Intelligent detection method and applied research of diabetic retinopathy based on residual attention network

Moye Yu (yumoye@126.com)
Fudan University Shanghai Cancer Center, Fudan University

Research Article

Keywords: Diabetic Retinopathy, Fundus Image, CNN, Attention Mechanism, Dilated Convolution, Deep Learning Assisted Diagnosis

Posted Date: June 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-646359/v1

Version of Record: A version of this preprint was published at International Journal of Imaging Systems and Technology on April 18th, 2022. See the published version at https://doi.org/10.1002/ima.22734.

Intelligent detection method and applied research of diabetic retinopathy based on residual attention network

Moye Yu*
Department of Information Center, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China, yumoye@126.com

Abstract

Diabetic Retinopathy (DR) is a late-stage ocular complication of diabetes. Proposing a high-accuracy automatic screening technology of fundus images based on deep learning is of great significance to delay the deterioration of DR. In this paper, we propose an end-to-end framework RAN for DR classification and diagnosis based on the ResNet, attention mechanism and dilated convolution was added to the framework. We implemented experiments on three DR datasets, Kaggle, Messidor and IDRid, analyzed and compared the experimental results. The focal loss function is added to solve the imbalance problem between DR datasets. The results show that the method RAN used mainly improves the results of the basic neural network when using the same dataset. Therefore, by optimizing the basic neural network, the classification and diagnosis effect of DR can be improved.

Keywords: Diabetic Retinopathy, Fundus Image, CNN, Attention Mechanism, Dilated Convolution, Deep Learning Assisted Diagnosis

1.Introduction

Diabetes Mellitus (DM) is a common endocrine system disease. The prevalence of diabetes in all ages was 2.8% in 2000 and 4.4% by 2030 of the world . The total number of diabetics is expected to increase from 171 million in 2000 to 366 million in 2030. Women have more medical records than men. The urban population of developing countries is expected to double between 2000 and 2030. The most significant demographic change in the prevalence of diabetes worldwide is the increase in the proportion of people over 65 years of age. Because of the rising

prevalence of obesity worldwide, the future prevalence of diabetes will continue to rise, and the burden of diabetes will also increase[1]. Diabetes is associated with life, the longer it is discovered and the prolonger diabetes is, the higher the risk of complications is. Eventually, complications of diabetes can be disabling and even life-threatening.

Diabetic retinopathy (DR) is a late manifestation of diabetes, and one of the most severe complications of diabetic microangiopathy. If it is not detected and treated early, it will cause irreversible visual impairment, and in severe cases it may cause blindness. Fundus image is an vital inspection method for early detection of DR lesions. Due to many ophthalmologists in the less developed areas are lacking, patients with diabetes lack early diagnosis and treatment of DR. Therefore, computerized screening technology based on fundus images is of great significance to delay the deterioration of DR.

Fundus changes in DR mainly include microaneurysms, hard exudates, bleeding spots, cotton-wool patches, Diabetic Macular Edema (DME), as shown in Figure 1.1. DR is divided into 5 levels, as is shown in Table1.1-1.2, Figure 1.2, for diabetic patients, regular DR screening is very essential. There are usually three methods for diagnosing DR: ①Ophthalmoscope or indirect ophthalmoscope can observe the fundus retina of diabetic patients after dilation of pupils; ②The patient's color retina images is also the primary method of early DR screening. Figure 1.1 shows the DR color retina images obtained by the fundus camera; ③Optical Coherence Tomography (OCT): can provide high-resolution cross-sectional images of the retina, showing its thickness.

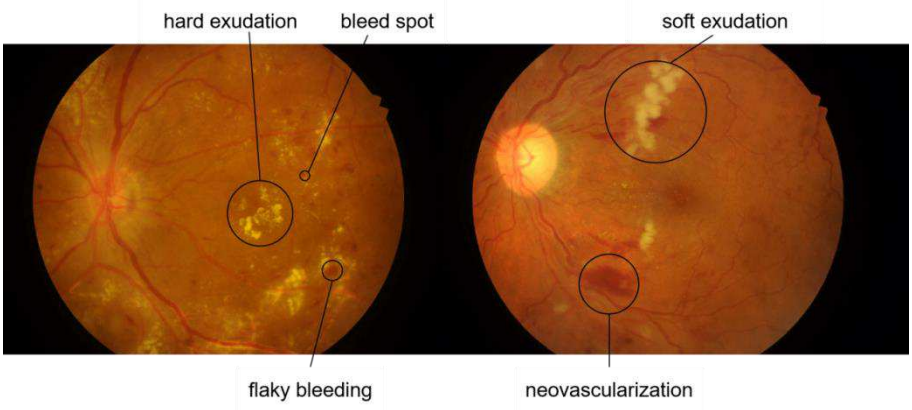

Figure 1.1 Fundus changes in DR lesions

Table 1.1 International DR clinical classification (see after examination of dilated pupils)[2]

DR Level	Fundus Examination
No obvious retinopathy	No abnormality
Nonproliferative DR, Mild	Microaneurysms only
Moderate	Besides microaneurysms, there are still a few hard exudative spots or small bleeding spots.
Severe	There are no signs of proliferative DR, but besides moderate lesions, there is still one of the following three (4, 2, 1 regulation)： More than 20 retinal vein beads in four quadrants, Two quadrants have clear retinal vein beads, One quadrant has obvious IRMA.
Proliferative DR	One or more of the following changes: 1. Neovascularization 2. Preretinal bleeding 3. Vitreous blood.

Figure 1.2 Schematic diagram of 5 grades of authentic clinical diabetic retina images[2]

Table 1.2 Clinical classification of diabetic macular edema (DME)[2]

DME Level	Fundus Examination
No obvious DME	No noticeable thickness of retina or hard exudate at the posterior pole.
There is obvious DME	A significant thickness of retina or hard exudate in the posterior pole.
Mild DME	The thickness of retina or hard exudates away from the fovea.
Moderate DME	The thickness of retina or hard exudates does not affect the fovea.
Severe DME	The thickness of retina or hard exudates affects the fovea.

High-quality color retina images can assist doctors in the diagnosis and judgment of retinopathy. However, the diagnosis of DR requires a clinically experienced ophthalmologist, and DR screening has not carried out in most grass-roots areas, which has significantly increased the risk of blindness due to diabetes[3]. Therefore, computer-assisted remote diagnostic technology in fundus images can effectively reduce the visual impairment of diabetic patients caused by insufficient medical resources. This study intends to use deep learning (DL) methods to process the fundus images, laying the foundation for the remote automatic fundus image screening system.

2. Related Work

At present, most of the work in the field of ophthalmology image analysis focuses on the DR classification, segmentation, and detection of retina structures, such as optic disc, macular, blood vessel, abnormal parts (hard osmosis, soft osmosis, bleeding spots, microaneurysms), Table 2.1. Common retina diseases use deep learning techniques to assist in diagnosis including DR, age-related macular degeneration (AMD), glaucoma, etc.

Pratt et al.[4] developed a network with CNN architecture and data augmentation, which can identify the intricate features involved in the classification task, trained the model on the Kaggle dataset, and achieved a sensitivity of 95% and an accuracy of 75% on 5,000 validation images. Chandrakumar et al.[5] proposed a Deep Convolution Neural Network (DCNN) method which gives high accuracy in the classification of retinal diseases through spatial analysis, using 35,126 Kaggle fundus images, the accuracy of normal, mild, moderate, NPDR and PDR was 99%, 83%, 79%, 84%, 91%, respectively. Rahim et al.[6] presents an automatic detection method of diabetic retinopathy and maculopathy in fundus images by employing fuzzy image processing techniques. A combination of fuzzy image processing techniques, the circular hough transform, and several feature extraction methods are implemented.

Eftekhari et al.[7] used a two-step process and two online datasets to train CNN, which can solve the problem of imbalance and reduce training time while accurately detecting. Seth et al.[8] used convolutional neural networks and linear support vector machines to train the network on the benchmark dataset EyePACS dataset. Experimental results show that the model has high sensitivity and specificity in detecting diabetic retinopathy. Dutta et al.[9] proposed an automatic knowledge model to identify critical prerequisites for disaster recovery. After testing using a CPU-trained neural network model, three types of back-propagation neural networks were used. The model was able to quantify the characteristics of different types of blood vessels, exudates, bleeding, and microaneurysms.

Benzamin et al.[10] proposed a deep learning algorithm based on CNN, which can detect hard exudates in fundus images and assist ophthalmologists in diagnosis. Adem et al.[11] used a CNN model with a circular Hough transform to apply it to retinal images. The results show that the using of the CNN model together with image processing methods can improve the accuracy, and the success rate of exudates

detection is 99.18%. Li et al.[12] developed a DL system for GON classification to classify GON on color fundus images automatically . The area under curve (AUC) of the DL system is 0.986, and the sensitivity is 95.6 %, the specificity is 92.0%. Based on these works, this article will combine the needs of ophthalmologists and diabetic patients, optimize the application of DL technology in ophthalmology clinics, improve the accuracy of model diagnosis, and assist clinicians in their work.

Table 2.1 Experimental methods, datasets, and results used in recent papers

	Method	Dataset	Performance
Gulshan et al.[13]	Inception v3	EyePACS Messidor	Specificity 93.4%，Sensitivity 97.5% Specificity 93.9%，Sensitivity 96.1%
Li et al.[14]	VggNet、GoogleNet	DR1 Messidor	Sensitivity 97.11%，Specificity 86.03% Accuracy 92.01%，AUC0.9834
Gargeya et al.[15]	Data driven DNN ResNet	EyePACS Messidor	AUC0.97 AUC0.94
Abramoffff et al.[16]	DCNN	Messidor	Sensitivity 96.8%,Specificity 87.0%, AUC 0.98
Ting et al.[17]	VGG-19	10 datasets recruited from 6 countries	Sensitivity 90.5%,Specificity 91.6%, AUC 0.936.
Li et al.[18]	Inception v3	Chinese color fundus images, Multi-racial color fundus images	Sensitivity 97.0%,Specificity 91.4%; Sensitivity 92.5%,Specificity 98.5%;
Abràmoffff et al.[19]	CNN	900 subjects enrolled in primary care clinics	Sensitivity 87.2%,Specificity 90.7%, Accuracy 96.1%.
Wang et al.[20]	DenseNet	EyePACS	Accuracy R0:0.92,R1:0.70, R2:0.64,R3:0.67,R4:0.69.
Zhou et al.[21]	Multi-grid multi-task CNN	EyePACS	Kappa 0.841
Doshi et al.[22]	5-layer CNN	EyePACS	Kappa 0.386

3.Method

3.1 Research Status of Deep Learning Methods

Based on the previous deep learning methods, the Residual Attention Network

proposed in this paper is mainly comprises of an encoder, a residual attention module, and dilated convolution.

3.2 Encoder

The primary function of the encoder is to extract image features with high-level semantic information. Generally, the deeper the network, the stronger the ability to extract features. But when the network increases to a certain depth, the problem of gradient disappearance will occur, which leads to the degradation of network performance. ResNet[23] solves this problem through residual connection, which can make the network deeper, and its ability to extract features is more stronger. It is a structure designed based on VGG[24]. The biggest part is adding a layer jump connection structure to achieve residual learning and increase identity mapping, making the depth of the network play a role.

From an intuitive perspective, the residual learning needs less content, and the learning difficulty is low. The residual unit can be expressed as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{3-1}$$

$$x_{l+1} = f(y_l) \tag{3-2}$$

The learning characteristics from shallow l to deep L are expressed as:

(3-3)

According to the chain rule, the gradient of the reverse process can be expressed as:

(3-4)

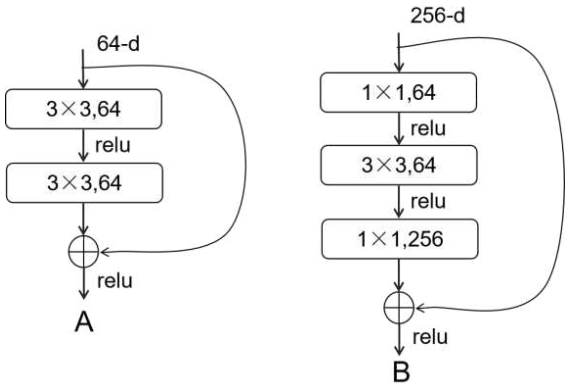

Figure 3.1 Examples of two-layer jumper connection methods (Figure A has the same channel number, Figure B has the different channel number)

As shown in Figure 3.1, Figure A corresponds to a shallow network, using identity mapping; Figure B corresponds to a deep network, using identity mapping when the input and output dimensions are consistent, and linear mapping when they are not. These two structures are aimed at ResNet18 / 34 (Figure A) and ResNet50 / 101 / 152 (Figure B).

3.3 Residual Attention Module:

The attention mechanism in computer vision is an imitation of human visual attention actually. The principle is that human brain can find the target area, and give more attention to the area, while assigning to the surrounding unimportant areas less attention, so as to obtain more useful information and suppress other useless information. In traditional image processing methods: salience detection, image feature extraction, and sliding window methods can also be regarded as attention mechanisms. The attention mechanism in deep learning also mainly includes two parts: learning weight distribution (different parts of the input image or feature map have different weights), task focus (divide the task, design different sub-networks, and focus to different subtasks, redistribute the learning ability of the network).

As shown in Figure 3.2, the attention guided module (AGM) is composed of two 1x1 convolution layers with different activation functions in the adaptive average pooling layer. The specific operation is as follows: First, the input feature map passes through an adaptive average pooling layer, and the output feature map dimension is $R^{1\times1\times c}$; then, after a 1x1 convolution layer with ReLU activation function, the output feature map dimension is $R^{1\times1\times c/r}$, and the number of channels It is reduced from C to $\mathrm{C/r}$; then, after a 1x1 convolution layer with sigmoid activation function, the number of channels is expanded from $\mathrm{C/r}$ to C, and a channel descriptor with dimension $R^{1\times1\times c}$ is obtained to recalibrate the original feature map. Among them, the hyper-parameter r can control the calculation amount of the AGM, which is set to 16 in the experiment. Finally, by multiplying the obtained channel descriptor and the input feature map, the recalibration of the feature map can be completed, and the importance of each channel can be recalibrated by integrating global information. The importance of different channels is different, which highlights important information

and suppresses background information.

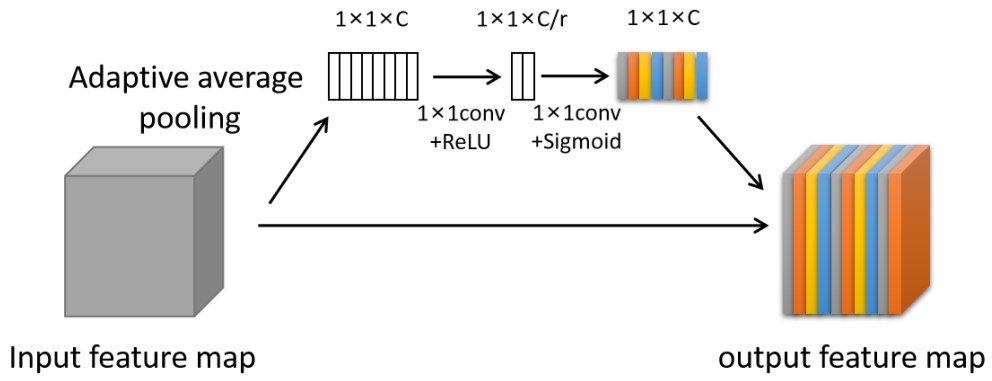

Figure 3.2 Attention guided module

In this DR classification experiment, the areas such as hard exudates, cotton velvet spots, bleeding films, and microaneurysms in the fundus image are the areas of focus[25]. The methods in the neural network can increase the abnormal area information of the lesion and suppress other background information, which can improve the accuracy of the model in the DR classification task.

Residual attention module structure, as shown in Figure 3.3, based on ResNet, the method of stacking attention structure to change the attention of features, as the network deepens, the attention mechanism module will make adaptive changes[26]. In each attention mechanism module, upsampling and downsampling structures are added.

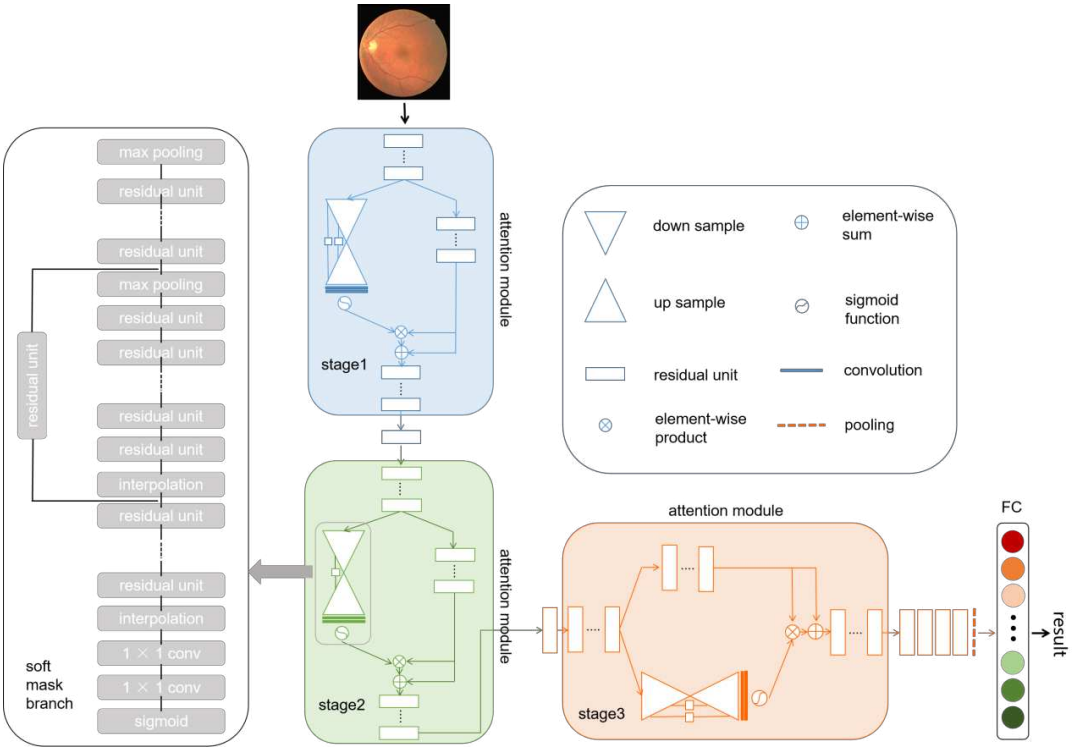

Figure 3.3 RAN structure

The core of the RAN idea is the attention mechanism. In the ordinary network, side branches are added. The side branches gradually extract high-level features and increase the receptive field of the model through a series of convolution and pooling operations. The corresponding activation position of the high-level features can reflect the area of attention, then up-sampling this feature map with attention features to make its size return to the size of the original feature map, the attention is corresponding to each position of the original image. Performing element-wise product operations with the original feature map is equivalent to a weighted action that enhances meaningful features and suppresses meaningless information.

Each attention mechanism module is divided into two branches, as Figure 3.4 shows, the soft mask branch (attention mechanism branch) and the trunk branch (original branch). The formula of the attention mechanism is:

$$H_{i,c}(x) = M_{i,c}(x) \times T_{i,c}(x) \tag{3-5}$$

T represents the main branch and M represents the mask branch. The mask branch uses several maximum pooling to increase the receptive field. After reaching the minimum resolution, a symmetric network structure is used to amplify the features back.

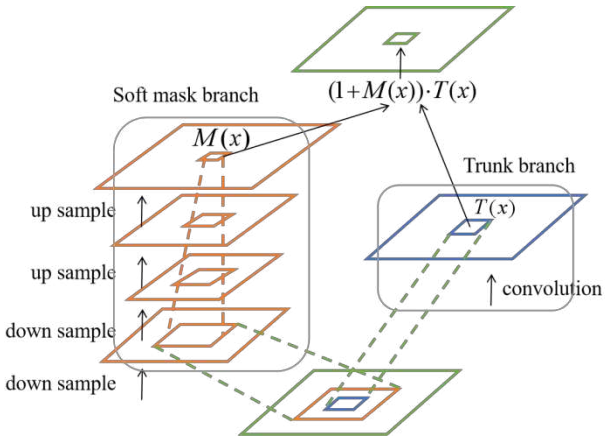

Figure 3.4 Soft mask branch and trunk branch

With the deepening of module stacking, different levels of attention information can be extracted from top to bottom, and the attention perception function from different modules will change adaptively. The added attention residual learning structure can train very deep residual attention networks, easily extended to hundreds of layers. Each attention module in the stack can be replaced with other structures (such as residual modules and inception modules), and can be easily connected to other networks to achieve a plug-and-play effect. By stacking this residual attention structure, the advantages of residual learning and attention mechanism can be thoroughly combined to achieve better results.

3.4 : Dilated Convolution Module[27] :

In order to expand the receptive field, this article also introduces a cavity convolution module. Deep features have high-level semantic information but lost resolution; shallow features have high-resolution, but the semantic level is low. The dilated convolution can expand the receptive field of the network without reducing the resolution in the case of 2 dimensions. The dilated convolution can be expressed as:

$$y_{[i]} = \sum_{k=1}^{K} x_{[i+d\cdot k]} \cdot w_{[k]} \tag{3-6}$$

In the formula, $y_{[i]}$ represents the output feature map, $x_{[i]}$ represents the input feature map, d represents the dilation rate, $W_{[k]}$ represents the $k-th$ parameter of the convolution kernel, and K represents the size of the convolution kernel.

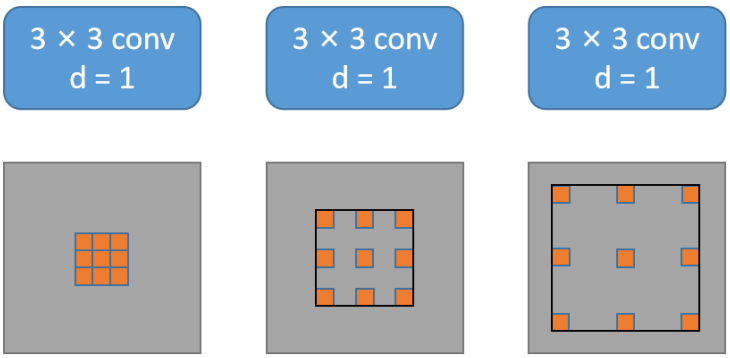

Figure 3.5 Dilated convolution

As shown in Figure 3.5, the dilated convolution is equivalent to filling d-1 dilations between adjacent convolution kernel parameters. When the dilation rate d=1,

the dilated convolution degenerates into a standard convolution, the larger d, the larger the receptive field of the convolution kernel. In this article, 1x1 standard convolution, 3x3 dilated convolution with dilation rate d=2, 3x3 dilated convolution with dilation rate d=3, 3x3 dilated convolution with dilation rate d=5, and global average pooling is used to extract features. Five levels of image information are extracted. The specific process of using global average pooling to extract features, is to use an adaptive average pooling layer to generate a 1x1x512 dimension feature map first. Second, use 1x1 convolution to change the number of channels to 256, and then use the bilinear interpolation algorithm to expand its size to 14x14. Third, the extracted feature maps of 5 levels are spliced with the original feature maps to obtain a 14x14x1792 dimension feature map, and finally use 1x1 convolution to change the number of channels to 512. After each convolution operation, there is a batch normalization layer (BN) and a ReLU activation function. Before each dilated convolution extracts features, the feature map perform a padding operation to ensure that the resolution of the feature map before and after does not change.

3.5 Loss Function

The loss function in the neural network is used to measure the gap between the predicted value obtained by the model and the actual value of the data, and it is also a standard used to measure the generalization ability of the model. The smaller the loss function, the better the performance of the model, and the loss function used by different models are generally different.

3.5.1 Cross Entropy[28]

In this experiment, the classification module performs the main task of DR classification. The goal of the classification task is to predict the label category of each input image. The most commonly used loss function is cross entropy. Cross entropy is also known as log-likelihood loss, logarithmic loss, and is also called logistic loss in the two-class classification. To describe the difference in probability distribution, the formula is:

（3-7）

y_i^c represents the original image label，$\hat{y}_i^c$ is the classifier predicting similar

values. Simultaneously, θ_{cls} represents the weight value in the classification module. Suppose there are m sets of training samples $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, where input features $x^{(i)} \in R^{n+1}$, the corresponding category is $y^{(i)} \in \{1,2,\ldots,k\}$, then the cross entropy is defined as follows:

（3-8）

$1\{\cdot\}$ is the indicator function, $\mathrm{p}(\mathrm{y}^{(\mathrm{i})} = \mathrm{j}|\mathrm{x}^{(\mathrm{i})}; \theta)$ represents a given sample $x^{(i)}$ and when model parameters is θ, probability of label is $y^{(i)}$. In this experiment, structural risk is added on the basis of cross entropy for optimization. Structural risk is based on empirical risk plus punishment items.

（3-9）

$\mathrm{J}(\theta)$ is the complexity of the model, the more complex the model f, the greater the penalty term model, $\lambda \geq 0$ represents weights penalties, regularization terms are L1 and L2 regularization.

3.5.2 Focal Loss[29]:

Since the imbalance problem generally exists in the DR datasets, focal loss, designed to solve the imbalance problem is introduced in this experiment. It is modified on cross entropy, and multiplies the original cross entropy by an index that weakens the contribution of the easily detectable object to the model training. So that focal loss successfully optimizes the imbalance problem between positive and negative samples, and relieves the problem that object detection loss are easily affected by a large number of negative samples. Focal Loss is defined as:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma} log(p_t) \quad （3\text{-}10）$$

γ is the focus parameter, $\gamma \geq 0$. $(1-p_t)^{\gamma}$ is called modulating factor, the purpose of adding modulation coefficient is to reduce the weight of samples that are easy to classified, so that the model focused more on the samples that are difficult to classified during training.

Focal loss has two important properties:①When a sample is wrong, p_t is very small, then the modulation factor $(1-p_t)$ is close to 1, the loss is not affected; when

$p_t \to 1$, the factor $(1-p_t)$ is close to 0, then the weight of the better sample is reduced. Therefore, the modulation coefficient tends to 1, which means that there is no significant change from the original loss. ②When $\gamma = 0$, focal loss can be written as cross entropy, and as γ increases, the modulation coefficient also increases.

3.6 Transfer Learning[30]

Transfer learning is a method of machine learning, which is to transplant the model obtained from one task training to the training of other tasks. Affected by transfer learning, in the case of insufficient training data, by loading the pre-trained EfficientNet weights on the ImageNet dataset, the model has a better weight initialization before starting to optimize the gradient, so as to train your own model. Considering the huge difference between the fundus image dataset and the ImageNet dataset, the training of the network layer during the experiment is restarted from each layer.

4. Materials and Approach

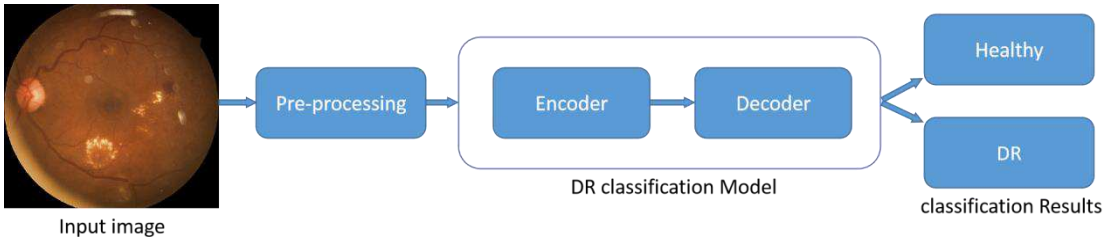

Figure 4.1 The workflow of the proposed work for classification of diabetic retinopathy

4.1 Datasets

In order to verify the effectiveness of RAN, comparative experiments were carried out on Kaggle, IDRid, and Messidor datasets, Figure 4.1 shows the workflow.

4.1.1 Kaggle Dataset[31]

The training set contains 35,125 fundus images released by the California Medical Foundation from eye-PACS users, including level 0 25809 (74%), level 1 2443 (7%), level 2 5292 (15%), level 3 873 (2%), level 4 708 (2%). The images in the dataset come from different models and types of cameras, which may affect the visual appearance and resolution of the images. Some images in the dataset contain artifacts, blurry focus, underexposure or overexposure, which may adversely affect the experimental results. In addition, due to the excessive number of normal fundus images in this dataset, 40% of the normal images were selected for training and testing during the second classification experiment, and only 20% of the normal images were selected for training during the five-class classification experiment and test.

4.1.2 Messidor Dataset[32]

The Messidor dataset is a research project funded by the French Ministry of Defense in the TECHNO-VISION program in 2004 to promote research on the computer-aided diagnosis of diabetic retinopathy. It consists of 1200 fundus images from three ophthalmology hospitals, 800 of which are images obtained after pupil dilation. According to the presence or absence of hard exudates, microaneurysms, hemorrhage, and neovascularization, each picture is marked with a DR lesion grade of 0-3, each picture also has a DME lesion grade of 0-2, Figure 4.2. Table 4.1 lists the number distribution. The image sizes in the dataset are 1440×960, 2240×1488, and 2304×1536, respectively, in tif format.

Table 4.1 Number distribution of DR and DME in Messidor dataset

Diseases	Lesion Grade	Picture Quantity (1200)	Ratio
DR	0	546	46%
	1	153	13%
	2	247	21%
	3	254	21%
DME	0	974	81%
	1	75	6%
	2	151	13%

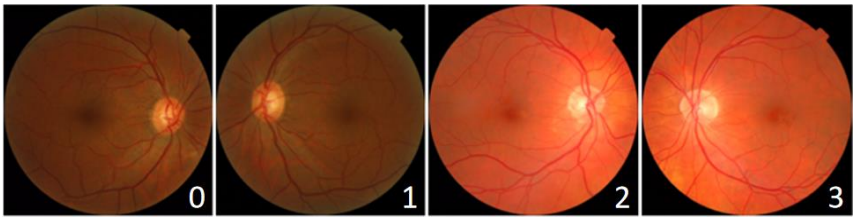

Figure 4.2 Fundus images of DR level 0-3 in the Messidor dataset from left to right

4.1.3 IDRid Dataset[33]

The dataset was published by the IEEE International Biomedical Imaging Symposium (ISBI-2018) Diabetic Retinopathy Segmentation and Classification Challenge, the purpose of this challenge is to evaluate algorithms for the automatic detection and grading of DR and DME using fundus images. Data includes: ①Lesion segmentation: DR-related retinal lesions are segmented into microaneurysms, hemorrhage, hard and soft exudates. ②Disease classification: classify fundus images according to the severity of DR and DME. ③Optic disc and fovea detection: automatic positioning of the central coordinates of the optic disc and fovea and segmentation of the optic disc. In this experiment, only disease classification data was used, including 413 pictures in the training set and 103 pictures in the test set. All pictures are 4288×2848, in jpg format. Table 4.2 shows the number and proportion distribution.

Table 4.2 Number and distribution ratio of DR and DME training sets in IDRid dataset

Diseases	Lesion Grade	Training Set (413)	Test Set (103)	Ratio
DR	0	134	34	33%
	1	20	5	5%
	2	136	32	33%
	3	74	19	20%
	4	49	13	12%
DME	0	177	45	43%
	1	41	10	10%
	2	195	48	47%

Because the amount of abnormal pictures in the Messidor dataset and IDRid dataset is too small, it is more meaningful for clinical application to do two-class classifications. It can also be seen from the above dataset that the most prominent feature of medical images is the imbalance distribution of data, that is, the number of samples in normal images is much higher than that of abnormal images, and the

amount of grading data with the severity of the disease is getting less and less. To solve this problem, the most commonly used method is data enhancement, to expand the lesion sample. In addition, improving the loss function or improving the network structure is also a widely used optimization method.

In the above three datasets, we randomly selected 60%, 15%, and 25% of the images in each dataset as the training set, validation set and test set.

4.2 Image Preprocessing

Since all the widely used DR public datasets have the problem of severe imbalance in data distribution, and image preprocessing is used in this experiment to increase the amount of data. The purpose of image enhancement is to process the acquired images so that the features of interest have better contrast and visibility. By making a series of random changes to the training image to produce similar but different training samples, the amount of training set is expanded. The robustness and generalization ability of the model trained by image enhancement can significantly improved. This method does neither reduce the capacity of the network, nor does it increase the computational complexity and the number of parameters. It is an implicit regularization method and is widely used in the current landing of medical image AI products. However, most of the image enhancement methods have a certain randomness, which may reduce the accuracy of the model while enhancing the robustness.

4.2.1 Data Augmentation

Commonly used image-enhancing methods, translation, rotation, cropping, scaling, noise addition, affine transformation, etc., usually do not change the type of object, are the earliest and most widely used type of image-enhancing method. Another way is to change the color. We can change the color of the image from four areas: brightness, contrast, saturation, and tone. In practical applications, multiple image-enhancing methods are usually superimposed, as shown in Figure 4.3.

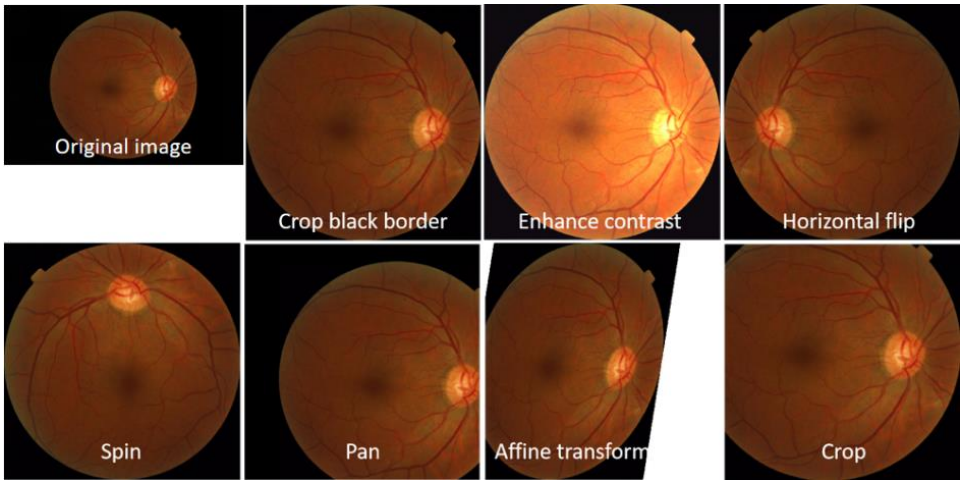

Figure 4.3 Messidor dataset enhanced results

4.2.2 CLAHE

Contrast Limited Adaptive Histogram Equalization (CLAHE), mainly limits the contrast in each small area by limiting the degree of contrast improvement of the adaptive histogram equalization, crops the histogram with a defined threshold, and limits the slope of the transform function. By cropping the histogram obtained from the statistics in the sub-blocks, the amplitude is within a certain interval, and the entire interval is evenly distributed, while also ensuring that the total area of the histogram remains unchanged. Figure 4.4 lists the changes of the image and its histogram after the histogram equalization and CLAHE processing of the fundus image in the Kaggle dataset.

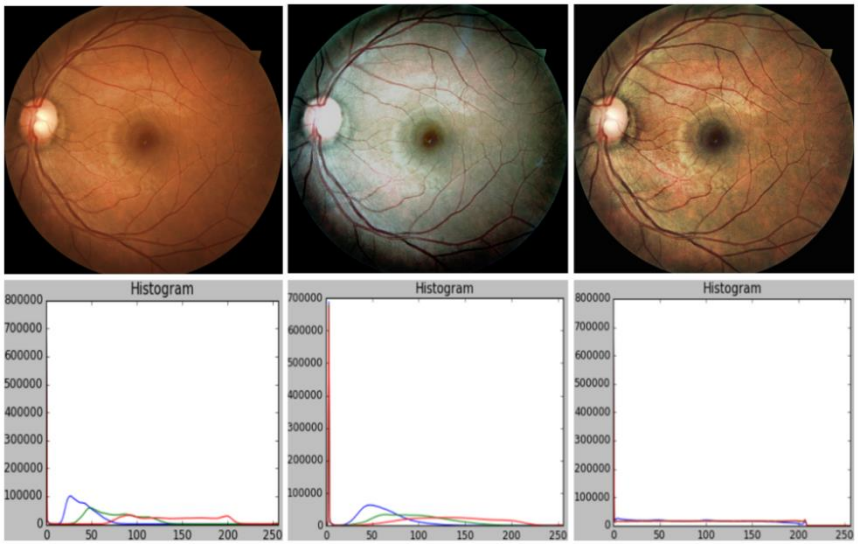

Figure 4.4 The three columns from left to right are the original image in the Kaggle dataset, after CLAHE processing, after equalization processing and its corresponding histogram

4.2.3 Normalizing

In order to reduce the difference between the different images of the dataset, before sending the image to the network for training, first standardize each image, specifically, Figure 4.5:

$$I'_{(i,j,k)} = \frac{I_{(i,j,k)} - m_k}{\sigma_k}$$

(4-1)

Where $I_{(i,j,k)}$ is the input image, $I'_{(i,j,k)}$ is the normalized image, iand j is the coordinates of the pixel points, k represents the three channels of the image (blue, green and red), m_k represents the average value of the $k-th$ channel pixel value, and σ_k represents the standard deviation of the $k-th$ channel pixel value.

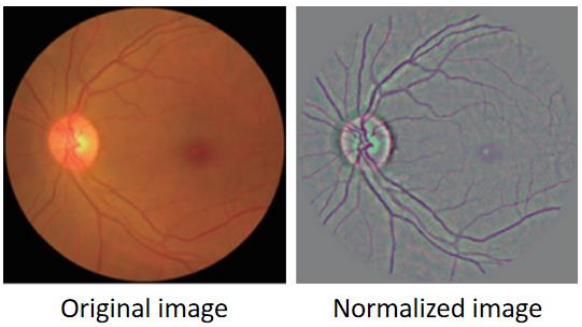

Figure 4.5 DR image before and after normalizing

In addition, due to limited computer performance, the image is first scaled to 224x224 pixels and then sent to the network for training and testing.

4.3 Implementation Details

The experiment uses the PyTorch deep learning framework and OpenCV image processing library, implemented on Ubuntu16.04 operating system, GeForce GTX 2080Ti graphics card, Adam optimizer initial learning rate is 0.001, the batchsize training phase is set to 8, the testing phase is set to 1, and a total of 60 epochs are trained. The test set is tested after every epoch of training, and we only output models and results with the highest sensitivity and accuracy values.

Table 4.3 Experimental environment

CPU	i7-9700k 8 cores3.6G
RAM	Kingston 16G DDR4 2666
Main Hard Drive	Samsung 256G M.2

GPU	RTX 2080ti 11GB
Operating System	Ubuntu 16.04
Software Environment	Python 3.6.2 PyTorch 1.0.0 Scikit-Learn 0.19.2

4.4 Evaluation Index

In this experiment, the relationship between the model prediction result and the true label of the data is evaluated by the following criteria (as shown in Figure 4.6): True Positive (TP), the number of true positive samples predicted as positive; False Negative (FN), The number of true positive samples predicted to be negative; False Positive (FP), the number of true negative samples predicted to be positive; True Negative (TN), the number of true negative samples predicted to be negative. We also adopted accuracy (ACC), sensitivity (SE), specificity (SP), receiver operating curve (ROC), and area under curve (AUC) to evaluate the experimental results.

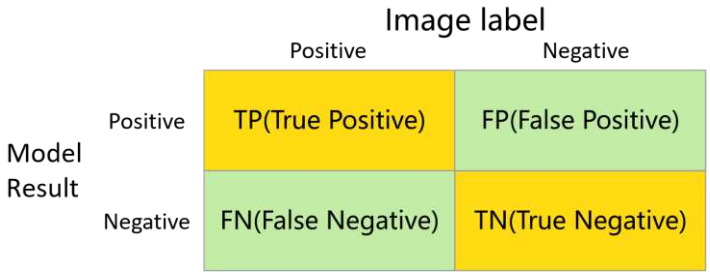

Figure 4.6 Evaluation criteria of DR grading experiment

SE represents the proportion of true positive samples that are predicted to be positive, and SP represents the proportion of true negative samples that are predicted to be negative. The calculation formula for the SE and SP is:

（4-2）

The higher the SE, the greater the probability of a DR image being diagnosed, the higher the SP, the greater the probability that a normal image is predicted to be normal. In clinical applications, the missed diagnosis has a greater adverse effect on patients, so the SE in DR classification is more significant. ACC represents the probability of correct classification of all samples, the calculation formula is as follows:

（4-3）

Taking multiple sets of SE values as the ordinate and SP values as the abscissa constitute the ROC curve, which can comprehensively evaluate the performance difference of multiple classifiers. AUC represents the area under the ROC curve, and the AUC also can express the model classification ability more intuitively. The larger the AUC, the better the classification performance. In the DR five-category experiment, the Kappa coefficient was also added as an evaluation criterion. The Kappa coefficient is calculated on the basis of the confusion matrix. It is a method for evaluating the consistency of the model results and can also be used to evaluate the accuracy and performance of the classifier. The formula for the Kappa coefficient is calculated as follows:

$$kappa = \frac{q_0 - q_i}{1 - q_i} \tag{4-4}$$

$$q_i = \frac{a_1 \times b_1 + a_2 \times b_2 + \ldots + a_i \times b_i}{n \times n} \tag{4-5}$$

q_i is the number of correctly classified samples in each category divided by the total number of samples, a_i is the number of samples of type i, b_i is the number of samples predicted to be the $\mathrm{i\_th}$ category.

5. Results and Discussion

In this paper, on the three DR datasets of Kaggle, Messidor, and IDRid, the commonly used deep learning methods and our proposed RAN are used respectively, and the loss function uses cross entropy and focal loss to perform classification and diagnosis experiments for DR and DME, then compared and analyzed them. The results are as follows.

5.1 Kaggle Results

It can be seen from Table 5.1 that in the Kaggle dataset, the specificity, sensitivity and AUC of the RAN proposed in this paper for DR two-class classification reached 0.894, 0.930 and 0.917, respectively, and the accuracy reached

0.892, which was 4.6%, 3.7%, 9.4% and 5.6% higher than VGG-16. RAN has reached an excellent level in the accuracy of DR classification.

Table 5.1 Two-class classification results of different models on the Kaggle dataset

Method	Accuracy	Specificity	Sensitivity	AUC
VGG-16	0.836	0.848	0.893	0.823
ResNet-50	0.844	0.862	0.907	0.876
ResNeXt-50[34]	0.866	0.867	0.919	0.892
DenseNet-121[35]	0.862	0.877	0.905	0.874
SE-ResNet-50[36]	0.889	0.889	0.924	0.896
DPN-92[37]	0.868	0.876	0.918	0.878
EfficientNet-b0[38]	0.886	0.882	0.925	0.903
RAN	**0.892**	**0.894**	**0.930**	**0.917**

It can be seen from Table 5.2 that in the Kaggle dataset, the accuracy of RAN for DR five-class classification reaches 0.815, which is 5.3% higher than that of VGG-16. The Kappa score reached 0.865, which was higher than the 0.829 achieved in the DR classification competition held on the Kaggle website in 2015[31], and the performance was improved.

Table 5.2 Five-class classification results of different models on Kaggle dataset

Method	Accuracy	Kappa Score
VGG-16	0.762	0.720
ResNet-50	0.782	0.781
ResNeXt-50	0.811	0.812
DenseNet-121	0.803	0.856
SE-ResNet-50	0.808	0.834
DPN-92	0.802	0.827
EfficientNet-b0	0.810	0.835
RAN	**0.815**	**0.865**

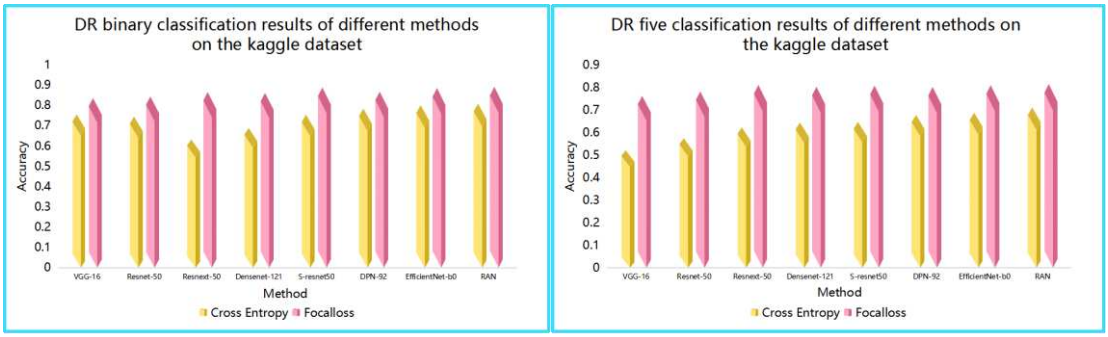

Figure 5.1 Comparison of the classification results of different models

with different loss functions on the Kaggle dataset

In order to improve the performance of the deep learning model and alleviate the problem of the small amount of data in the Messidor and IDRid datasets, this paper transfers the RAN model trained on the Kaggle dataset to the experiments on the Messidor and IDRid datasets.

5.2 Messidor Results

It can be seen from Table 5.3 that in the Messidor dataset, the specificity and sensitivity of the RAN proposed in this paper for DR two-class classification reached 0.887 and 0.931, and the accuracy reached 0.898, which was 3.7%, 4.9% and 7.5% higher than VGG-16. The AUC result of EfficientNet is 0.912, which is 0.5% higher than RAN.

Table 5.3 The results of DR two-class classification of different models on the Messidor dataset

Method	Accuracy	Specificity	Sensitivity	AUC
VGG-16	0.823	0.850	0.882	0.830
ResNet-50	0.861	0.875	0.892	0.887
ResNeXt-50	0.868	0.864	0.904	0.893
DenseNet-121	0.854	0.878	0.895	0.864
SE-ResNet-50	0.858	0.884	0.914	0.894
DPN-92	0.864	0.876	0.908	0.869
EfficientNet-b0	0.884	0.881	0.923	**0.912**
RAN	**0.898**	**0.887**	**0.931**	0.907

In the Messidor dataset, Table 5.4, the accuracy, sensitivity and AUC of RAN for DME two-class classification reached 0.881, 0.912, and 0.882, respectively, which were 4.2%, 4% and 4.4% higher than VGG-16. The specificity result of EfficientNet is 0.862, which is 0.4% higher than RAN, also has a good performance.

Table 5.4 DME two-class classification results of different models on the Messidor dataset

Method	Accuracy	Specificity	Sensitivity	AUC
VGG-16	0.839	0.824	0.872	0.838
ResNet-50	0.852	0.833	0.876	0.851
ResNeXt-50	0.861	0.833	0.883	0.858
DenseNet-121	0.852	0.836	0.882	0.844
SE-ResNet-50	0.866	0.843	0.897	0.863

DPN-92	0.853	0.832	0.889	0.858
EfficientNet-b0	0.871	**0.862**	0.904	0.880
RAN	**0.881**	0.858	**0.912**	**0.882**

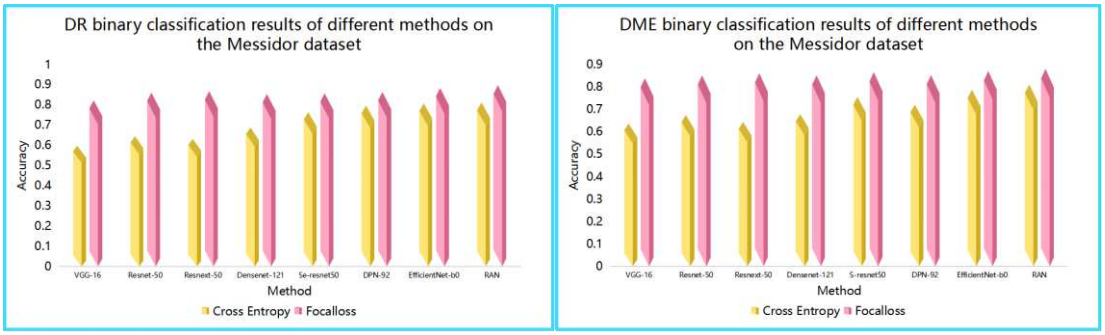

Figure 5.2 Comparison of the classification results of different models with different loss functions on the Messidor dataset

5.3 IDRid Results

Table 5.5 DR two-class classification results of different models on IDRid dataset

Method	Accuracy	Specificity	Sensitivity	AUC
VGG-16	0.678	0.684	0.724	0.667
ResNet-50	0.693	0.705	0.734	0.684
ResNeXt-50	0.702	0.714	0.748	0.695
DenseNet-121	0.694	0.706	0.736	0.685
SE-ResNet-50	0.701	0.716	0.745	0.711
DPN-92	0.692	0.708	0.734	0.702
EfficientNet-b0	0.704	0.715	0.747	0.715
RAN	**0.715**	**0.728**	**0.752**	**0.726**

Table 5.6 DME two-class classification results of different models on IDRid dataset

Method	Accuracy	Specificity	Sensitivity	AUC
VGG-16	0.702	0.714	0.748	0.693
ResNet-50	0.725	0.736	0.768	0.714
ResNeXt-50	0.746	0.749	0.785	0.738
DenseNet-121	0.735	0.740	0.779	0.737
SE-ResNet-50	0.742	0.751	0.792	0.736
DPN-92	0.728	0.742	0.787	0.718
EfficientNet-b0	0.756	0.763	0.807	0.739
RAN	**0.785**	**0.796**	**0.813**	**0.772**

It can be seen from Table 5.5 that in the IDRid dataset, the accuracy, specificity, sensitivity and AUC of RAN for DR two-class classification reached 0.715, 0.728,

0.752 and 0.726. From Table 5.6 we can see that the accuracy, specificity, sensitivity, and AUC of the RAN for the IDRid dataset DME two-class classification reached 0.785, 0.796, 0.813 and 0.772. Compared with other methods, the performance of RAN in each index in the IDRid dataset has an improvement of about 1 to 10 points.

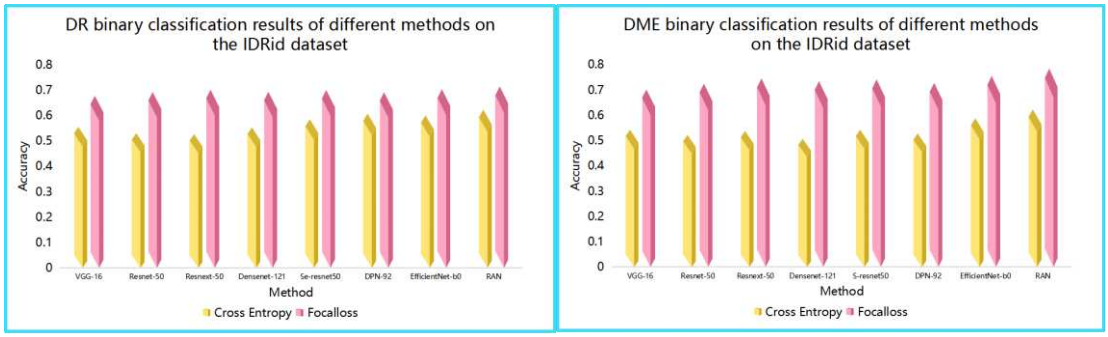

Figure 5.3 Comparison of the classification results of different models with different loss functions on the Messidor dataset

As can be seen from Figure 5.1-5.3, because of the imbalance problem in the DR datasets, in each classification task, using focal loss as the loss function is more suitable than cross entropy as the loss function, accuracy has been greatly improved.

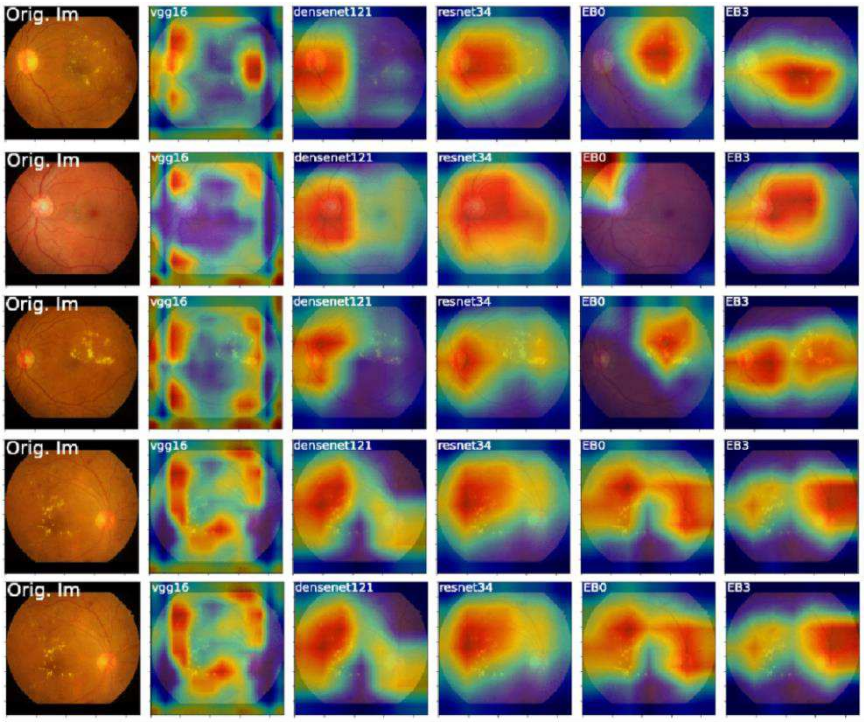

Figure 5.4 Visualization of Grad-CAM[39] DR classification

In addition, we also used Grad-CAM to visualize the attention heat map during

the fundus image DR classification process. As shown in Figure 5.4, we can clearly see that the optimization method is more focused on abnormal parts than the basic neural network structure.

The above experimental results show the intense competitiveness of CNN in clinical diagnostic applications, and RAN has achieved good results in completing the DR classification task. The image-enhancing method used in this experiment can make the amount of data in each classification of DR reach a relatively balanced state, and the loss function optimization method has also achieved satisfactory results in alleviating the problem of data imbalance. The model added an attention mechanism, which can pay more attention to the features in the fine-grained image during classification, and play an active auxiliary role in the feature extraction of the network. Using optimization methods such as dilated convolution can also improve the results of the neural network. In short, using our RAN can enhance the accuracy of DR classification and diagnosis on most fundus images.

6. Conclusion

This paper proposes a classification algorithm, Residual Attention Network (RAN), combining attention mechanism and dilated convolution for diaetic retinopathy (DR) detection. The classification effect of the model is verified on Kaggle, Messidor and IDRid competition data. Since the imbalance between data categories will lead to overfitting during model training, data augmentation and focal loss are used. Aiming at the problem of minor differences between DR categories, we performed a series of preprocessing on the original retinal image to make the bleeding and exudation in the fundus image more obvious. Then, an attention mechanism is added to the network to extract features of fine-grained images, so that the network can better distinguish the differences between the types of lesions, and we also used dilated convolution in the network to increase the receptive field. Through this combination of residual network designed based on ResNet, attention mechanism and dilated convolution, the accuracy of the classification task of diabetic retinopathy can be improved. However, the increase in accuracy of this method is not significant enough. Therefore, in the future work, we will integrate the prior knowledge of age, blood glucose, blood pressure, intraocular pressure, and past history into the DR classification model to integrate more information related to disease and effectively

improve the diagnosis effect. In addition, multi-task experiments will mutually promote the improvement of experimental results. How to integrate the results of optic disc, macular detection, and blood vessel segmentation into the DR classification model will also be the focus of future work. It is the common aspiration of algorithm engineers and clinicians to build a robust and accurate deep learning model for DR screening. This desire cannot be achieved without the joint efforts and cooperation of both parties.

Funding. National Natural Science Foundation of China(81971708); National key R&D project(2018YFC1314900,2018YFC1314902).

Disclosures. The authors declare that there are no conflicts of interest related to this article.

References

[1] Zhang Xiaofeng, Huang Shuren. Diagnosis and treatment of diabetic retinopathy[J].Chinese Medical Journal,2003(07):12-15.

[2] Zhang Chengfen. Fundus Medicine (Second Edition)[C]. 2010:149.

[3] Yau J W, Rogers S L, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy[J]. Diabetes Care, 2012, 35(3):556-64.

[4] Pratt H, Coenen F, Broadbent D M, et al. Convolutional neural networks for diabetic retinopathy[J]. Procedia Computer Science, 2016: 200-205.

[5] Chandrakumar T, Kathirvel R. Classifying Diabetic Retinopathy using Deep Learning Architecture[J]. International Journal of Engineering & Technical Research, 2016,5(6).

[6] Rahim S S , Palade V , Shuttleworth J , et al. Automatic screening and classification of diabetic retinopathy and maculopathy using fuzzy image processing[J]. Brain Informatics, 2016.

[7] Eftekhari N, Pourreza H R, Masoudi M, et al. Microaneurysm detection in fundus images using a two-step convolutional neural network[J]. BioMedical Engineering OnLine, 2019, 18(1).

[8] Seth S, Agarwal B. A hybrid deep learning model for detecting diabetic retinopathy[J]. Journal of Statistics and Management Systems, 2018, 21(4): 569-574.

[9] Dutta S, Manideep B C, Basha S M, et al. Classification of diabetic retinopathy images by using deep learning models[J]. International Journal of Grid and Distributed Computing, 2018, 11(1): 99-106.

[10] Benzamin A, Chakraborty C. Detection of hard exudates in retinal fundus images using deep learning[C]. International Conference on Informatics Electronics and Vision, 2018: 465-469.

[11] Adem K. Exudate detection for diabetic retinopathy with circular hough transformation and convolutional neural networks[J]. Expert Systems With Applications, 2018: 289-295.

[12] Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs[J]. Ophthalmology, 2018:S0161642017335650.

[13] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. JAMA, 2016, 316(22): 2402-2410.

[14] Li X, Pang T, Xiong B, et al. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification[C]. International Congress on Image and Signal Processing, 2017: 1-11.

[15] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep Learning[J]. Ophthalmology, 2017:S0161642016317742.

[16] Abramoff M D, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning[J]. Investigative Ophthalmology & Visual Science, 2016, 57(13): 5200-5206.

[17] Ting D S W , Cheung Y L , Lim G , et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes[J]. JAMA, 2017, 318(22):2211.

[18] Li Z, Keel S, Liu C, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs[J]. Diabetes Care, 2018, 41(12): 2509-2516.

[19] Abràmoff Michael D, Lavin Philip T, Birch Michele, Shah Nilay, Folk James C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices.[J]. NPJ digital medicine,2018,1.

[20] Wang Y, Wang G A, Fan W, Li J. A deep learning based pipeline for image grading of diabetic retinopathy[J]. Lecture Notes in Computer Science, vol 10983. Springer, Cham.

2018.
[21] Zhou K, Gu Z, Liu W, et al. Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading[J]. 2018.
[22] Doshi D, Shenoy A, Sidhpura D, et al. Diabetic retinopathy detection using deep convolutional neural networks[C]. International Conference on Computing Analytics and Security Trends, 2016: 261-266.
[23] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Computer Vision and Pattern Recognition, 2016: 770-778.
[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. Computer Vision and Pattern Recognition, 2014.
[25] Wang Z , Yin Y , Shi J , et al. Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection[J]. Springer, Cham, 2017.
[26] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification[C]. Computer Vision and Pattern Recognition, 2017: 6450-6458.
[27] Yu F , Koltun V , Funkhouser T . Dilated Residual Networks[J]. IEEE Computer Society, 2017.]
[28] Schmidhuber J. Deep learning in neural networks[J]. Neural Networks, 2015: 85-117.
[29] Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. International Conference on Computer Vision, 2017: 2999-3007.
[30] Pan, Sinno, Jialin. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2010.
[31] https://www.kaggle.com/c/diabetic-retinopathy-detection/
[32] http://www.adcis.net/en/third-party/messidor/
[33] https://idrid.grand-challenge.org/leaderboard/
[34] Xie S, Girshick R, Dollar P, et al. Aggregated residual transformations for deep neural networks[C]. Computer Vision and Pattern Recognition, 2017: 5987-5995.
[35] Huang G, Liu Z, Der Maaten L V, et al. Densely connected convolutional networks[C]. Computer Vision and Pattern Recognition, 2017: 2261-2269.
[36] Hu J, Shen L, Sun G, et al. Squeeze-and-Excitation Networks[J]. arXiv: Computer Vision and Pattern Recognition, 2017.
[37] Chen Y, Li J, Xiao H, et al. Dual Path Networks[J]. arXiv: Computer Vision and Pattern Recognition, 2017.
[38] Tan M,Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[J]. 2019.
[39] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. International Conference on Computer Vision, 2017: 618-626.