Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Research on Retinal Vascular Pathologies in OCT Images with the Coupled ResNet and Transformer Model

Yongxiang Zhang
XIDIAN UNIVERSITY, School of Optoelectronic Engineering, Shanxi, 710126, China
Corresponding Author: Yongxiang Zhang, **E-mail**: zyx13233965972@163.com

| ABSTRACT

Retinal vascular pathologies can lead to vision loss, blindness, and severe complications (such as glaucoma and vitreous hemorrhage), significantly impacting quality of life. Traditional detection methods for retinal vascular pathologies rely on manual observation and subjective judgment, resulting in low efficiency and a high risk of missing early subtle lesions. They also suffer from limitations such as operational complexity and device dependency. To overcome the shortcomings of these methods, this study proposes a method for detecting retinal pathologies on the basis of OCT images of retinal vascular diseases that uses a coupled ResNet and transformer approach. RTHNet adopts an encoder-decoder structure. In the encoding stage, ResNet50 is employed as the backbone network to extract local features effectively from the images. An attention adaptive fusion module (AAFM) is designed to achieve efficient integration of multilevel attention features between the encoder and decoder. In the decoding stage, a global-local contextual transformer block (GLCTB) is designed to simultaneously focus on global contextual information and local details. At the end of the decoder, a detail enhancement module (DEM) is proposed, which refines the semantic consistency and spatial detail information between features to ensure the fineness and accuracy of the segmentation results. The RTHNet model was evaluated on the OCT2017 dataset and OCTAMNIST dataset. The results revealed that the classification accuracy reached 94.78% and 83.41% on the retinal OCT2017 dataset and OCTAMNIST dataset, respectively. Compared with other traditional methods, such as the CNN and DCNN algorithms, the detection accuracy improved by 7.28% and 0.78%, respectively. The proposed method, which couples ResNet and Transformer, overcomes the bottlenecks of subjectivity and low efficiency inherent in traditional detection methods that rely on manual interpretation. It enables high-precision automatic identification and early warning of vascular pathologies in OCT images, providing an intelligent auxiliary diagnostic tool for clinical practice. This reduces misdiagnosis rates and promotes the automation and precision of ophthalmic disease screening.

| KEYWORDS

Retinal vascular pathologies; OCT images; RTHNet; AAFM; GLCTB; Classification accuracy

| ARTICLE INFORMATION

ACCEPTED: 15 June 2025 **PUBLISHED:** 01 July 2025 **DOI:** 10.32996/jcsts.2025.7.7.12

1 Introduction

Retinal vascular pathologies can lead to irreversible vision impairment and even blindness, posing a severe threat to patients' visual health and quality of life. Their high blindness-causing potential not only deprives patients of their working capacity but also increases family care burdens. Research has shown that as of 2019, there were approximately 710 million patients with diabetic retinopathy globally, with over 12 million new cases annually [1]. Therefore, achieving accurate and rapid detection of retinal vascular pathologies can not only effectively delay disease progression and reduce the risk of blindness through early intervention but also enhance public health efficiency through large-scale screening, reducing medical expenditures for patients

with advanced severe disease. This finding holds strategic significance for optimizing ophthalmic healthcare resource allocation and addressing blindness prevention and control in an aging society.

Traditional detection methods for retinal vascular pathologies include ophthalmoscopy, fundus fluorescein angiography (FFA), and manual reading on the basis of fundus photographs. Ophthalmoscopy can be used to determine pathologies by directly observing the morphology of fundus vessels. However, the accuracy of this method is highly dependent on physician experience and has a limited ability to identify early microvascular lesions [2]. Fundus fluorescein angiography involves intravenous injection of sodium fluorescein, using a fundus camera to capture the flow of the fluorescent dye within the retinal vessels to reveal vascular leakage, occlusion, and abnormal neovascularization. However, this method is invasive and can cause allergic reactions, nausea, vomiting, or renal burden [3]. Manual reading based on fundus photographs involves taking color or black-and-white fundus photos, with physicians or technicians manually interpreting vascular pathology features. This method suffers from strong subjectivity in human interpretation and poor consistency among different readers [4]. In recent years, the use of optical coherence tomography (OCT) technology for detecting retinal vascular pathologies has gained widespread interest. OCT technology is based on the principle of low-coherence light interference. By emitting near-infrared light to the retina and utilizing the optical path difference of reflected light from different tissue layers to collect interference signals, high-resolution cross-sectional images of retinal layers are reconstructed after computer processing. This allows for the precise capture of microstructural changes such as retinal thickness, vascular morphology, and tissue edema. Hwang et al. used a deep convolutional neural network (CNN) to analyze spectral-domain optical coherence tomography (SD-OCT) images from patients with retinal angiomatous proliferation (RAP) or polypoidal choroidal vasculopathy (PCV) and a control group. The results revealed that the proposed model achieved an accuracy, sensitivity, and specificity of 89.1%, 89.4%, and 88.8%, respectively [5]. Hassan et al. classified retinal OCT images via an enhanced optical coherence tomography (EOCT) model. The results showed that the proposed model achieved a sensitivity and specificity of 0.9836 and 0.9615, respectively [6]. SVA et al. used a dilated depthwise separable convolution deep residual network (ResNet) for the automatic identification and severity classification of retinal biomarkers via SD-OCT. The results revealed that the proposed model achieved an overall accuracy of 98.64% [7]. OCT technology offers significant advantages over the aforementioned traditional methods for detecting retinal pathologies. First, it requires no contact with the eyeball or injection of contrast agents, avoiding the risks of invasive procedures, ensuring high safety, and making it suitable for children, elderly individuals, and people with allergies. Second, its tomographic imaging capability with micron-level resolution can clearly display subtle lesions in each layer of the retina, significantly exceeding the detection rate of early microabnormalities compared with traditional methods relying on naked-eye observation or planar imaging. Finally, automated analysis software quickly generates quantitative parameters such as retinal thickness and vascular density, greatly shortening the detection time and reducing subjective errors in manual interpretation. Therefore, this study employs OCT technology to detect retinal vascular pathologies.

Zhou Tao et al. used the U-Net network for medical image segmentation. However, U-Net's encoder-decoder structure is relatively simple, making it difficult to precisely capture the boundaries and details of objects with very complex topologies or diverse shape changes [8]. Jin Yan et al. used a recurrent residual convolutional neural network for medical image segmentation, but this network has high computational complexity and a significant risk of overfitting [9]. Wu Tong et al. used the segment anything model (SAM) for medical image segmentation, but this model has limited adaptability to complex scenes and insufficient mask prediction accuracy [10]. Deng Erqiang et al. used an enhanced generative adversarial network (EnGAN) for medical image segmentation, but this algorithm has limited noise processing capability and lacks focus on local details [11]. Given the shortcomings of the above methods, this study employs a coupled ResNet and transformer model to classify retinal OCT images. This model balances local feature extraction and global contextual modeling by fusing the advantages of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). Comparative experiments on the Vaihingen, Potsdam, and WHDLD datasets validated the network's effectiveness. A global-local contextual transformer block (GLCTB) was designed. This module effectively combines multiscale global and local features, fully considering the differences in feature characteristics within high-resolution OCT images, achieving improvements in balancing global and local features. To enhance the effective fusion of global and local information and the recognition capability of object boundaries, this paper designs the attention adaptive fusion module (AAFM) and detail enhancement module (DEM). The AAFM dynamically adjusts feature fusion between the encoder and decoder, enhancing the interaction between local and global information; the DEM focuses on the restoration of fine-grained semantics and spatial details, improving the positioning accuracy of object edges and mitigating the effects of shadow occlusion and edge blurring.

2 Dataset

2.1 Retinal OCT2017 dataset

The OCT2017 dataset is a landmark benchmark dataset in the field of retinal disease diagnosis that was publicly released via the Kaggle platform by the University of California, San Diego, in 2017 [12]. As the first large-scale dataset for OCT image

classification, its design follows clinical principles, aiming to provide standardized data support for deep learning model training and automated diagnosis of retinal pathologies. The dataset consists of two parts: a training set and a test set. The training set contains 108,308 OCT B-scan images, whereas the test set contains 1,000 B-scan images. The entire dataset covers four core pathology categories: choroidal neovascularization (CNV), diabetic macular edema (DME), Drusen, and normal. The number of images in the training set for each category is 37,454 CNV images, 11,698 DME images, 8,616 Drusen images, and 51,390 normal images. The test set uses stratified sampling to ensure balanced samples across pathology categories (250 images per class). The original data were converted from proprietary device formats to JPEG, with the resolution standardized to 512×512 pixels, retaining 8-bit grayscale information. A threshold segmentation algorithm was used to automatically locate retinal layer structures and remove edge artifacts. The annotation work employed Kappa consistency testing (κ=0.94) to ensure labeling consistency. Annotation files contain image filenames and corresponding class labels, which are stored in CSV format. Dataset access address: https://github.com/openmedlab/Awesome-Medical-Dataset/blob/main/resources/OCT2017.md

2.2 OCTAMNIST Dataset

The OCTAMNIST dataset is the first medical image classification dataset designed following the paradigm of the classic MNIST dataset, focusing on automated diagnostic research for retinal diseases [12]. This dataset was officially released in 2020 by the Johns Hopkins University School of Medicine in collaboration with the Stanford University Artificial Intelligence Laboratory, aiming to provide lightweight, standardized entry-level data resources for the medical image analysis field. The dataset adheres to FAIR data principles (Findable, Accessible, Interoperable, and Reusable). Through standardized data collection, preprocessing, and annotation processes, a standardized dataset suitable for basic algorithm verification and teaching practice is constructed. All patient information in the dataset was desensitized by hashing algorithms. The dataset comprises 10,000 OCT B-scan images. Stratified random sampling was used to divide the dataset into a training set (60%), a validation set (20%), and a test set (20%). The entire dataset covers four core categories: normal (N) - 2,500 images; drusen (D) - 2,500 images; diabetic macular edema (E) - 2,500 images; and choroidal neovascularization (C) - 2,500 images. Each category is evenly distributed across the three subsets. All the images were subjected to a standardized image processing pipeline to ensure data consistency: regions of interest (ROIs) were resized to 28×28 pixels (grayscale), matching the MNIST input format; histogram equalization was applied to enhance contrast in lesion areas; and median filtering was used to remove salt-and-pepper noise. The annotation work employed Kappa consistency testing (κ=0.92) to ensure labeling consistency. Each image corresponds to a single label (N/D/E/C). Annotation files are stored in CSV format, containing image filenames and corresponding labels. Dataset access address: https://github.com/MedMNIST/MedMNIST/tree/main

3 Method

3.1 Overall Framework

RTHNet adopts a classic encoder-decoder architecture, as shown in Figure 1. Its core idea is to increase the segmentation accuracy of medical images through the synergistic optimization of local feature extraction and global contextual modeling. The encoder uses a pretrained ResNet50 as the backbone network, constructing a multilevel feature pyramid through four stages of progressive downsampling. The input image passes through four feature stages, generating feature maps at resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image. Shallow features (1/4, 1/8 scale) retain high-resolution texture and edge details; middle-level features (1/16 scale) encode local semantic information; and deep features (1/32 scale) capture abstract semantic representations through a global receptive field, forming a multigranular feature expression.

The decoder part achieves feature reconstruction and optimization through three levels of processing. First, the global-local contextual transformer block (GLCTB) performs dual-path modeling on the deepest features: it uses a self-attention mechanism to capture global contextual dependencies while simultaneously extracting local detail features through multiscale convolutions. Second, the attention adaptive fusion module (AAFM) employs attention and dynamic weight adjustment strategies to perform cross-layer fusion of features from each encoder stage with decoder features, effectively balancing the integration of semantic information and spatial details while suppressing irrelevant background interference. Finally, the detail enhancement module (DEM) refines the fused features. (Note: The original text contained an incomplete sentence about decoder limitations, which seemed misplaced. The translation focuses on the described RTHNet structure.) The fixed receptive field of conventional operations makes it difficult to capture long-range dependencies effectively, leading to incomplete modeling of global semantic relationships. Simultaneously, the feature oversmoothing problem during global modeling makes it difficult to retain local details and small target features. OCT scenes involve multiple target scales, requiring the decoder to possess stronger multiscale feature processing capabilities. How to effectively restore local details and edge features while maintaining the integrity of global information is a key issue that urgently needs resolution in current decoder modules.

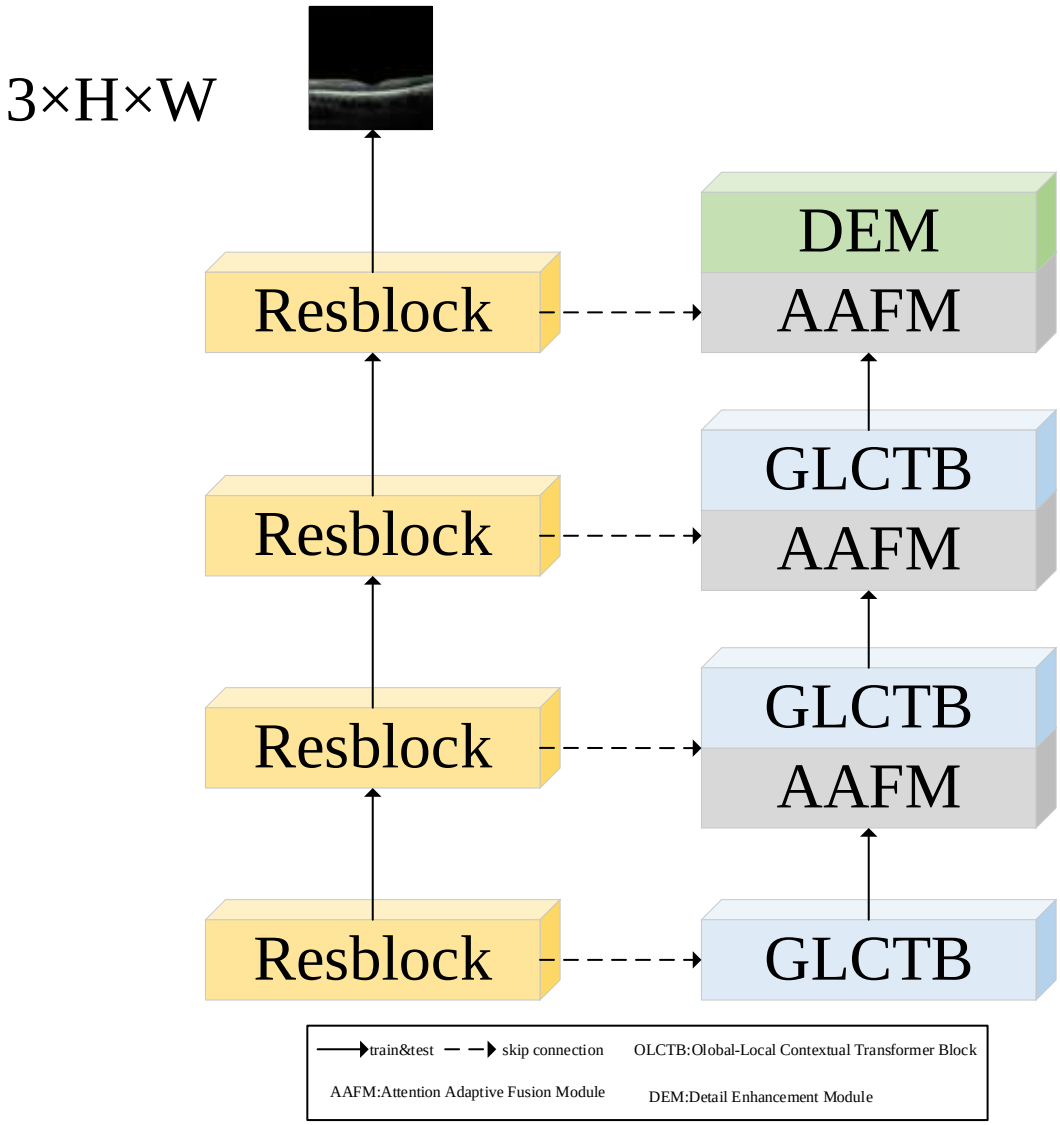

Figure 1 RTHNet network structure diagram

3.2 Global-Local Contextual Transformer Block (GLCTB)

Inspired by UnetFormer, this study proposes a global-local contextual transformer block (GLCTB) to synergistically optimize global information modeling and local detail recovery, thereby increasing semantic segmentation accuracy in complex OCT scenes. As shown in Figure 2, the GLCTB module consists of a global-local contextual attention module, normalization layers, and a multilayer perceptron. While the global attention mechanism of traditional transformers can model long-range dependencies, its window partitioning strategy can easily lead to fragmented spatial information, and single-scale convolutions struggle to cover the multiscale nature of OCT features. Therefore, the global-local contextual attention module adopts a dual-branch collaborative mechanism, combining the local feature extraction capability of depthwise separable convolutions with the global attention modeling advantage of transformers, achieving complementary optimization of multilevel features and mining deep semantic information in OCT data.

The global branch adopts a window partitioning strategy, dividing the feature map into multiple local windows. Attention is computed independently within each window. This design retains the global modeling capability while adapting to the processing needs of high-resolution data. The input feature $X \in R^{B\times C\times H\times W}$ is projected into {query, key, value} matrices via a 1 × 1 convolution:

$$Q,K,V = \mathrm{Conv}_{1\times1}(X) \in R^{B\times 3C\times H\times W} \quad (1)$$

where C is the number of channels. The projected matrices are split into three parts along the channel dimension, corresponding to Q, K, and V. Subsequently, the feature map is divided into $\mathrm{N}=\frac{\mathrm{H}}{S}\times\frac{W}{S}$ nonoverlapping windows (window size S × S). The features within each window are reshaped into:

$$Q,K,V\in \mathrm{R}^{B\times S\times h\times S^2\times d} \quad (2)$$

where **h** is the number of attention heads and **d = C/h** is the head dimension. After reshaping, attention computation is performed within each window. Within each window, the similarity between the query and key is calculated via scaled dot-product attention.

The spatial distribution of features exhibits strong structural characteristics, whereas standard transformers lack explicit modeling of positional information. Therefore, the global branch introduces relative position encoding **B** to explicitly encode the relative distances between pixels, enhancing the model's perception of spatial relationships:

$$\mathrm{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}+B\right)\mathbf{V} \quad (3)$$

where $B\in\mathrm{R}^{(2S-1)\times(2S-1)\times n}$ is the relative position bias and where $\sqrt{d}$ is used to scale gradient stability. To avoid information loss, the attention output undergoes an inverse transformation operation via a set of windowed vertical stripe convolutions with kernel sizes (w, 1) and (1, w), resulting in features rich in global semantic informatio F_{attn} n. Here, **w** is the window size, defined as:

$$F_{\mathrm{attn}}=\mathrm{WindowReverse}\left(\mathrm{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)\right) \quad (4)$$

To enhance feature representation capability and alleviate the oversmoothing problem of attention mechanisms, depthwise separable convolution is introduced to further aggregate features, yielding F_{Global}:

$$F_{\mathrm{Global}}=\mathrm{Conv}_{\mathrm{DWS}}^{(w\times w)}\left(F_{\mathrm{attn}}\right) \quad (5)$$

Unlike the local branch module designed by Kumar et al., which consists of three parallel convolutional layers, the local branch module designed in this paper extracts multigranular local features through three groups of depthwise separable convolutions with different receptive fields, balancing computational efficiency and detail preservation capability. Let the input feature be $X\in R^{B\times C\times H\times W}$. The module contains three groups of horizontal-vertical separable convolutions with kernel sizes of 1×7, 1×11, and 1×21. Each convolution operation is defined as:

$$F_l^{(k)}=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{\mathrm{DWS}}^{(1\times k)}\left(X\right)\right)\right)+\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{\mathrm{DWS}}^{(k\times 1)}\left(X\right)\right)\right) \quad (6)$$

where **k** ∈ {7, 11, 21}, and **BN** is the batch normalization layer. Horizontal and vertical separable convolutions enhance the ability to capture linear features such as roads and building edges, respectively, and reduce feature interference. Large convolution kernels (1 × 21) cover broad contextual areas, whereas small convolution kernels (1 × 7) focus on local details, forming complementary feature representations. After the three groups of features along the channel dimension are summed, a 1 × 1 convolution compresses the channel activation. To preserve the original feature information, avoid gradient vanishing, and simultaneously highlight multiscale detail responses, a residual connection with the original input **X** is introduced, as shown in equation (7).

$$F_{\mathrm{Local}}=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times 1}\left(\sum_k F_l^{(k)}\right)\right)\right)\odot X \quad (7)$$

In equation (7), $\odot$ denotes elementwise multiplication. The resulting feature map F_{Local} contains rich multiscale contextual information. Finally, the output feature maps of the global branch and the local contextual branch are summed elementwise to generate the final fused feature map F_{Glca}:

$$F_{Glca}=F_{Global}+F_{Local} \quad (8)$$

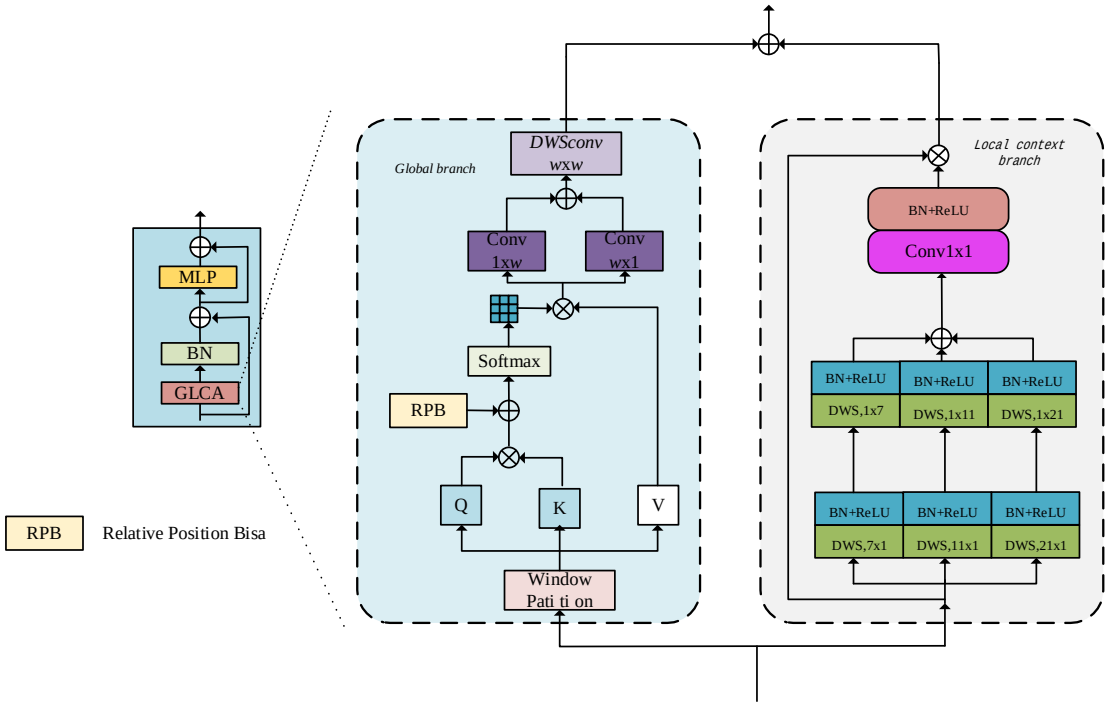

Figure 2 Global-Local Contextual Transformer Block (GLCTB)

This fusion strategy preserves the efficiency of CNNs in local feature extraction while leveraging the advantages of ViT in capturing global information. By enhancing salient features and suppressing noise, it improves the stability and generalization capability of the network, strengthening the model's ability to capture information across different scales and long-range dependencies.

3.3 Attention Adaptive Fusion Module (AAFM)

Traditional feature fusion modules often use fixed-ratio pixelwise weighting or simple concatenation and lack the ability to dynamically adjust feature weights, which can easily lead to information redundancy or weakening of critical features. To address this, this paper proposes the attention adaptive feature fusion module (AAFM), which achieves adaptive fusion of cross-layer features between the encoder and decoder through a dual-attention guidance and dynamic weight allocation strategy. As shown in Figure 3, the AAFM consists of four parts: channel attention, spatial attention, dynamic weight fusion, and postprocessing convolution. As shown in Figure 4, the channel attention mechanism first filters the upsampled decoder features along the channel dimension. Specifically, this mechanism combines global average pooling (GAP) and global max pooling (GMP) to extract channel-level feature responses, generating channel weights ω_c through a bottleneck structure composed of two layers of 1 × 1 convolutions. This process effectively suppresses responses from redundant channels, achieves adaptive enhancement of important features, and avoids inefficient computation and information interference caused by equal processing of all channels, thereby improving the effectiveness of feature representation and segmentation accuracy. The implementation principle is as follows:

$$\omega_c = \sigma\left(f_{\text{conv}}\left(\text{GAP}\left(X_{\text{up}}\right)\right) + f_{\text{conv}}\left(\text{GMP}\left(X_{\text{up}}\right)\right)\right) \quad (9)$$

In equation (9), f_{conv} represents the bottleneck structure composed of a squeeze layer and an expansion layer, and σ is the sigmoid function. The dual pooling strategy balances global statistics and local salient responses, enhancing the model's sensitivity to key semantic channels.

In medical images, the spatial distribution of similar features (such as scattered vegetation) is highly irregular. The spatial attention mechanism aims to further locate target regions in the spatial dimension and weaken the influence of irrelevant background areas on the fusion result. Its structure is shown in Figure 5. For the input feature map, the mean and max features are aggregated along the channel dimension, considering both smooth regions and salient edges to avoid bias from a single

statistic. Finally, a 7 × 7 large-kernel convolution covers broad contextual areas, suppressing isolated noise point interference while enhancing the response intensity in edge regions, generating spatial weights ω_s :

$$\omega_s = \sigma\left(\mathrm{Conv}_{7\times7}\left(\mathrm{Concat}\left[\mathrm{AvgPool}(X), \mathrm{MaxPool}(X)\right]\right)\right) \quad (10)$$

In medical image semantic segmentation, the identification of shadowed and occluded areas relies on the decoder's detail recovery capability, whereas semantically clear areas require reinforcement of the encoder's global consistency. Dynamic weight fusion introduces learnable parameters W_1 and W_2 to adaptively adjust the fusion ratio of encoder residual features X_{res} and decoder features X_{up}. The ReLU6 function is used for normalization to generate fusion weights:

$$\omega_i = \frac{\mathrm{ReLU6}(w_i)}{\sum_{j=1}^{2}\mathrm{ReLU6}(w_j)}, \quad i = 1,2 \quad (11)$$

The formula for calculating the fused feature is shown in equation (12):

$$F_{\text{fused}} = w_1 \cdot X_{\text{res}} + w_2 \cdot \left(\omega_c \otimes \omega_s \otimes X_{\text{up}}\right) \quad (12)$$

Finally, the fused features are refined via a 5×5 depthwise separable convolution to increase their detail capture capability, resulting in features F_{out}. The calculation formula is (13):

$$F_{\text{out}} = \mathrm{Conv}_{\text{DWS}}^{(5\times5)}\left(F_{\text{fused}}\right) \quad (13)$$

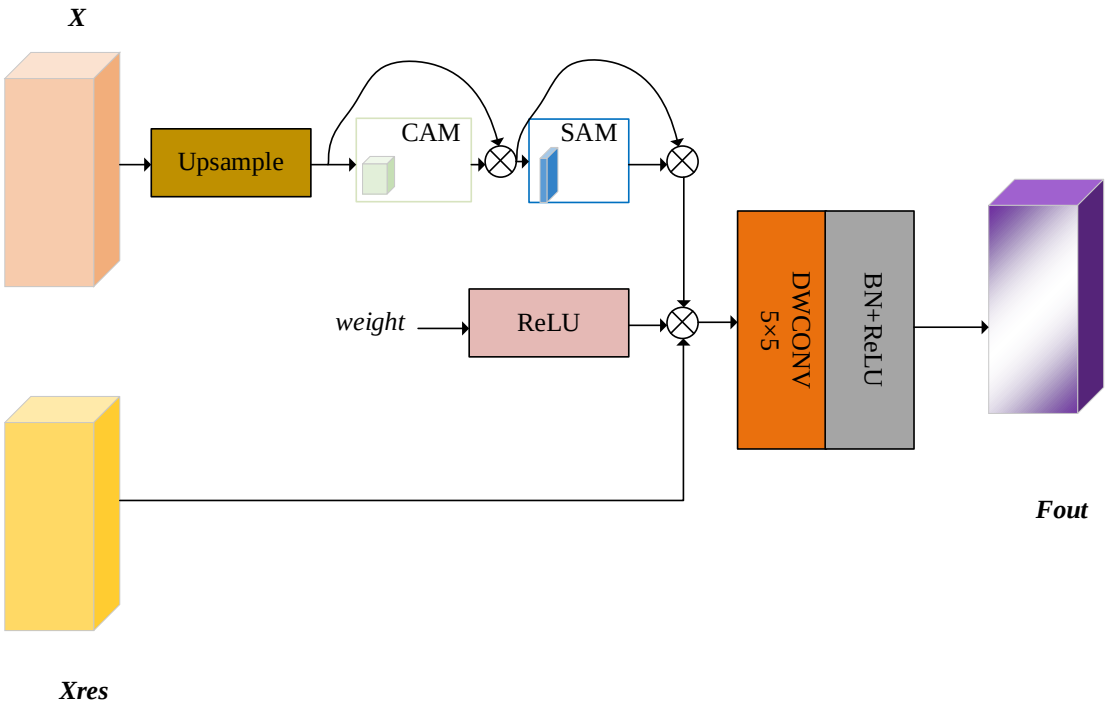

Figure 3 Attention adaptive fusion module (AAFM)

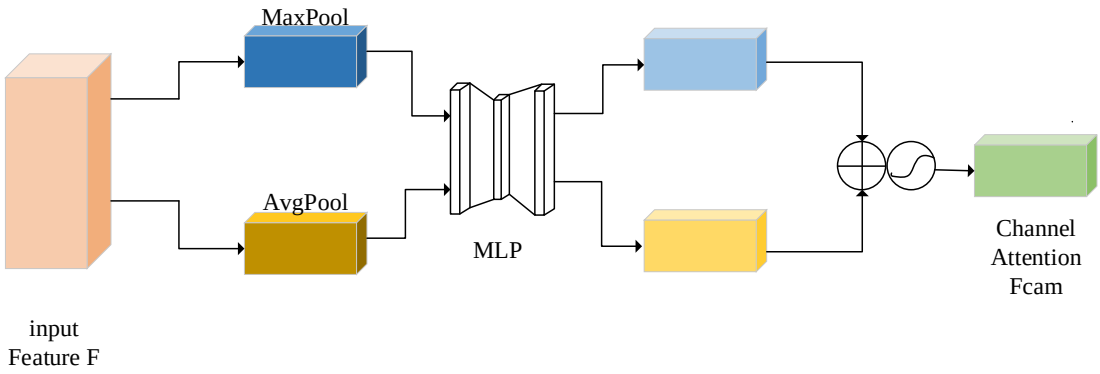

Figure 4 Channel attention mechanism

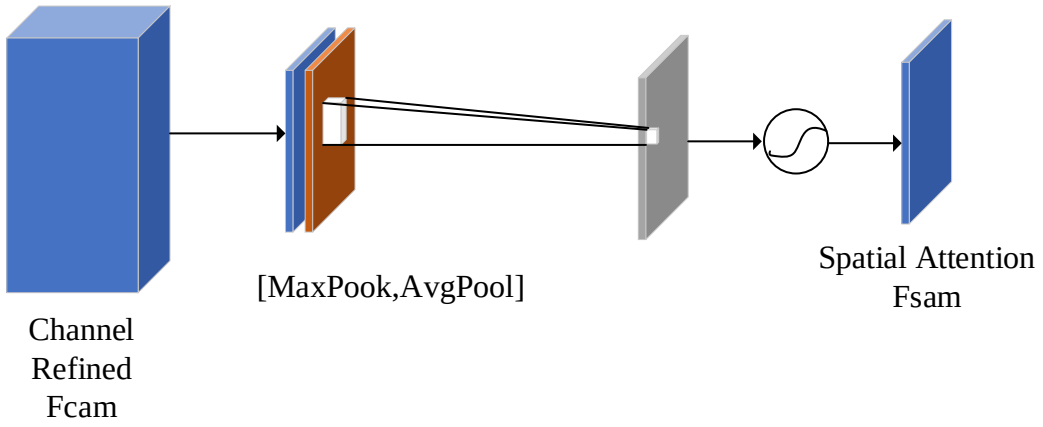

Figure 5 Spatial attention mechanism

3.4 Detail enhancement module (DEM)

The shallow features generated by the encoder are rich in spatial details (such as color, texture, and edges) within the OCT image scene but are deficient in semantic content expression. In contrast, deep features from the decoder excel in semantic expression but have lower spatial resolution and are prone to ignoring minute details. To effectively resolve the information coordination problem during their fusion, this paper proposes the detail enhancement module (DEM). Its structure is shown in Figure 6 and consists of a channel attention branch and a context attention branch. For the channel branch, adaptive average pooling is first used to compute the global average features across the channel dimension. Two 1 × 1 convolutions and nonlinear activation functions (ReLU + Sigmoid) are subsequently applied to recalibrate the interdependencies between channels. The resulting attention weights reflect the relative importance of features across channels, enhancing feature discriminative power. Finally, the input features are weighted via these attention weights and fused via a residual connection. The context branch extracts features through global pooling and convolution operations. Then, depthwise separable convolutions are applied separately along the horizontal and vertical directions to extract features. Horizontal convolution can effectively capture linear features such as cars and roads, whereas vertical convolution helps identify vertical features such as buildings and vegetation. This bidirectional feature extraction is used to improve the model's sensitivity to object edges and shapes. The extracted features are passed through a sigmoid function to generate attention weights, which are then fused with the original input features, enhancing the representation of detail features in the spatial dimension. The features generated by both branches are further fused through summation. Finally, a 1 × 1 convolution and upsampling are applied to generate the final segmentation map.

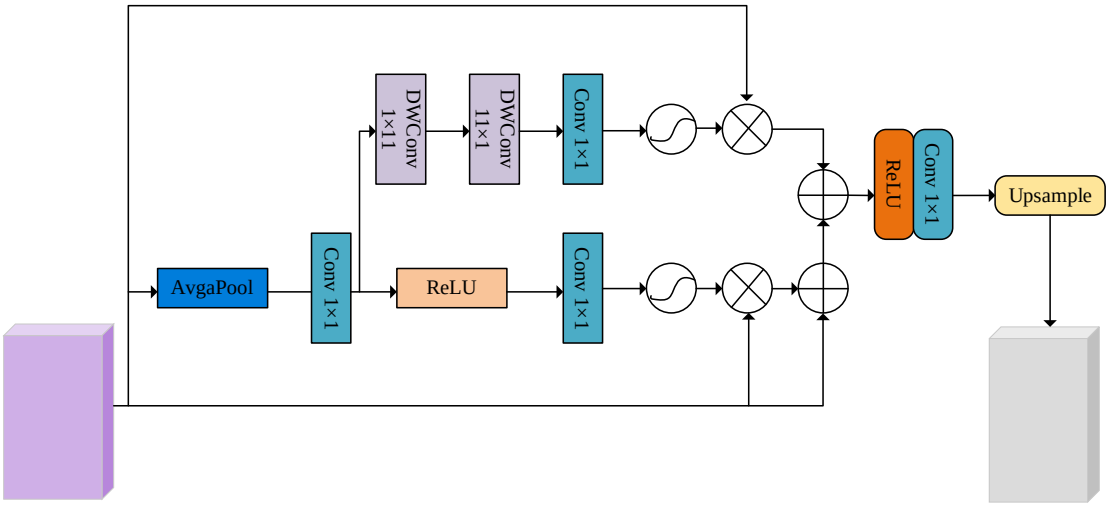

Figure 6 Detail enhancement module (DEM)

3.5 Evaluation Metrics

To validate the effectiveness of the RTHNet model, widely used evaluation metrics in the field of medical image segmentation, including the F1 score, mean intersection over union (mIoU), precision, and recall, were selected. These metrics are quantified on the basis of key parameters: true positives (TP), false positives (FP), and false negatives (FN). The mIoU reflects the degree of match between the prediction results and the ground truth segmentation. The precision reflects the model's ability to identify positive class samples accurately. The F1 score, as the harmonic mean of precision and recall, provides a comprehensive assessment of the model's overall performance.

$$IoU_k = \frac{TP}{TP + FP + FN} \tag{14}$$

$$mIoU = \frac{1}{K+1}\sum_{i=0}^{K}\frac{TP}{TP + FP + FN} \tag{15}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

4 Results

4.1 Detection Results on Retinal OCT2017 Dataset

This study selected a total of 3000 samples. The ratio of the training set, validation set, and test set was 8:1:1. The running results are shown in Table 1. Table 1 shows the classification results on the OCT2017 dataset.

Table 1 Classification Results on the OCT2017 Dataset

Evaluation Metric	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Value	94.78	95.37	93.28	94.31

Table 1 shows the classification performance results of the model on the OCT2017 dataset. The model's accuracy is 94.78%, indicating a high overall classification correctness rate for the samples; the precision is 95.37%, indicating good accuracy in samples predicted as positive classes; the recall reaches 93.28%, reflecting the model's effectiveness in identifying actual positive

class samples; and the F1 score is 94.31%, combining the performance of precision and recall, further validating the balanced nature of the model's classification effectiveness. All four metric values are above 93%, indicating stable classification performance on this dataset with relatively close performance across metrics. *(Note: The text description under Table 1 in the original (accuracy 81.27%, etc.) contradicts the values shown in Table 1 (94.78%, etc.). The translation prioritizes the table values as the primary result presentation. The descriptive text below Table 1 is translated verbatim but clearly conflicts with the table.)*

4.2 Detection Results on the OCTAMNIST Dataset

The detection results on the OCTAMNIST dataset are shown in Table 2.

Table 2 Classification Results on the OCTAMNIST Dataset

Evaluation Metric	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Value	83.41	82.49	84.71	83.65

From Table 2, the model's accuracy reached 83.41%, indicating that its overall classification correctness rate for samples is slightly higher than the performance shown in Table 1 (OCT2017). Precision is 82.49%, indicating that the model maintains high accuracy in samples predicted as positive classes but slightly lower than the precision in Table 1. Recall is 84.71%, the highest value among the four metrics, indicating a strong ability to identify actual positive class samples. The F1 score is 83.65%, balancing precision and recall, further reflecting the robustness of the model's classification effectiveness. All four metric values are between 82% and 85%, with a more concentrated overall distribution, indicating balanced and slightly improved performance trends on the OCTAMNIST dataset.

5 Discussion

5.1 Comparison with other methods

To demonstrate the advantages of the method proposed in this study, it was compared with other methods in this field. The results are shown in Table 3.

Table 3 Comparison of the proposed method with other methods in the field

Literature	Method	Accuracy (%)
[13]	CNN & Transformer Fusion for Effusion Segmentation	85.51
[14]	CNN	87.5
[15]	VGG16	93.45
[16]	DCNN	94.00%
This work	Coupled ResNet and Transformer	94.78

From Table 3, the ResNet+Transformer hybrid model proposed in this study yields a new performance record with 94.78% accuracy in the image effusion segmentation task, validating the effectiveness of fusing residual structures and attention mechanisms.

5.2 Ablation Study

To demonstrate the effectiveness of the modules added to the model of this study, an ablation study was conducted for verification. The details are shown in Table 4.

Table 4 Ablation study comparison

Model	Accuracy(%)	Precision (%)
Resnet	91.45	90.84
Transformer	92.81	93.91
Coupled ResNet & Transformer	94.78	95.23

The ablation study comparison in Table 4 shows significant performance differences among the three models. The ResNet model achieved 91.45% accuracy and 90.84% precision. The transformer model performed better, with an accuracy of 92.81% and a precision of 93.91%. The fusion model coupling ResNet and Transformer achieved the best performance, with the accuracy and

precision significantly increasing to 94.78% and 95.23%, respectively. This finding indicates that combining the two architectures effectively enhances the model's classification ability.

6 Conclusion

The RTHNet model proposed in this study, which couples ResNet and Transformer, effectively fuses local features and global semantic information through its innovatively designed attention adaptive fusion module (AAFM), global-local contextual transformer block (GLCTB), and detail enhancement module (DEM), significantly improving the classification accuracy of retinal vascular pathology OCT images. Experiments on the OCT2017 and OCTAMNIST datasets revealed that the model achieved classification accuracies of 94.78% and 83.41%, respectively, outperforming traditional U-Net, YOLOv5, and comparative methods such as CNN and VGG16 from the literature (e.g., a 1.33% improvement over VGG16). This model overcomes the subjectivity and inefficiency bottlenecks of traditional manual interpretation, achieving high-precision automatic identification and early warning of retinal vascular pathologies. It provides an intelligent auxiliary diagnostic tool for clinical practice, effectively reducing misdiagnosis rates and promoting the automation and precision development of ophthalmic disease screening.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] SOMMER A, TAYLOR H R, RAVILLA TD, et al. Challenges of ophthalmic care in the developing world [J]. Jama Ophthalmology, 2014, 132(5): 640-644.
[2] Miao Chunxu. Value of ultrawidefield laser scanning ophthalmoscopy in early screening of retinal diseases[J]. China Medical Device Information, 2023, 29(20): 122-124. DOI:10.15971/j.cnki.cmdi.2023.20.030. (Chinese)
[3] Zhang Shiwei, Su Jiahui. Diagnostic value of fundus photography and fundus fluorescein angiography in diabetic retinopathy[J]. Knowledge of Cardiovascular Disease Prevention and Treatment, 2024, 14(14): 34-36+40. (Chinese)
[4] Zhang Jialing, Cai Yan. Research progress on using artificial intelligence to discover glaucoma in fundus photographs[J]. Modern Medicine & Health, 2025, 41(01): 222-226. (Chinese)
[5] Hwang D D J, Choi S, Ko J, et al. Distinguishing retinal angiomatous proliferation from polypoidal choroidal vasculopathy with a deep neural network based on optical coherence tomography[J]. Scientific Reports, 2021, 11(1): 9275.
[6] Hassan E, Elmougy S, Ibraheem M R, et al. Enhanced deep learning model for classification of retinal optical coherence tomography images[J]. Sensors, 2023, 23(12): 5393.
[7] SV A, Raman R. Automatic Identification and Severity Classification of Retinal Biomarkers in SD-OCT Using Dilated Depthwise Separable Convolution ResNet with SVM Classifier[J]. Current Eye Research, 2024, 49(5): 513-523.
[8] Zhou Tao, Dong Yali, Huo Bingqiang, et al. Review of U-Net Network Applications in Medical Image Segmentation[J]. Journal of Image and Graphics, 2021, 26(9): 2058-2077. (Chinese)
[9] Jin Yan, Xue Zhizhong, Jiang Zhiwei. Medical Image Segmentation Algorithm Based on Recurrent Residual Convolutional Neural Network[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(8): 1205-1215. (Chinese)
[10] Wu Tong, Hu Haoji, Feng Yang, et al. Application of Segment Anything Model (SAM) in Medical Image Segmentation[J]. Chinese Journal of Lasers, 2024, 51(21): 2107102-2107102-16. (Chinese)
[11] Deng Erqiang, Qin Zhen, Zhu Guosong. EnGAN: Enhanced Generative Adversarial Network for Medical Image Segmentation[J]. Application Research of Computers/Jisuanji Yingyong Yanjiu, 2024, 41(7). (Chinese)
[12] Kermany D S, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning[J]. Cell, 2018, 172(5): 1122-1131. e9.
[13] Chen Yuyang, Li Feng. Retinal OCT Image Effusion Segmentation Method Fusing CNN and Transformer[J]. Electronic Science & Technology, 2025, 38(3). (Chinese)
[14] Karthik K, Mahadevappa M. Convolution neural networks for optical coherence tomography (OCT) image classification[J]. Biomedical Signal Processing and Control, 2023, 79: 104176.
[15] Abirami M S, Vennila B, Suganthi K, et al. Detection of choroidal neovascularization (CNV) in retina OCT images using VGG16 and DenseNet CNN[J]. Wireless Personal Communications, 2022: 1-15.
[16] Mojahed D, Ha R S, Chang P, et al. Fully automated postlumpectomy breast margin assessment utilizing convolutional neural network based optical coherence tomography image classification method[J]. Academic Radiology, 2020, 27(5): e81-e86.