scientific reports

OPEN

Diabetic retinopathy classification using a multi-attention residual refinement architecture

Zijian Wang[1,2], Yi Wang[1], Chun Ma[1], Xuan Bao[1] & Ya Li[1]✉

Diabetic Retinopathy (DR) is a complication caused by diabetes that can destroy the retina, leading to blurred vision and even blindness. We propose a multi-attention residual refinement architecture that enhances conventional CNN performance through three strategic modifications: class-specific multi-attention for diagnostic feature weighting, space-to-depth preprocessing for improved spatial information preservation, and Squeeze-and-Excitation blocks for enhanced representational capacity. Our framework demonstrates universal applicability across different CNN architectures (ResNet, DenseNet, EfficientNet, MobileNet), consistently achieving 2-5% performance improvements on the EyePACS dataset while maintaining computational efficiency. The attention mechanism provides interpretable visualizations that align with clinical pathological patterns, validating the model's diagnostic reasoning.

Keywords Diabetic retinopathy, Attention mechanism, Deep learning model

Diabetes is a lifelong disease in which the amount of glucose in the blood increases due to a lack of control over insulin levels[1]. It affects different body parts, such as the nerves, kidneys, heart, and retinas[1]. The retina is the light-sensitive lining at the innermost layer of the eye, which lines the posterior part of the eye[2]. The function of the retina is to process visual information and coordinate with the brain by transmitting light through nerve signals[3]. The retina receives blood nourishment through microvessels[4], and it is necessary to maintain blood glucose levels with uninterrupted blood flow. Even in pre-diabetes, high blood sugar levels can damage microvessels. A complication of diabetes may cause progressive damage (swell, leak fluids, and bleed) to the blood vessels in the retina, which is called Diabetic Retinopathy (DR)[5,6]. DR causes progressive vision loss and is one of the most common causes of permanent blindness[7]. The International Diabetes Federation estimated that about 537 million diabetic adults worldwide in 2021, and almost 30% suffer from DR[8]. Early diagnosis of DR is essential since conditions are more treatable in the early stages[9], and the progression of visual impairment can be slowed or avoided if DR is detected in time[10]. DR is a progressive eye disease classified by four stages (No DR is Class 0):

- Mild nonproliferative diabetic retinopathy (Mild NPDR, Class 1);
- Moderate nonproliferative diabetic retinopathy (Moderate NPDR, Class 2);
- Severe nonproliferative diabetic retinopathy (Severe NPDR, Class 3);
- Proliferative diabetic retinopathy (PDR, Class 4).

Mild DR is characterized by swelling of retinal blood vessels and is usually asymptomatic. During the moderate DR stage, there is a significant increase in microvascular swelling. Swelling will interfere with blood flow from blood vessels to the retina, prevent proper nourishment, and lead to blurred vision. In the severe DR stage, abnormal vascular growth can significantly reduce retinal blood flow. Late DR shows hyperplastic retinal detachment with significant retinal breaks that can lead to complete vision loss. The diagnosis of DR requires ophthalmologists to identify the presence of retinal vascular abnormalities associated with related lesions. However, as the number of DR patients increases, most areas with large populations still lack professional knowledge resources from clinicians. With the development of deep learning in recent years, many works have made considerable progress in DR automatic screening using image classification[9–12], image detection[13–15], and image segmentation[16–18].

Medical image classification is a challenging task for practical computer vision applications because medical images contain a large amount of irrelevant information. Therefore, it is necessary to utilize attention algorithms to concentrate on the critical parts of the image for classification decisions. However, the research methods in this field are complicated with heavy calculations, and the lack of intuitive explanation makes it difficult to

[1]School of Medical Information Engineering, Anhui University of Chinese Medicine, Hefei, China. [2]Hefei University of Technology, Hefei, China. ✉email: liya1985724@163.com

generalize. In order to notice the different spatial regions corresponding to various lesions more effectively, this paper uses a class-specific residual multi-attention algorithm and applies it to the current popular algorithm model. By proposing a spatial attention score, class-specific residual multi-attention generates specific features for each category, and average pooling is then used for feature fusion. The algorithm is easy to implement with small computation, can be directly fused to the fully-connected layer of deep learning models, and has multiple advantages such as solid interpretability and easy visualization. The deep learning models used in this paper all use pre-training techniques[19], and the proposed methods have been trained, validated, and tested on publicly available datasets from EyePACS[20]. The experimental results show that the proposed method has considerable performance improvement on the original basis. The methodology presented in this paper has the following contributions:

- A novel class-specific multi-attention mechanism is developed to dynamically weight spatial features according to their diagnostic relevance for each DR severity level, resulting in improved classification accuracy and model interpretability.
- An enhanced preprocessing approach utilizing space-to-depth layers is proposed to preserve spatial information during downsampling, effectively expanding the receptive field compared to conventional methods.
- Squeeze-and-Excitation blocks are strategically integrated into the residual framework to optimize channel-wise feature representations, enhancing model capacity while maintaining computational efficiency.

Related work

Attention model

The attention model is derived from the human biological system, which can focus on the most critical points when processing high-resolution images or large amounts of text information[21]. The attention mechanism has become essential in deep learning models in recent years. Several studies have shown that the attention mechanism can visually express the understanding of model learning[21]. For medical image classification, this mechanism can directly highlight the most concerned part of the model and further explain the decision-making principle of the model. Attention mechanisms can be used in various tasks such as image classification, machine translation, image segmentation, etc. It can solve the underfitting caused by the diversification of information, focus on the more valuable information in big data, and filter out the useless parts. It has recently attracted lots of awareness from researchers, and its design philosophy is gradually improved. For example, multi-attention mechanisms can focus on multiple vital parts of the images simultaneously, which is better than single-attention mechanisms[22].

Residual network refinement

The residual network[23] is a famous architecture with high training efficiency and better accuracy. Numerous recent studies have proposed architecture optimization methods based on residual neural networks[24]. These optimization methods usually vary the number of layers of residual neural networks to obtain better performance than the original model. Alternatively, the bottleneck in the original model can be redesigned to generate model variants to improve the theoretical stability boundary. The feature extraction ability can be significantly enhanced by introducing new architectures and increasing the model depth. Due to the limitation of computational resources, many researchers focus on the trade-off between computational complexity and accuracy. For example,[25] provided an explicit mechanism for establishing model dynamics and nonlinear channel dependence to simplify the learning process and enhance the ability of network representation.

Methods

This section focuses on the method of the proposed model. We first introduced the class-specific residual attention module and then proposed a multi-attention mechanism to improve the ability of small object detection (lesions detection). Then, we present the refinements of the model, such as using the Space-to-depth layer to improve the receptive field and redesigning the residual block to enhance the information flow in the model.

Class-specific multi-attention module

Medical image classification requires focused attention on discriminative regions that vary across different pathological conditions. To address this challenge, we propose a class-specific multi-attention mechanism that dynamically weights spatial features according to their relevance for each diagnostic category, enabling more precise and interpretable classification decisions.

Class-specific residual attention module

The feature matrix $\mathrm{x} \in \mathbb{R}^{d \times h \times w}$ of the image *I* is extracted using the feature extractor (convolutional neural network). *d*, *h*, and *w* represent the dimension, height, and width of the feature matrix, respectively. The feature extraction process can be represented by the function θ:

$$\mathrm{x} = \phi(I; \theta). \tag{1}$$

Assume that the dimension of the feature matrix from the function θ is 2048×7×7. Firstly, the feature matrix is decoupled into a group of the position feature matrix: $x_1, x_2, \cdots, x_{49}$ ($\mathrm{x_j} \in \mathbb{R}^{2048}$). Then, a fully-connected layer (1×1 convolution) is used as a classifier. Note that each class has its specific classifier, and the parameter of the classifier for the class *i* is $\mathrm{m}_i \in \mathbb{R}^{2048}$. The constant 49 here can be changed according to different model settings.

Define the class-specific residual attention score for class *i* and position *j*:

$$s_j^i = \frac{\exp\left(T\mathrm{x}_j^T\mathrm{m}_i\right)}{\sum_{k=1}^{49}\exp\left(T\mathrm{x}_k^T\mathrm{m}_i\right)}, \tag{2}$$

where T $(T > 0)$ is a temperature factor that controls the sensitivity of the attention score. It represents the probability of the class *i* appearing in the position *j*.

For class *i*, the weight corresponding to the position feature vector $\mathrm{x_j}$ is s_j^i. All the groups of the position feature matrix are weighted and summed to get the class-specific feature vector:

$$\mathrm{a}^i = \sum_{k=1}^{49} s_k^i \mathrm{x}_k. \tag{3}$$

In traditional average pooling, the entire group of position feature vectors is directly averaged without considering the corresponding weights of the position feature vectors:

$$\mathrm{g} = \frac{1}{49}\sum_{k=1}^{49}\mathrm{x}_k. \tag{4}$$

Since average pooling has been widely used in practice and achieved superior results, we fuse vectors $\mathrm{a}^i = \sum_{k=1}^{49} s_k^i \mathrm{x}_k$ and $\mathrm{g} = \frac{1}{49}\sum_{k=1}^{49}\mathrm{x}_k$. Finally, the class-specific residual attention f^i for the class *i* is:

$$\mathrm{f}^i = (1-\lambda)\mathrm{g} + \lambda\mathrm{a}^i, \tag{5}$$

where λ is the hyperparameter (λ=0.2). The above formulas constitute the whole class-specific residual attention module.

Multi-attention module

We introduced the multi-attention mechanism to extend the class-specific residual attention module. With the introduction of the multi-attention mechanism, there is no need to finetune temperature hyperparameters T manually. Multi-attention uses different temperature parameters to calculate and output multiple probabilities sequentially. In the experimental design of multi-attention, most proposed methods usually follow the convention that the probability value of the output is gradually controlled and reduced according to the orders, and the last attention output tends to be 0. We also follow this design and define a rule for the temperature factor T and the number of attention heads H, as shown in Table 1.

Multiple logical probability outputs can be obtained successively using multi-attention: $\hat{\mathrm{y}}_{T_1}, \hat{\mathrm{y}}_{T_2}, \ldots, \hat{\mathrm{y}}_{T_H}$. Note that each output here is from a class-specific residual attention module. In this paper, the logical output of each head is summed directly to get the final multi-attention logic output:

$$\widehat{\mathrm{y}}_o = \sum_{i=1}^{H}\widehat{\mathrm{y}}_{T_i} = \sum_{i=1}^{H}\left(y^1, y^2, \ldots, y^C\right)|_{T_i} = \sum_{i=1}^{H}\left(\mathrm{m}_1^T\mathrm{f}^1, \mathrm{m}_2^T\mathrm{f}^2, \ldots, \mathrm{m}_C^T\mathrm{f}^C\right)|_{T_i}, \tag{6}$$

where C is num of classes, T_i is the temperature factor of the class *i*. The class-specific residual attention module incorporates the attention features of all locations. Intuitively, when complex textures exist in the medical image, this module has a clear advantage over global average pooling or maximum pooling alone.

Residual refinement method

To enhance the performance of convolutional neural networks while maintaining computational efficiency, we propose a systematic residual refinement approach that operates on both the input preprocessing stage and the feature representation enhancement. Our methodology introduces two key architectural modifications: (1) a

Number of Attention Heads	Temperature Factors
$H = 2$	$T_0 = 1, T_1 = 99$
$H = 3$	$T_{0:1} = 1, 2, T_2 = 99$
$H = 4$	$T_{0:2} = 1, 2, 3, T_3 = 99$
$H = 5$	$T_{0:3} = 1, 2, 3, 4, T_4 = 99$
$H = 6$	$T_{0:4} = 1, 2, 3, 4, 5, T_5 = 99$
$H = 7$	$T_{0:5} = 1, 2, 3, 4, 5, 6, T_6 = 99$
$H = 8$	$T_{0:6} = 1, 2, 3, 4, 5, 6, 7, T_7 = 99$

Table 1. The rule maps of temperature factor *T* and attention heads *H*. The probability value of the output is gradually controlled and reduced according to the orders, and the last attention output tends to be 0 (*T*=99).

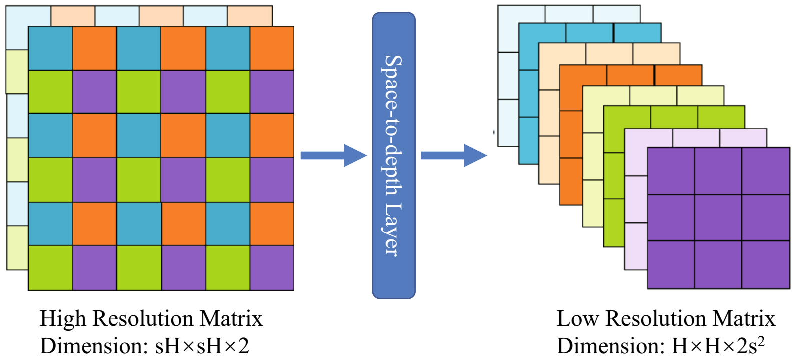


Figure 1. The transformation process of the Space-to-depth Layer. The high-resolution matrix is converted into a deeper matrix with lower resolution, while retaining more pixel information for later layers. *s* is the upscaling factor, *H* represents the height of the matrix.

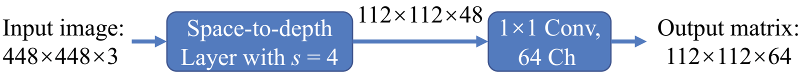


Figure 2. The proposed down-sampling method. The Space-to-depth layer is followed by 1×1 convolution, which can replace the traditional 7×7 convolution down-sampling method, thus preserving more information.

space-to-depth preprocessing layer that preserves spatial information while expanding the receptive field, and (2) channel-wise attention mechanisms via Squeeze-and-Excitation blocks to boost representational capacity.

Improving the extent of receptive field

Deep learning networks usually first resize the input image and down-sample the spatial dimension to 224× 224×3. Sandler et al.[26] prove that simply increasing the image resolution does not significantly improve the model's performance because the high resolution only plays a limited role in the first few layers of the network. The accuracy of the model is affected by the receptive field other than high resolution. The model accuracy can gradually increase by increasing the image resolution in a specific receptive field. However, when the image resolution exceeds the receptive field, the increasing trend of model accuracy will slow down or even decline. In this paper, the Space-to-depth layer is used to increase the extent of the receptive field, which can retain more pixel information for later layers. On the contrary, most popular networks use the convolution module to reduce image resolution directly. Figure 1 shows the transformation of the Space-to-depth layer: the high-resolution matrix is converted into a deeper matrix with lower resolution, while the original image information is not lost.

We use the Space-to-depth layer followed by 1×1 convolution to match the required number of channels, which can replace the traditional 7×7 convolution down-sampling method, thus preserving more information, as seen in Fig. 2. In this research, we used the input resolution higher than the classic image resolution, which is 448×448. After the Space-to-depth layer and 1×1 convolution processing, we can get the output matrix with 112×112×64 dimensions.

Boosting the representation power of the model

To improve the representation of the model, we introduce the Squeeze-and-Excitation block [25]to optimize our network. For a given convolutional transformation:

$$\mathrm{F}_{tr}:\mathrm{X}\rightarrow\mathrm{U},\mathrm{X}\in\mathbb{R}^{H'\times W'\times C'},\mathrm{U}\in\mathbb{R}^{H\times W\times C},\tag{7}$$

where X is input matrix, H', W', and C' represent the height, width, and channel of the input matrix, U is output matrix from the convolution or a set of the convolutions, *H*, *W*, and *C* represent the height, width, and channel of the output matrix. First, we passed the feature matrix U through to the Squeeze operation, which aggregates its spatial dimensions $H \times W \times C$ into embedding space Z with $1 \times 1 \times C$. The embedding space can respond to the global feature distribution channel-wise, conveying the information flow to the bottom layer and improving the overall model representation ability. The global average pooling is used to perform the squeeze operation, and Z is calculated by:

$$\mathrm{Z}=\mathrm{F}_{sq}(\mathrm{U})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathrm{U}(i,j)\tag{8}$$

To fully capture the channel-wise dependence, the excitation operation is used to learn the nonlinearity between the channels and ensure that the multiple channels are activated. We utilize the gating mechanism with a sigmoid activation to implement the excitation process, and excitation output S is calculated by:

$$\mathrm{S} = \mathrm{F}_{ex}(\mathrm{Z}, \mathrm{W}) = \sigma(g(\mathrm{Z}, \mathrm{W})) = \sigma\left(\mathrm{W}_2 \delta\left(\mathrm{W}_1 \mathrm{Z}\right)\right), \tag{9}$$

where σ is the ReLU function, $\mathrm{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $\mathrm{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, and r is hyperparameter. The actual module construction of the Excitation process uses two fully-connected layers to stack. The first fully-connected layer compresses the number of channels according to the parameter *r* to reduce the number of channels to *C*/*r*, and the second expands the channels to the original size *C*. Finally, the feature matrix U is reweighted according to the excitation output S to generate the output of the entire Squeeze-and-Excitation block, which will feed in subsequent layers, as shown in Fig. 3. The output of the Squeeze-and-Excitation block $\tilde{\mathrm{X}}$ is calculated by:

$$\tilde{\mathrm{x}}_c = \mathrm{F}_{\text{ scale }}\left(\mathrm{u}_c, s_c\right) = s_c \cdot \mathrm{u}_c, \tilde{\mathrm{X}} = [\tilde{\mathrm{x}}_1, \tilde{\mathrm{x}}_2, \ldots, \tilde{\mathrm{x}}_c], \tag{10}$$

where s_c is the *c*-th element of S,the $\mathrm{F}_{\mathrm{scale}}$ means the channel-wise multiplication of the scalar S and matrix U.

Overall architecture

The novel Squeeze-and-Excitation block was introduced to design the ResNet50 enhanced model to obtain better model representation capabilities based on the classical residual network design concept. We proposed a new down-sampling method based on the Space-to-depth layer followed by 1×1 convolutions to increase the size of the receptive field. For the classifier of the model, the proposed multi-attention method is used to set different classifier parameters according to the category to improve the accuracy of the model. The block design, width, and quantity of each layer of ResNet50 enhanced model are shown in Table 2.

Computational complexity

Our refinement framework is specifically designed to maintain computational efficiency while enhancing model capability. The modular design ensures that each enhancement component contributes meaningfully to performance improvement while minimizing computational overhead. We present a detailed breakdown of FLOPs consumption for each architectural modification when applied to the ResNet50 backbone with 448×448 input resolution Table 3.

Experiments

This section discusses the evaluation metrics, comparisons from different models, and ablation experiments. The proposed and currently popular medical classification models are pre-trained on open source datasets[28] and then fine-tuned on the EyePACS dataset[20]. The EyePACS dataset is presently the largest RA public classification dataset, containing 88702 classified color fundus images. This dataset can be divided into five categories: 0 (No DR), 1 (Mild DR), 2 (Moderate), 3 (Severe), and 4 (Proliferative DR). The dataset is relatively close to the needs of the actual clinical environment. The dataset contains noise from images and labels, such as the image may be underexposed and out of focus. Due to the large dataset size and the near-realistic data noise, the accuracy and robustness of each model can be compared appreciably. The officially divided training set (46032) and test set (42670) were used in this experiment on different models.

Due to the unavailability of public labels for the original Kaggle competition's EyePACS test set, we implemented a comprehensive data split strategy to ensure robust and fair evaluation. We combined two data sources: the EyePACS dataset and clinical images from our institution. The EyePACS training set was randomly divided in an 8:2 ratio to create training and validation partitions. Additionally, we collected 1,800 fundus images from the ophthalmology outpatient clinic of the First Affiliated Hospital of Anhui University of Chinese Medicine. These clinical images were evenly distributed across all three splits (training, validation, and test) in a 1:1:1 ratio to ensure balanced representation. All clinical images were independently labeled and cross-validated

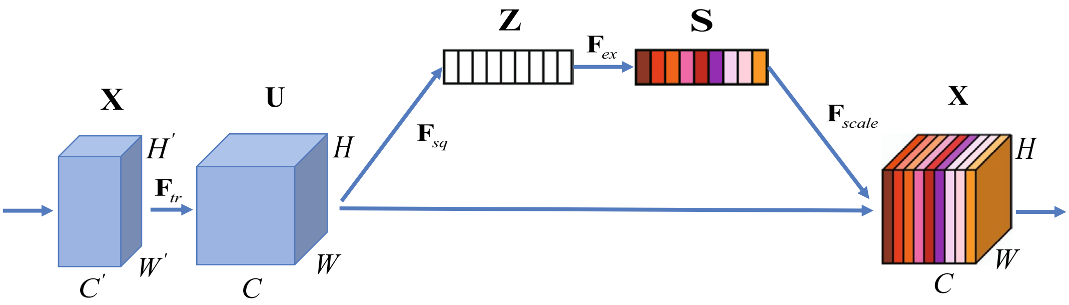


Figure 3. The design of the Squeeze-and-Excitation block. First, we passed the feature matrix U through to the squeeze operation to generate the embedding space Z with $1 \times 1 \times C$, then we utilized the gating mechanism with a sigmoid activation to implement the excitation process to get excitation output S. Finally, the feature matrix U is reweighted according to the excitation output S to get the output $\tilde{\mathrm{X}}$ of the Squeeze-and-Excitation block.

Layers name	**Output size**	**ResNet50 enhanced model**
Space-to-depth layer	**112×112**	**Space-to-depth layer, *s*=4**
Expansion layer	112×112	3×3 max pool
		$\begin{bmatrix} 1\times 1\,\text{conv},\,64 \\ 3\times 3\,\text{dwise} \\ \text{SE}\,layer, r=6 \\ 1\times 1\,\text{conv},\,128 \end{bmatrix} \times 4$
Residual_block1	28×28	$\begin{bmatrix} 1\times 1\,\text{conv},\,128 \\ 3\times 3\,\text{dwise} \\ \text{SE}\,layer, r=6 \\ 1\times 1\,\text{conv},\,256 \end{bmatrix} \times 4$
Residual_block2	14×14	$\begin{bmatrix} 1\times 1\,\text{conv},\,256 \\ 3\times 3\,\text{dwise} \\ \text{SE}\,layer, r=6 \\ 1\times 1\,\text{conv},\,512 \end{bmatrix} \times 6$
Residual_block3	7×7	$\begin{bmatrix} 1\times 1\,\text{conv},\,512 \\ 3\times 3\,\text{dwise} \\ \text{SE}\,layer, r=6 \\ 1\times 1\,\text{conv},\,1024 \end{bmatrix} \times 4$
Multi_attention	1×5	Multi-attention module

Table 2. The overall architecture of the ResNet50 enhanced model. The size of the input image is 448×448×3. All convs and dwises from the residual blocks are preceded by the BN + ReLU layer; the SE layer represents the Squeeze-to-Excitation block; dwise means depthwise separable convolution[27].

Component	**ResNet50 Baseline (M-FLOPs)**	**ResNet50 Enhanced (+Ours) (M-FLOPs)**
Standard ResNet50 backbone	16,400	16,400
Space-to-depth preprocessing	0	5
Expansion layers	0	1350
SE blocks (distributed)	0	670
Multi-attention module	0	12
Total	16,400	18,437

Table 3. Detailed FLOPs analysis of the proposed architectural enhancements. The baseline represents standard ResNet50 with 448×448 input resolution (16.4×10^9 FLOPs), and our method shows the additional computational cost introduced by each component.

by three experienced professional ophthalmologists to ensure label accuracy and provide objective assessment on real-world clinical data. We applied standard data augmentation techniques including horizontal flipping (p=0.5) and random rotation ($\pm 15^{\circ}$).

The models we constructed are all built on the PyTorch library, and other deep learning libraries also can implement these models. We used stochastic gradient descent for learning and fine-tuned all parameters of models after pre-training, with learning rate set to 0.001, weight decay set to 0.001, momentum set to 0.9, and batch size set to 32. We do not use dropout. Training was conducted for a maximum of 200 epochs with early stopping enabled (patience=10 epochs) based on validation loss to prevent overfitting. The training environment was Ubuntu 18.04 workstation with an Intel(R) Xeon(R) Silver 4114 CPU and 1080Ti 8GB GPU.

Evaluation metrics

We calculated the mean values of sensitivity, specificity, and AUC for the five categories in the test set. The optimal threshold is determined according to the maximum sum of the sensitivity and specificity, also known as the Youden index. We also used the Kappa coefficient to evaluate the model. The mathematical calculations of these metrics are as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{11}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{12}$$

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e}, \tag{13}$$

where TP represents the number of the sample which correctly classified as positive, FP is incorrectly classified as positive, TN is correctly classified as negative, FN is incorrectly classified as negative, p_0 indicates the relative

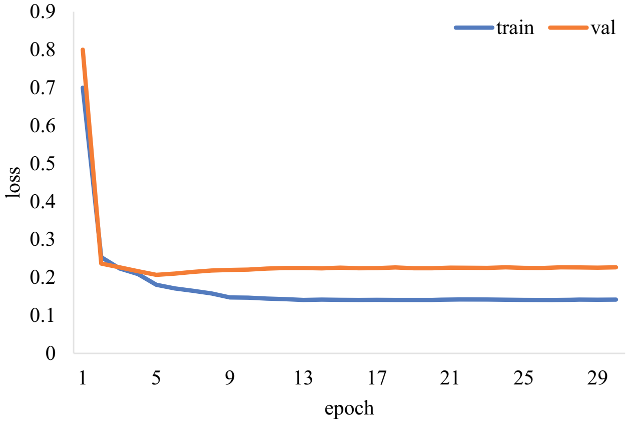


Figure 4. The fine-tuning process of the ResNet50 enhanced model in the test dataset.

Methods	Baseline models				Enhanced models (+ours)			
	Sens.	Spec.	Kappa	AUC	Sens.	Spec.	Kappa	AUC
MobileNet V3[27]	0.76	0.74	0.76	0.81	**0.78**	**0.77**	**0.79**	**0.84**
ResNet50[29]	0.78	0.77	0.78	0.84	**0.80**	**0.81**	**0.82**	**0.89**
ResNet101[29]	0.79	0.78	0.79	0.85	**0.81**	**0.82**	**0.83**	**0.90**
DenseNet121[30]	0.80	0.78	0.78	0.86	**0.82**	**0.81**	**0.82**	**0.91**
EfficientNet-B0	0.77	0.76	0.77	0.83	**0.79**	**0.79**	**0.80**	**0.86**

Table 4. Performance comparison of baseline models and their enhanced versions using our proposed refinement framework on the EyePACS test dataset. The enhancement includes space-to-depth preprocessing, expansion layers, SE blocks, and multi-attention mechanism ($H = 4$).

observed agreement among raters, and p_e represents the probability of each observer seeing each category randomly from the observed data.

Comparison

The fine-tuning process of the ResNet50 enhanced model is shown in Fig. 4. It can be seen that the training loss and validation loss slowly decline with the iteration of epochs, and the model does not suffer from overfitting problems. We selected the optimal training model with the lowest validation loss. The metrics of the different models in the test set are compared in Table 4. It shows that our ResNet50 enhanced model is better than other famous models and may indicate the effectiveness of our method in improving the size of the receiving field and representation capability.

To further analyze the ability of the ResNet50 enhanced model to discriminate between different diabetes categories, ROC curves are provided in Fig. 5. According to the ROC curves, the comprehensive classification performance of this model was good, but label 4 (proliferative DR) was inadequate. The normalized confusion matrix is provided in Fig. 6, and it can be seen that the part of proliferative DR samples is classified as severe or moderate DR. This indicates that the algorithm needs to be improved in the discrimination of proliferative DR. The algorithm has good discrimination ability for other DR classifications (No DR, Mild DR, Moderate DR, and Severe DR) and has specific clinical application value. Figure 7 shows attention heatmaps that validate our model's training effectiveness by demonstrating focused attention on pathological features in moderate NPDR cases.

Ablation study

We performed ablation experiments on the ResNet50 enhanced model to better understand its relative importance. The experimental results are shown in Table 5. It shows that the multi-attention module, the proposed down-sampling method, and the Squeeze-and-Excitation block all affect the model's performance. Among these methods, the Squeeze-and-Excitation block has the most significant impact on the model performance, suggesting that the higher complexity of the model may have a higher feature extraction ability. The multi-attention module has a specific influence on the results, which means that multi-attention can improve the model's accuracy. Compared with the traditional 7×7 convolution down-sampling method, the model accuracy of the proposed down-sampling method is improved, which indicates that the proposed down-sampling method can theoretically improve the model performance by improving the receiver domain of the model.

We also performed ablation experiments to explore the effect of the number of multi-attention heads on the model's accuracy. As seen in Table 6, increasing the number of heads will improve the model's accuracy to

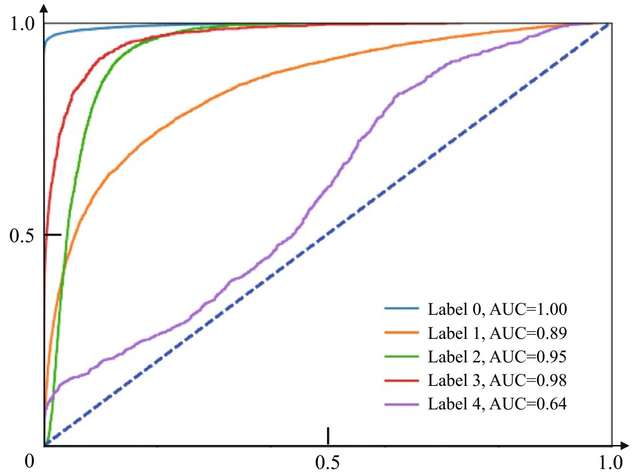
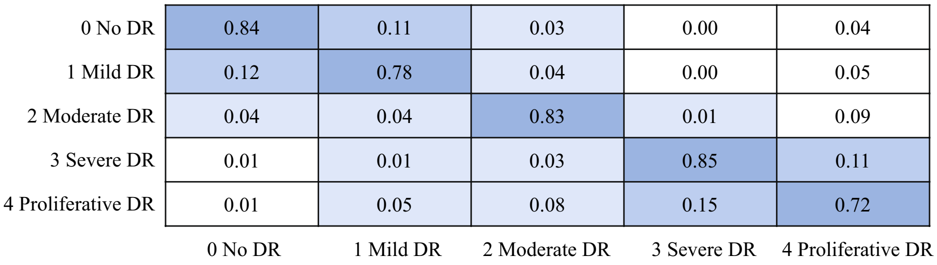


Figure 5. ROC curve of the ResNet50 enhanced model on our test datasets.



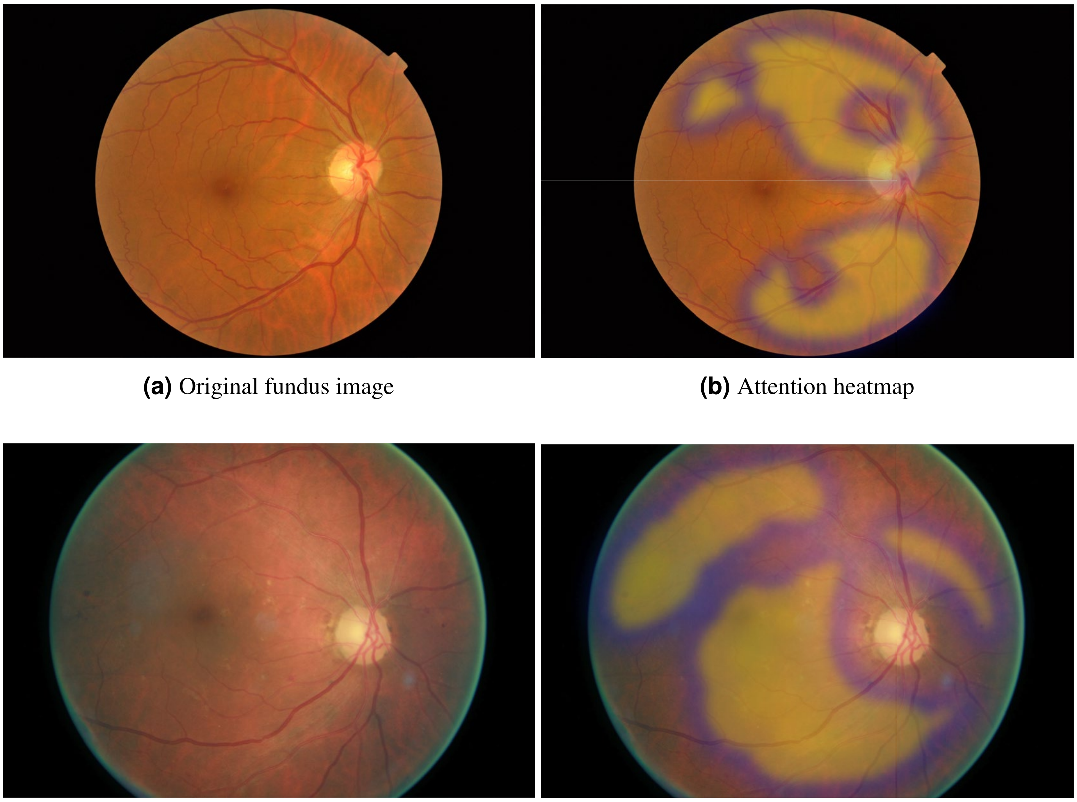
	0 No DR	1 Mild DR	2 Moderate DR	3 Severe DR	4 Proliferative DR
0 No DR	0.84	0.11	0.03	0.00	0.04
1 Mild DR	0.12	0.78	0.04	0.00	0.05
2 Moderate DR	0.04	0.04	0.83	0.01	0.09
3 Severe DR	0.01	0.01	0.03	0.85	0.11
4 Proliferative DR	0.01	0.05	0.08	0.15	0.72

Figure 6. Normalized confusion matrix of the ResNet50 enhanced model on the test dataset.

a certain extent, but the precision will decrease if the number of heads exceeds 5. This may indicate that the appropriate number of multi-attention heads may lead to better performance.

Conclusions

This study presents a multi-attention residual refinement framework that systematically enhances CNN architectures for diabetic retinopathy classification through class-specific attention, space-to-depth preprocessing, and Squeeze-and-Excitation blocks. Our experimental evaluation demonstrates consistent 2-5% performance improvements across multiple architectures (ResNet, DenseNet, EfficientNet, MobileNet) while maintaining computational efficiency and providing interpretable attention visualizations. Although our results suggest potential for clinical applications, we acknowledge important limitations including the need for external validation on independent datasets, formal clinical evaluation by additional ophthalmologists, and assessment of real-world deployment challenges beyond technical performance metrics. Future work should focus on prospective clinical validation and evaluation across diverse healthcare settings to establish the practical utility of our approach in actual clinical workflows.



(a) Original fundus image **(b)** Attention heatmap

(c) Original fundus image **(d)** Attention heatmap

Figure 7. ResNet50 enhanced model effectiveness demonstration on Grade 2 moderate NPDR cases. Original fundus images (**a**, **c**) and corresponding attention heatmaps (**b**,**d**) show strong correlation between pathological features and model focus regions.

Ablation study	Sensitivity	Specificity	Kappa	AUC
Replace the multi-attention module with an FClayer	0.77	0.80	0.80	0.87
Replace the proposed down-sampling method with a 7×7 convolution	0.79	**0.81**	0.81	0.88
Remove the squeeze-and-excitation block	0.76	0.79	0.81	0.86
ResNet50 enhanced (+Ours)	**0.80**	**0.81**	**0.82**	**0.89**

Table 5. Performance of ResNet50 enhanced model indifferent ablation experiments on our test dataset.

Number of attention heads	Sensitivity	Specificity	Kappa	AUC
$H = 2$	0.78	0.80	0.80	0.87
$H = 3$	0.79	0.80	0.81	0.88
$H = 4$	**0.80**	**0.81**	**0.82**	**0.89**
$H = 5$	0.79	**0.81**	**0.82**	0.88
$H = 6$	0.78	0.80	0.80	0.87
$H = 7$	0.77	0.80	0.80	0.86
$H = 8$	0.77	0.79	0.80	0.85

Table 6. Performance of the ResNet50 enhanced model used different numbers of attention heads on the test dataset.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 1 April 2025; Accepted: 6 August 2025
Published online: 10 August 2025

References

1. Association, A. D. et al. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **28**, S37 (2005).
2. Peate, I. Retinal detachment. *Br. J. Healthc. Assist.* **16**, 236–241 (2022).
3. Wässle, H. Parallel processing in the mammalian retina. *Nat. Rev. Neurosci.* **5**, 747–757 (2004).
4. Del Pinto, R. et al. Arterial hypertension and the hidden disease of the eye: Diagnostic tools and therapeutic strategies. *Nutrients* **14**, 2200 (2022).
5. Pan, W. W., Lin, F. & Fort, P. E. The innate immune system in diabetic retinopathy. *Prog. Retin. Eye Res.* **84**, 100940 (2021).
6. Selvachandran, G., Quek, S. G., Paramesran, R., Ding, W. & Son, L. H. Developments in the detection of diabetic retinopathy: A state-of-the-art review of computer-aided diagnosis and machine learning methods. *Artif. Intell. Rev.* **56**, 915–964 (2023).
7. Willis, J. R. et al. Vision-related functional burden of diabetic retinopathy across severity levels in the united states. *JAMA Ophthalmol.* **135**, 926–932 (2017).
8. Sun, H. et al. Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **183**, 109119 (2022).
9. Al-Antary, M. T. & Arafa, Y. Multi-scale attention network for diabetic retinopathy classification. *IEEE Access* **9**, 54190–54200 (2021).
10. Kassani, S. H. et al. Diabetic retinopathy classification using a modified xception architecture. In *2019 IEEE international symposium on signal processing and information technology (ISSPIT)*, 1–6 (IEEE, 2019).
11. Gayathri, S., Gopi, V. P. & Palanisamy, P. A lightweight cnn for diabetic retinopathy classification from fundus images. *Biomed. Signal Process. Control* **62**, 102115 (2020).
12. Kandel, I. & Castelli, M. Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review. *Appl. Sci.* **10**, 2021 (2020).
13. Alyoubi, W. L., Shalash, W. M. & Abulkhair, M. F. Diabetic retinopathy detection through deep learning techniques: A review. *Inf. Med. Unlocked* **20**, 100377 (2020).
14. Tsiknakis, N. et al. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Comput. Biol. Med.* **135**, 104599 (2021).
15. Ishtiaq, U. et al. Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues. *Multimedia Tools Appl.* **79**, 15209–15252 (2020).
16. Zhou, Y., Wang, B., Huang, L., Cui, S. & Shao, L. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE Trans. Med. Imaging* **40**, 818–828 (2020).
17. Sambyal, N., Saini, P., Syal, R. & Gupta, V. Modified u-net architecture for semantic segmentation of diabetic retinopathy images. *Biocybern. Biomed. Eng.* **40**, 1094–1109 (2020).
18. Porwal, P. et al. Idrid: Diabetic retinopathy-segmentation and grading challenge. *Med. Image Anal.* **59**, 101561 (2020).
19. Chakraborty, S. et al. Efficient conditional pre-training for transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4241–4250 (2022).
20. Bhaskaranand, M. et al. Eyeart+ eyepacs: automated retinal image analysis for diabetic retinopathy screening in a telemedicine system. In *Ophthalmic Medical Image Analysis International Workshop*, vol. 2 (University of Iowa, 2015).
21. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
22. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
23. Ridnik, T. et al. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1400–1409 (2021).
24. Zhang, L. & Schaeffer, H. Forward stability of resnet and its variants. *J. Math. Imaging Vis.* **62**, 328–351 (2020).
25. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
26. Sandler, M., Baccash, J., Zhmoginov, A. & Howard, A. Non-discriminative data or weak model? on the relative importance of data and model resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0 (2019).
27. Howard, A. et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324 (2019).
28. Irvin, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *In Proceedings of the AAAI conference on artificial intelligence* **33**, 590–597 (2019).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
30. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

Acknowledgements

We would like to acknowledge the First Clinical Medical College of Anhui University of Chinese Medicine for infrastructure and radiologist evaluation support.

Author contributions

Zijian Wang designed the study, collected and analyzed the data and wrote the manuscript. Yi Wang assisted with the data analysis, contributed to the data analysis and interpretation, and critically reviewed the manuscript. Chun Ma and Xuan Bao contributed to the study design, provided guidance on the data analysis and critically reviewed the manuscript. Ya Li provided guidance on the study design, contributed to the interpretation of the results, critically reviewed the manuscript and provided expertise in a specific area of the study. Each author has read and approved the final version of the manuscript and has agreed to be accountable for their contributions.

Funding

This work was supported by the Anhui Provincial Key Natural Science Research Project for Higher Education Institutions (No.2023AH050780), the Key Teaching Research Project of Anhui Provincial Department of Education (No.2022jyxm858), the Anhui Provincial Key Natural Science Research Project for Higher Education Institutions (No.2023AH050846), and the Teaching Reform Research Project of Anhui University of Chinese Medicine (No. 2023xjjy_yb016).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Anhui University of Chinese Medicine. Informed consent was obtained from all patients whose fundus images were included in this study. All patient data were anonymized prior to analysis.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.