17(8):909–917. https://doi.org/10.2337/diacare.17.8.909 PMID: 7956643
7. Krans Hendrik Michiel Jan and Porta M and Keen Harry. Diabetes care and research in Europe: the St. Vincent Declaration action programme: implementation document. 1992.
8. Hendra Timothy J and Sinclair Alan J. Improving the care of elderly diabetic patients: the final report of the St Vincent Joint Task Force for Diabetes. Age and ageing. 1997; 26(1):3–6. https://doi.org/10.1093/ageing/26.1.3 PMID: 9143430
9. Litjens Geert and Kooi Thijs and Bejnordi Babak Ehteshami and Setio Arnaud Arindra Adiyoso and Ciompi Francesco and Ghafoorian Mohsen et al. A survey on deep learning in medical image analysis. Medical image analysis. 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
10. Huang David and Swanson Eric A and Lin Charles P and Schuman Joel S and Stinson William G and Chang Warren et al. Optical coherence tomography. science. 1991; 254(5035):1178–1181. https://doi.org/10.1126/science.1957169 PMID: 1957169
11. Krizhevsky Alex and Sutskever Ilya and Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012; 25:1097–1105.
12. Oulhaj, Hind and Amine, Aouatif and Rziza, Mohammed and Aboutajdine, Driss. Noise Reduction in Medical Images-comparison of noise removal algorithms. 2012 International Conference on Multimedia Computing and Systems. 2012;344–349.

13. Rasti Reza and Rabbani Hossein and Mehridehnavi Alireza and Hajizadeh Fedra. Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE transactions on medical imaging. 2017; 37(4):1024–1034. https://doi.org/10.1109/TMI.2017.2780115

14. Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.

15. Srinivasan Pratul P and Kim Leo A and Mettu Priyatham S and Cousins Scott W and Comer Grant M and Izatt Joseph A et al. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomedical optics express. 2014; 5 (10):3568–3577. https://doi.org/10.1364/BOE.5.003568 PMID: 25360373

16. Wang Yu and Zhang Yaonan and Yao Zhaomin and Zhao Ruixue and Zhou Fengfeng. Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images. Biomedical optics express. 2016; 7(12):4928–4940. https://doi.org/10.1364/BOE.7.004928 PMID: 28018716

17. Vidyasagar Mathukumalli. Machine learning methods in the computational biology of cancer. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2014; 470 (2167):20140081. https://doi.org/10.1098/rspa.2014.0081 PMID: 25002826

18. Hall Mark A. Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato. 1998.

19. Karri Sri Phani Krishna and Chakraborty Debjani and Chatterjee Jyotirmoy. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomedical optics express. 2017; 8(2):579–592. https://doi.org/10.1364/BOE.8.000579 PMID: 28270969

20. Li Feng and Chen Hua and Liu Zheng and Zhang Xuedian and Wu Zhizheng. Fully automated detection of retinal disorders by image-based deep learning. Graefe's Archive for Clinical and Experimental Ophthalmology. 2019; 257(3):495–505. https://doi.org/10.1007/s00417-018-04224-8 PMID: 30610422

21. Wang Depeng and Wang Liejun. On OCT image classification via deep learning. IEEE Photonics Journal. 2019; 11(5):1–14.

22. Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

23. Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;2818–2826.

24. Yang, Yibo and Zhong, Zhisheng and Shen, Tiancheng and Lin, Zhouchen. Convolutional neural networks with alternately updated clique. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018;2413–2422.

25. Chen, Yunpeng and Li, Jianan and Xiao, Huaxin and Jin, Xiaojie and Yan, Shuicheng and Feng, Jiashi. Dual path networks. arXiv preprint arXiv:1707.01629. 2017.

26. Huang, Gao and Liu, Zhuang and Van Der Maaten, Laurens and Weinberger, Kilian Q. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;4700–4708.

27. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;770–778.

28. Xie, Saining and Girshick, Ross and Dollár, Piotr and Tu, Zhuowen and He, Kaiming. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;1492–1500.

29. Ding, Xiaohan and Zhang, Xiangyu and Ma, Ningning and Han, Jungong and Ding, Guiguang and Sun, Jian. Repvgg: Making vgg-style convnets great again. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021;13733–13742.

30. Hu, Jie and Shen, Li and Sun, Gang. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018;7132–7141.

31. Li, Xiang and Wang, Wenhai and Hu, Xiaolin and Yang, Jian. Selective kernel networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019;510–519.

32. Kermany Daniel S and Goldbaum Michael and Cai Wenjia and Valentim Carolina CS and Liang Huiying and Baxter Sally L et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2019; 172(5):1122–1131. https://doi.org/10.1016/j.cell.2018.02.010