Journal of Imaging Science and Technology® 69(6): 1–11, 2025.

A Multiscale Attention Feature based Transformer–Residual Combined Network for Retinal Vessel Segmentation

Mingwei Zhang, Lixian Shi, and Xiaoyan Zhang
Department of Ophthalmology, Shouguang People's Hospital, Shouguang, Shandong 262799, China

Yonghua Zhan
School of Life Science and Technology, Xidian University, Xi'an, Shaanxi 710126, China

Getao Du
School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China
E-mail: gtdu@xupt.edu.cn

Abstract. *Accurate segmentation and recognition of retinal vessels is a very important medical image analysis technique, which enables clinicians to precisely locate and identify vessels and other tissues in fundus images. However, there are two problems with most existing U-net-based vessel segmentation models. The first is that retinal vessels have very low contrast with the image background, resulting in the loss of much detailed information. The second is that the complex curvature patterns of capillaries result in models that cannot accurately capture the continuity and coherence of the vessels. To solve these two problems, we propose a joint Transformer–Residual network based on a multiscale attention feature (MSAF) mechanism to effectively segment retinal vessels (MATR-Net). In MATR-Net, the convolutional layer in U-net is replaced with a Residual module and a dual encoder branch composed with Transformer to effectively capture the local information and global contextual information of retinal vessels. In addition, an MSAF module is proposed in the encoder part of this paper. By combining features of different scales to obtain more detailed pixels lost due to the pooling layer, the segmentation model effectively improves the feature extraction ability for capillaries with complex curvature patterns and accurately captures the continuity of vessels. To validate the effectiveness of MATR-Net, this study conducts comprehensive experiments on the DRIVE and STARE datasets and compares it with state-of-the-art deep learning models. The results show that MATR-Net exhibits excellent segmentation performance with Dice similarity coefficient and Precision of 84.57%, 80.78%, 84.18%, and 80.99% on DRIVE and STARE, respectively.*
Keywords: *retinal vessels, fundus images, segmentation, Transformer–Residual, multiscale attention feature mechanism*

[DOI: 10.2352/J.ImagingSci.Technol.2025.69.6.060502]

1. INTRODUCTIONS

The retina is one of the most important tissues in the eye with a rich vessel network that can reflect early pathological changes caused by many diseases (e.g., diabetes retinopathy, macular degeneration) [1–3]. If these diseases cannot be detected and treated in a timely manner, it will bring about great health hazards to patients. Therefore, the analysis of vessel characteristics plays a crucial role in the early diagnosis of ophthalmic diseases. At present, in clinical diagnosis, doctors use artificial visual recognition of retinal vessels in fundus images to detect and diagnose various ophthalmic diseases. However, this manual tracking method for individual retinal vessels is relatively subjective and time-consuming, and fundamentally cannot guarantee the accuracy of segmentation. Therefore, developing a fast and accurate algorithm for segmenting retinal vessels is crucial to achieving precise extraction of vessel features from fundus images. However, the fundus images of retinal vessels have limitations. On the one hand, the contrast difference between the retinal vessel area and background noise is small, especially for capillaries, which increases the difficulty of segmentation [4, 5]. On the other hand, due to the complex curvature of retinal vessels, it is prone to discontinuous segmentation, which makes it impossible to completely segment retinal vessels. To address these limitations, many studies have been conducted so far. For example, Memari et al. proposed an automatic retinal vessel segmentation method that combines matching filtering technology with the AdaBoost classifier [6]. Jiang et al. used a morphology-based global thresholding method to map the structure of retinal vessels and performed centerline detection on the capillaries [7]. Although these methods have greatly improved the task of fundus image segmentation, the limitations of fundus images have not been fundamentally resolved.

In recent years, with the emergence and continuous development of deep learning (DL), it can learn the inherent rules and representation hierarchy of image data, effectively extracting representative features from the retinal vessel area and background of an image [8, 9]. Ronneberger et al. proposed a U-net network based on FCN [10] in the 2015 ISBI Cell Tracking Challenge, which applied end-to-end training to medical image segmentation and attracted widespread attention [11]. However, directly applying U-net to retinal vessel segmentation poses certain challenges, mainly due to (1) the low contrast between the vessel region and the background as indicated by the yellow arrow in Figure 1A, which

Received Nov. 3, 2024; accepted for publication Mar. 5, 2025; published online May 09, 2025. Associate Editor: Ning Yu.

1062-3701/2025/69(6)/060502/11/$25.00

can easily increase segmentation errors and reduce segmentation accuracy in fundus images; (2) the complexity of vessel curvature in fundus images (as shown in Fig. 1B). In response to these two problems, Qu et al. first proposed a global feature selection mechanism, which can autonomously select the most important features for segmentation tasks from the features of each layer of the network, thereby enhancing the segmentation ability for low contrast vessels [12]. Liu et al. proposed a feature enhancement cascade module based on Deformable Convolution v3, which can flexibly adapt to and capture the intricate and constantly changing connections of retinal vessel morphology, ensuring the continuity of vessel segmentation [13]. Although these methods have somewhat improved the accuracy of vessel segmentation, they are all centered on examining the variations in local vessel features while ignoring the significance of global features of retinal vessels, which makes it difficult to handle situations with extremely low contrast. To address this problem, the introduction of the Vision Transformer model not only provides powerful global contextual information but also demonstrates extraordinary adaptability when extensively pretrained downstream tasks [14]. However, these models frequently lose efficacy when working with small datasets like medical image datasets because of the abundance of factors and find it difficult to understand the positional information of retinal vessels. Therefore, how to capture both local and global features of vessels during the segmentation process will be the key to improving accurate segmentation of retinal vessels. Second, in response to the complex curvature of vessels, Sun et al. applied lightweight attention modules and dual attention modules in the decoder section, effectively improving the feature extraction ability of U-net for complex shaped small vessels and retinal lesion images [15]. Although the attention mechanism can effectively focus on spatial differences in images, it ignores the key differences between channels in the feature map, thereby preventing deep networks from effectively focusing on the complex features of retinal vessels.

Driven by the above challenges, we propose a combined Transformer–Residual network based on the multiscale attention feature (MSAF) mechanism, namely MATR-Net, to address the challenges of retinal vessel segmentation in fundus images. Unlike previous vessel segmentation models, the proposed MATR-Net implements a dual encoder branch that combines Residual and Transformer to simultaneously capture local and global feature information. The purpose is to use MATR-Net to learn foreground and background region features separately during the training process and then increase the weight of foreground features by suppressing the background feature response in order to segment the complete retinal vessel region from the fundus image and improve the accuracy of segmentation. We propose an MSAF module to address the complex curvature of vessels. By fusing features extracted by different convolution kernels, we capture different details and structural information of fundus images and combine attention mechanisms to explore deeper feature information and compensate for lost detail

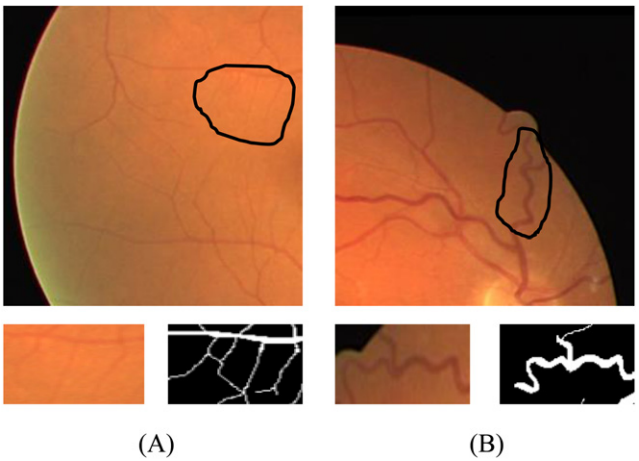

Figure 1. Examples of retinal vessel fundus image segmentation: (A) examples of low contrast between vessels and background noise in fundus images (black box marked area); (B) examples of complex vessel curvature in fundus images (black box marked area).

pixels. In addition, the joint loss function used in this study is composed of Dice loss and cross-entropy loss, and Dice loss can solve the problem of imbalanced positive and negative samples in fundus images. The main contributions of this work include the following:

1. In order to improve the feature extraction ability of the model when the contrast between vessels and background is low, we adopt a dual encoder module consisting of Transformer and Residual layers in MATR-Net. This module can combine local features of retinal vessels with global features to enhance the model's feature extraction ability in low contrast situations. Relative to the traditional convolutional neural network (CNN), the model can minimize gradient dispersion and explosion issues during training by utilizing Residual layers.
2. We propose an MSAF module based on different convolutional kernels to address the complex curvature of vessels. The MSAF module can capture features of vessels and capillaries with various complex curvature shapes, thereby improving segmentation accuracy.
3. The experimental results on two publicly available datasets show that MATR-Net achieves better segmentation performance and generalization ability with fewer model parameters compared to the most advanced retinal vessel segmentation methods currently available.

2. METHODS

This section provides a detailed introduction to the proposed MATR-Net for retinal vessel segmentation. First, we describe the overall structure of MATR-Net. Then, the key components of the network are described in detail, namely the Transformer encoder branch and the MSAF module. By using a dual encoder branch consisting of Residual and Transformer, not only can rich local feature information and important global feature information be captured but

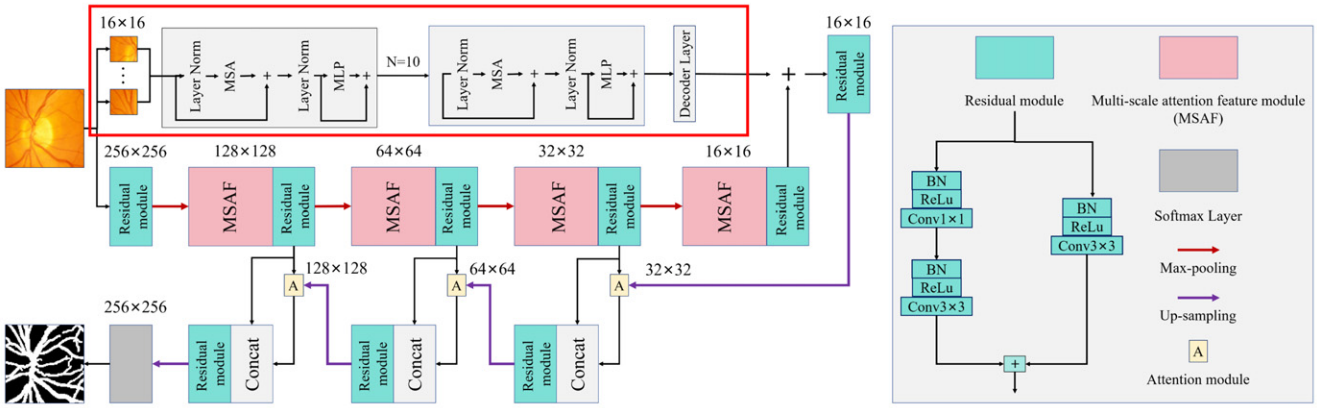

Figure 2. Segmentation network architecture of MATR-Net.

also vessel regions can be identified from low contrast images. In addition, we have developed an MSAF module and embedded it into the encoder branch to obtain deep feature information of vessels. The proposed segmentation framework is shown in Figure 2.

2.1 *Network Architecture*

Inspired by the powerful representation capabilities of the CNN and Transformer, this study proposes an end-to-end retinal vessel segmentation framework, namely MATR-Net, to accurately and reliably segment retinal vessel fundus images. The overall architecture is shown in Fig. 2. MATR-Net mainly consists of three parts: a dual encoder for enhancing feature encoding, an MSAF module for effectively extracting deep features, and a background noise suppression attention module. Specifically, by merging efficient Residual and Transformer branches into MATR-Net, rich local features and important global contextual information for retinal vessel segmentation can be extracted. Second, MSAF can effectively capture the features of vessels and capillaries with various complex curvature shapes, compensating for the loss of deep vessel feature information during down-sampling. Finally, we enhance the feature response of vessels by adopting an attention mechanism while reducing the impact of background noise on segmentation results. In the following section, we discuss in detail the proposed Transformer encoder branch and the MSAF module.

2.2 *Residual and Transformer Dual Encoder Branches*

2.2.1 *Transformer Encoder Branch*

Solving the low contrast problem of fundus images is crucial to improving the accuracy of vessel segmentation. Low contrast is a common issue in fundus image processing, which can affect the clarity of vessels and other important features, thereby affecting the accuracy of subsequent automatic segmentation and diagnosis. Therefore, we have implemented a dual encoder consisting of Residual and Transformer, which captures both local features of vessels and global contextual information. The red border in Fig. 2 shows the structure of the Transformer encoder branch, where the multihead self-attention (MSA) mechanism is the core component of the Transformer, consisting of 16 heads. First, three sequence vectors (query vector, key vector, and value vector) are calculated from the output of the previous Transformer layer. Then, the attention score is obtained by dot-multiplying the query vector with the key vector. The generated vector is normalized using a softmax activation function to ensure that all values are positive. Finally, each value vector is multiplied with the normalized vector value and the weighted value vectors are added to obtain the output vector. The MSA projects the same query vector, key vector, and value vector into different subspaces of the original high-dimensional space for self-attention calculation, and concatenates multihead with self-attention scores. The workflow of the Transformer encoder branch is as follows. We first split the fundus image with a size of $H \times W$ into multiple patches for input, where each patch has a size of $P \times P$ and the number of patches in a sequence is $N = (HW/P^2)$. Then, the vectorized patches are mapped onto a D-dimensional embedding space through trainable linear projection, and the spatial information of the patches is encoded. Finally, the encoder features are decoded using a progressive up-sampling method. To preserve its location information, the learned location information is embedded into the patch using the following formula:

$$y = [Z_P^1 E, \ldots, Z_P^N E] + E_{\text{pos}}, \tag{1}$$

where E represents the position information of each patch, Z_P represents a patch, and E_{pos} represents the position embedding information.

2.2.2 *Residual Encoder Branch*

In order to capture contextual features and preserve certain spatial details through convolutional neural networks, this study uses Residual as the backbone network of the CNN encoder. The Residual encoder branch consists of two 3×3 convolutional layers and a 1×1 convolutional layer, which are processed using batch normalization (BN) and rectified linear unit (ReLU) activation functions before the convolution operation. The Residual encoder passes gradients freely to lower layers through skip connections to

prevent gradients from vanishing or exploding. A kernel size of 3 × 3 reduces parameters and speeds up computation, which is beneficial for extracting local features. The BN operation is used to maintain the same distribution of network inputs, improve network training speed, and prevent gradient vanishing problems. Replacing traditional CNN layers with Residual layers increases the model's depth to improve segmentation accuracy while alleviating problems such as gradient vanishing and exploding caused by depth increase, ensuring good performance of the model.

2.3 *MSAF Module*

The original U-net generates multilayer feature maps through the encoder, which are transmitted to the corresponding decoder layers through skip connections, allowing the decoder to obtain more high-resolution information during up-sampling and thus more accurately restore the details in the original image. However, due to the complex curvature of retinal vessels, especially for capillaries, a large amount of detail information will be lost during down-sampling, resulting in discontinuous vessel segmentation. Therefore, we propose an MSAF module, where each layer consists of a multiscale feature fusion module and attention. The MSAF utilizes three types of convolutions—1 × 1, 3 × 3, and 5 × 5—to extract feature information at different scales and captures different details and structural information of fundus images through fusion operations. However, although multiscale processing can improve the model's ability to understand images, in actual processing of large amounts of image information, the model may be disturbed by too many details, leading to the loss or misunderstanding of key information. To address this issue, we introduce an attention mechanism in the multiscale fusion module. The specific workflow is as follows. First, the multiscale fusion features are convolved by 1 × 1 and ReLU to obtain Feature Map A. Then, Feature Map A is normalized using 1 × 1 convolution and the sigmoid function with the aim of outputting attention coefficient values in the range [0,1]. Finally, the attention coefficient is multiplied with the multiscale fusion features to enable the model to focus more on the vessel regions in fundus images, thereby improving its ability to recognize and locate vessels. The detailed structure of MSAF is shown in Figure 3.

2.4 *Loss Function*

In this study, due to the use of the overlap method for cropping fundus images, the number of negative samples in some images is much larger than that of positive samples. This problem of sample imbalance may make the training process difficult, resulting in the model's overprediction of non-vessel areas. To solve this problem, we adopt a combined loss function consisting of cross-entropy loss and Dice loss. Dice loss makes the model more inclined to explore the foreground region by taking the intersection and union of the segmentation results with the ground truth, reducing the influence of most negative pixels. The formula for the joint loss function is as follows:

$$\text{Loss} = Loss_{\text{cross}} + Loss_{\text{Dice}} \tag{2}$$

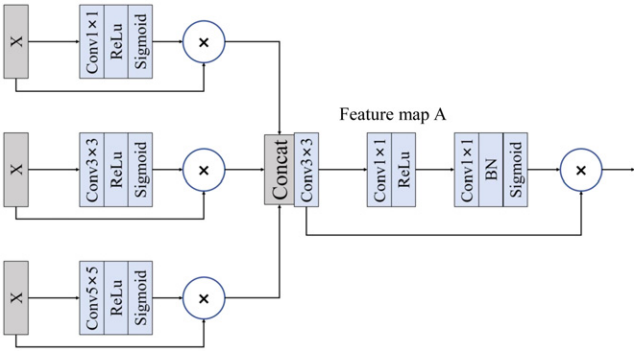

Figure 3. Structure of the MSAF module.

$$Loss_c = -y_t \log(y_{\text{pred}}) - (1 - y_t) \log(1 - y_p) \tag{3}$$

$$Loss_{\text{Dice}} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|}. \tag{4}$$

Among them, Loss represents the joint loss function, $Loss_c$ represents the cross-entropy loss, y_t is the target value, y_p is the predicted value, $Loss_{\text{Dice}}$ represents Dice loss, X is the ground truth, and Y is the segmentation result.

3. EXPERIMENTS

3.1 *Dataset*

3.1.1 *DRIVE Dataset [16]*

This dataset contains 40 retinal fundus images with a resolution of 584 × 565, and each image has a corresponding ground truth, of which 20 are used for model training and 20 are used for model testing. Ground truth is the baseline in the process of model training and evaluation.

3.1.2 *STARE dataset [17]*

This dataset contains 20 retinal fundus images with a resolution of 700 × 605 and 20 ground truth images, of which 16 are used for model training and 4 are used for model testing.

Considering the limited number of images in these two datasets, overfitting may occur during network training. To address this issue, we employed data augmentation techniques (Figure 4). Specifically, the overlap clipping method has an overlap rate of 102 pixels for the DRIVE dataset and 82 pixels for the STARE dataset; (1) random rotation within the range of [−10,10]; (2) horizontal flipping; (3) vertical flipping; (4) histogram equalization CLAHE. Table I provides detailed information on the number of images, training set, testing set, and resolution before and after cropping for each dataset.

3.2 *Training Parameters*

The MATR-Net segmentation model was developed on GeForce GTX 3090 and Intel Core i9-10900K, using PyTorch 1.8.0 as the DL framework. The optimizer uses Adam and employs a poly learning decay strategy. The initial learning rate is set to 0.004, and the batch size and epoch are set to 20 and 100, respectively. Table II provides detailed information

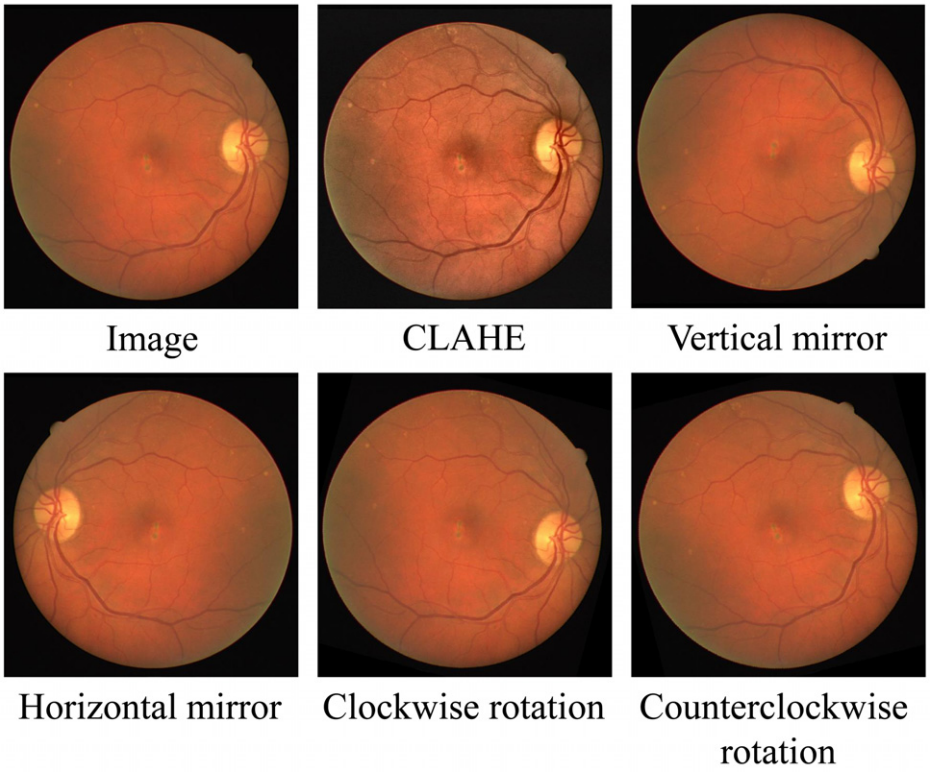

Figure 4. Examples of data amplification.

Table I. Details of the two datasets.

Datasets	DRIVE	STARE
Total number	40	20
Train set number	20	16
Test set number	20	4
Resolution (pixel)	565 × 584	605 × 700
Resized (pixel)	256 × 256	256 × 256

Table II. Hyperparameters of the proposed MATR-Net model.

Items	Value
Input size	256 × 256 × 3
Epochs	100
Initial learning rate	0.004
Batch size	20
Optimizer	Adam
Loss function	Cross-entropy loss + Dice loss

on image input size, total epochs, initial learning rate, batch size, optimizer, and loss function.

3.3 *Evaluation Metrics*

In order to comprehensively evaluate the segmentation performance of the MATR-Net proposed in this paper, we quantitatively analyzed the experimental results using four commonly used medical image segmentation evaluation metrics: Dice similarity coefficient (DSC), Intersection over Union (IoU), Precision, and Recall. The four formulas are as follows:

$$\mathrm{DSC} = \frac{2 * (X \cap Y)}{|X| + |Y|} \tag{5}$$

$$\mathrm{IoU} = \frac{X \cap Y}{X \cup Y} \tag{6}$$

$$\mathrm{Precision} = \frac{X \cap Y}{X} \tag{7}$$

$$\mathrm{Recall} = \frac{X \cap Y}{Y}, \tag{8}$$

where X represents the ground truth and Y represents the segmentation result.

4. RESULTS

4.1 *Comparisons with Other State-of-the-Art Methods*

To compare the segmentation performance of our proposed MATR-Net with other state-of-the-art network models, we trained nine segmentation models on two datasets and manually adjusted all hyperparameters for optimal performance, including TransU-net [18], U-net [11], UNeXt [19], ResU-net [20], MCDAU-Net [21], Attention U-net [22], IMFF-Net [23], DCNet [24], and MATR-Net. Then, we tested the test dataset and evaluated the segmentation performance of the model using four metrics: DSC, IoU, Precision, and Recall. Due to the lack of code provided in the original text

for Curv-Net and MCSE-U-Net models, we only referred to data from the relevant literature and used the symbol '-' to indicate that there was no available experimental data in the original literature. The specific results of the experimental comparison between the two datasets are shown in Tables III and IV.

On the DRIVE dataset, as shown in Table III, MATR-Net achieved excellent results with DSC and Precision metrics of 84.57% and 80.78%, respectively, surpassing the compared methods. Of note, our method outperforms MCSE-U-Net, Curv-Net, and DCNet in DSC metrics by 0.27%, 1.05%, and 2.58%, respectively. This indicates that our method has good recognition ability, which helps to more accurately distinguish between vessel and non-vessel areas. In addition, Precision metrics can effectively measure the accuracy of the model. From Table III, it can be seen that only MATR-Net achieved over 80%, indicating that the MATR-Net segmentation model can better distinguish between positive and negative samples.

On the STARE dataset, as shown in Table IV, MATR-Net achieved excellent results of 84.18% in DSC, which is 3.1% higher than MCSE-U-Net. This indicates that our proposed segmentation model can more accurately segment retinal vessels from fundus images and display the integrity of retinal vessels than the most advanced segmentation models currently available. In terms of Precision metrics, MATR-Net outperforms nine state-of-the-art segmentation models, achieving 80.99%. For example, compared to DCNet, MATR-Net has improved its DSC metrics by 3.66% but decreased its recall metrics by 5.17%. This indicates that improving the segmentation ability of positive sample vessel pixels is feasible while ignoring some positive sample background pixels is also necessary.

Table III. Performance comparison with the most state-of-the-art segmentation methods on the DRIVE dataset.

Method	Year	DSC (%)	IoU (%)	Precision (%)	Recall (%)
UNeXt [19]	2022	72.35	57.17	63.23	85.93
TransU-net [18]	2021	73.36	58.43	61.50	**92.36**
U-net [11]	2015	76.93	62.90	67.84	89.86
ResU-net [20]	2018	79.72	66.56	75.11	85.72
MCDAU-Net [21]	2023	80.43	67.53	74.98	87.34
Attention U-net [22]	2019	81.22	68.63	76.87	86.68
IMFF-Net [23]	2024	81.24	68.67	77.55	85.87
DCNet [24]	2024	81.99	69.72	79.96	84.88
Curv-Net [25]	2022	83.52	—	—	—
MCSE-U-Net [26]	2023	84.30	**84.73**	—	—
MATR-Net		**84.57**	73.63	**80.78**	89.19

Table IV. Performance comparison with the most state-of-the-art segmentation methods on the STARE dataset.

Method	Year	DSC (%)	IoU (%)	Precision (%)	Recall (%)
U-net [11]	2015	75.36	60.91	70.40	82.07
UNeXt [19]	2022	76.29	62.14	71.44	82.59
TransU-net [18]	2021	77.36	63.44	68.01	90.60
Attention U-net [22]	2019	78.33	64.73	71.70	87.19
MCDAU-Net [21]	2023	79.04	65.61	75,63	83.40
ResU-net [20]	2018	79.73	66.60	75.44	85.41
DCNet [24]	2024	80.52	67.82	71.52	**93.16**
IMFF-Net [23]	2024	80.73	67.95	76.39	86.27
MCSE-U-Net [26]	2023	81.08	**80,81**	—	—
MATR-Net		**84.18**	73.19	**80.99**	87.99

4.2 *Visual Comparisons with Other State-of-the-Art Methods*

In order to demonstrate the segmentation performance of the proposed MATR-Net model, this study presents five segmentation models for visual comparison with MATR-Net. We have selected the five most representative segmentation models, namely DCNet, TransU-net, UNeXt, U-net, and MCDAU-Net. Among them, TransU-net is a segmentation model based on Transformer, MCDAU-Net and DCNet are the best performing models in the field of retinal vessel segmentation for the past two years, and UNeXt and U-net are typical medical image segmentation methods. In these two datasets, we use red borders to mark the local segmentation results of vessels and enlarge the local images for display. In Figures 5 and 6, the five segmentation models exhibit issues such as incompleteness, discontinuity, and omission in capillaries while MATR-Net can accurately segment the capillary region. In addition, when dealing with densely distributed areas of vessels, other methods may encounter problems of missing elements or oversegmentation, such as TransU-net missing many vessel regions on the DRIVE dataset and U-net identifying background regions as vessel regions on the STARE dataset, both of which are erroneous segmentation. However, compared to these five segmentation models, MATR-Net focuses more on the feature response of vessel regions and suppresses the influence of background noise on segmentation results. From the last column of Figs. 5 and 6, it can be seen that the proposed method exhibits significant advantages in terms of vessel segmentation performance. It not only has high accuracy but also can capture the details of small vessels more comprehensively, providing us with more accurate and reliable vessel segmentation results. Overall, compared to comparative methods, MATR-Net demonstrates excellent segmentation performance for capillaries and large vessels.

4.3 *Ablation Experiments*

To demonstrate the effectiveness of each module in the MATR-Net network, this study conducted experiments using different combinations of MATR-Net. All ablation experiments were conducted on the DRIVE dataset and STARE dataset. Tables V and VI show the results of comparing our proposed method with models containing different components.

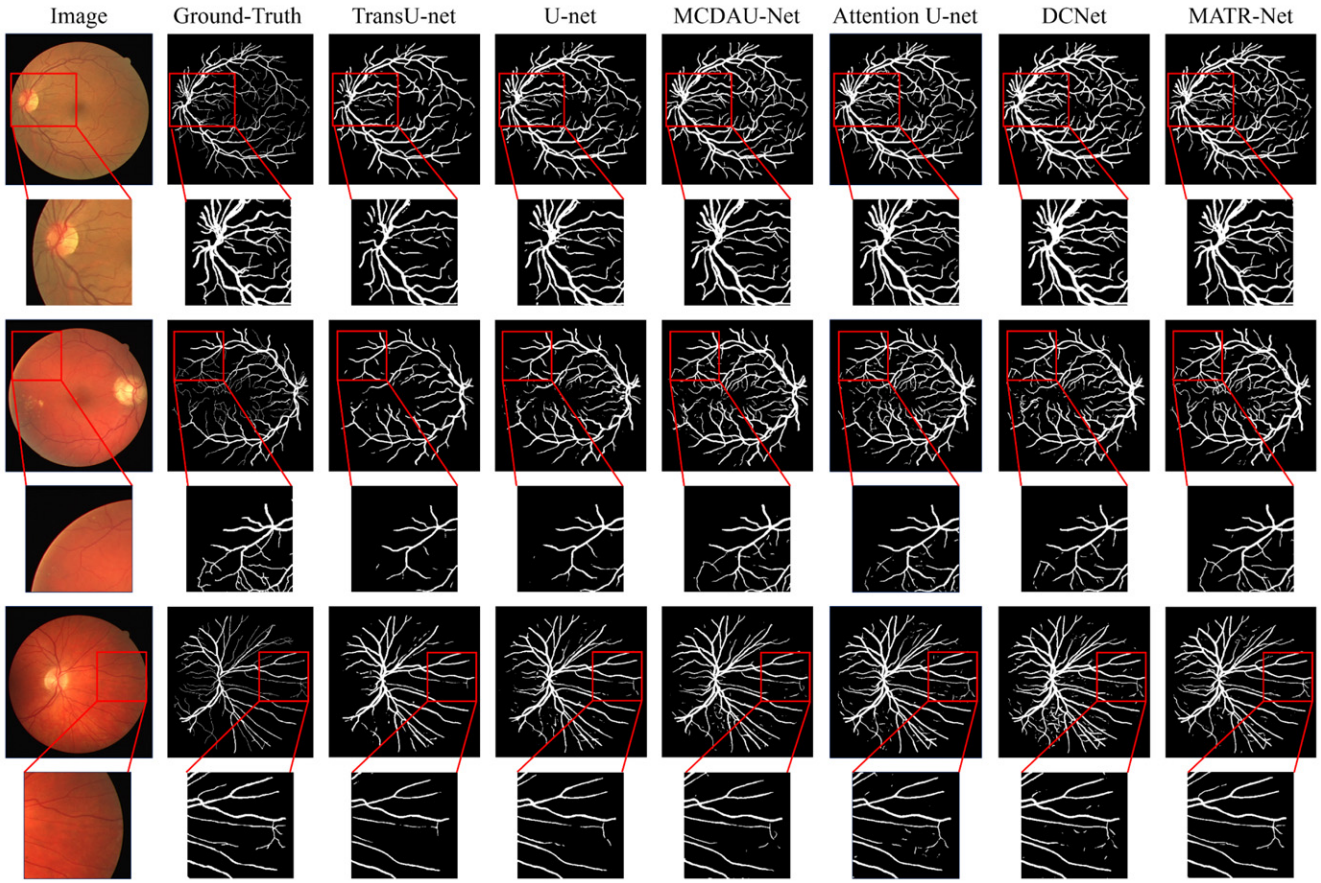

Figure 5. Visual comparison of segmentation results with different state-of-the-art segmentation methods on DRIVE dataset.

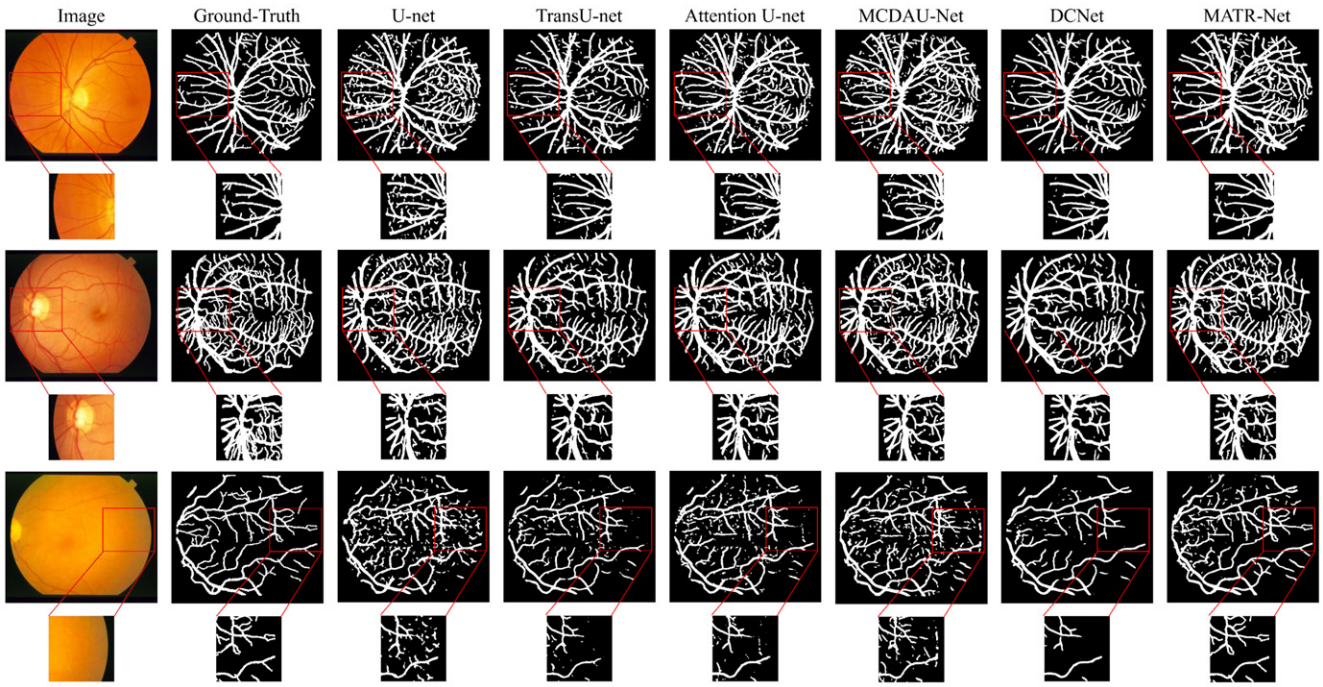

Figure 6. Visual comparison of segmentation results with different state-of-the-art segmentation methods on STARE dataset.

Table V. Quantitative results of different components on the DRIVE dataset.

Method	DSC (%)	IoU (%)	Precision (%)	Recall (%)
Our w/o Attention	82.88	71.39	78.17	88.72
Our w/o Transformer	83.24	71.71	79.52	87.79
Our w/o MSAF module	83.53	72.12	79.72	88.26
MATR-Net	**84.57**	**73.63**	**80.78**	**89.19**

Table VI. Quantitative results of different components on the STARE dataset.

Method	DSC (%)	IoU (%)	Precision (%)	Recall (%)
Our w/o MSAF module	80.78	68.15	73.36	**90.54**
Our w/o Transformer	82.03	69.84	76.26	89.37
Our w/o Attention	83.82	72.34	80.12	88.19
MATR-Net	**84.18**	**73.19**	**80.99**	87.99

4.3.1 *Ablation Experiments for Transformer*

To verify the effectiveness of the Transformer branch, we removed it from MATR-Net. Combining Tables V and VI, it can be seen that compared to MATR-Net, the DSC and IoU of the Transformer module on the DRIVE dataset decreased by 1.33% and 1.92%, respectively, while on the STARE dataset, the DSC and IoU decreased by 2.15% and 3.35%, respectively. This indicates that a single Residual encoder branch does not have the ability to capture global contextual information while the dual encoder branch composed of Transformer and Residual modules can simultaneously consider both local and global features of vessels. By integrating these two types of feature information, the problem of low contrast in fundus images can be solved to obtain the best segmentation results. In order to demonstrate the effectiveness of Transformer more intuitively, it can be seen from the red boxes in Figure 7 that segmentation models without Transformer cannot extract complete vessel regions under low contrast conditions. However, MATR-Net can effectively solve the problem of low contrast in fundus images.

4.3.2 *Ablation Experiments for Attention*

In order to better reduce the impact of non-vessel areas on segmentation results, a feature attention module is added to the skip connection of each layer in the MATR-Net

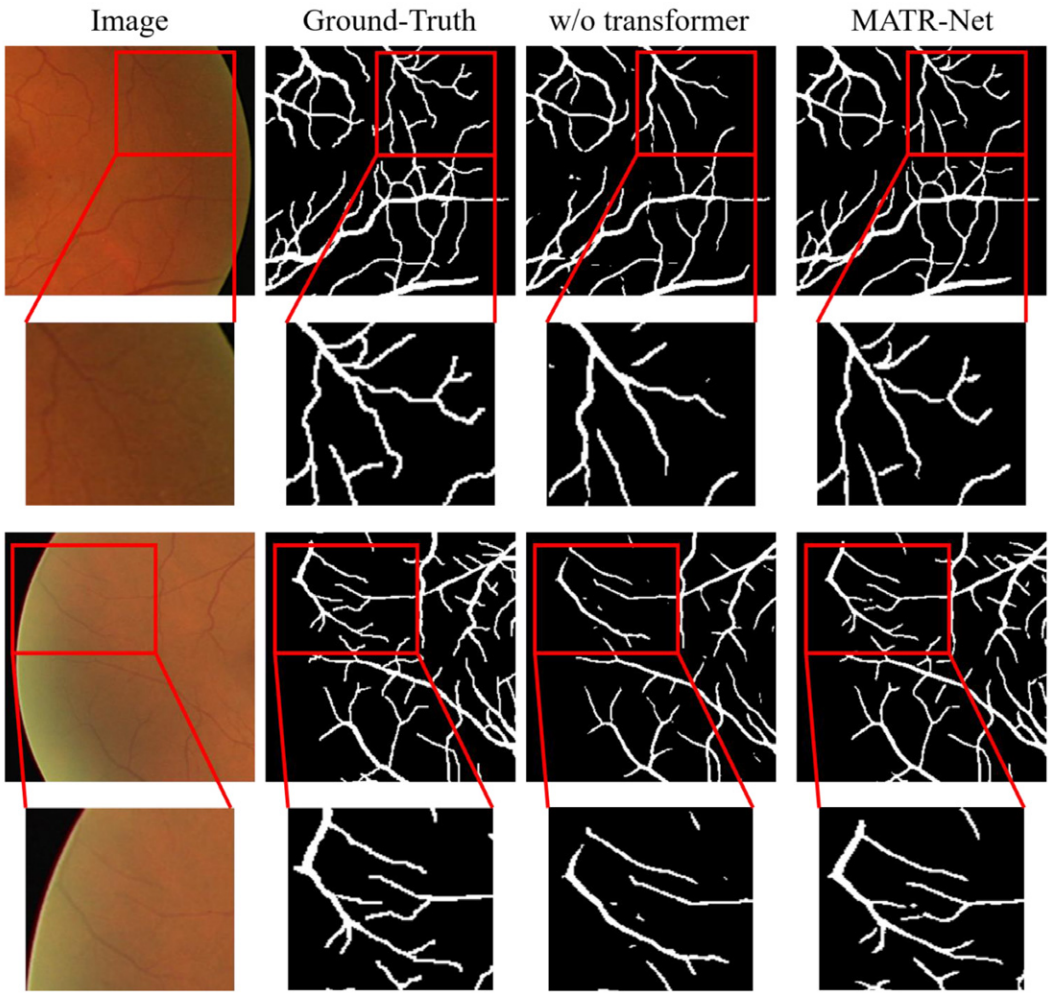

Figure 7. Visual comparison of the impact of segmentation results without and with Transformer modules.

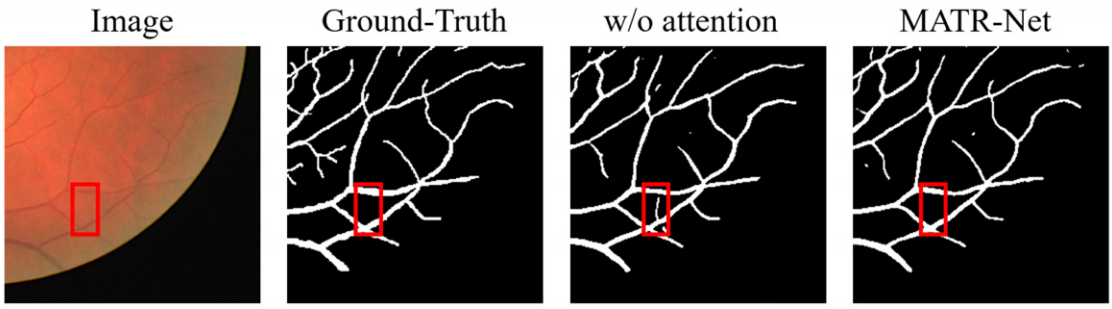

Figure 8. Visual comparison of the impact of segmentation results without and with attention modules.

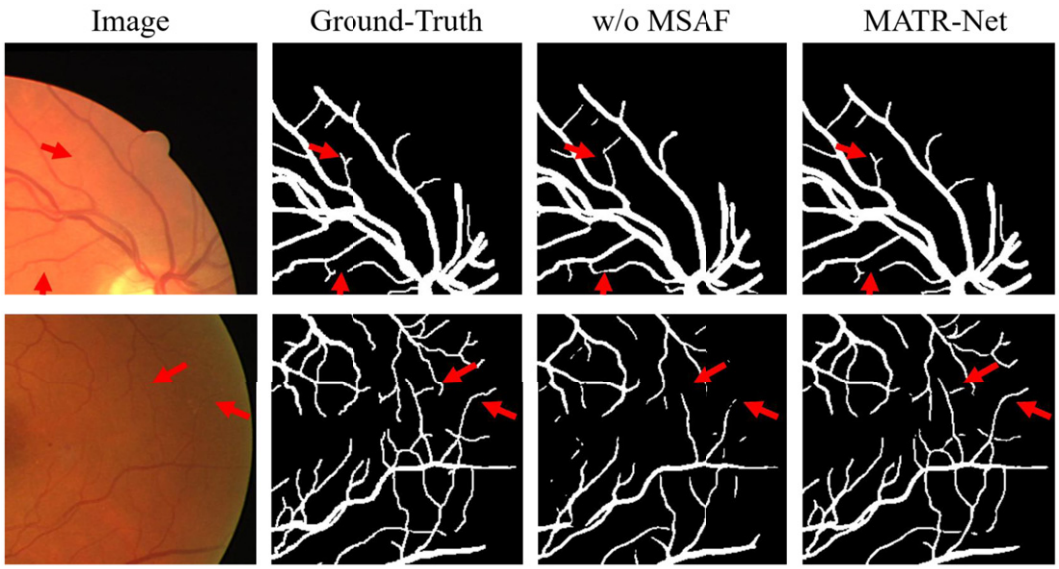

Figure 9. Visual comparison of the impact of segmentation results without and with MSAF modules.

model. To further validate its performance, we compared the impact of using and not using the feature attention module on segmentation performance on the DRIVE and STARE datasets separately. Tables V and VI provide quantitative comparisons of these variables, from which it can be seen that adding the feature attention module does indeed improve segmentation performance, with DSC increasing from 82.88% and 83.82% to 84.57% and 84.18%, respectively. This indicates that attention can improve the segmentation performance of vessels by significantly highlighting the feature response of vessel regions. In order to demonstrate the effectiveness of the attention module more intuitively, it can be seen from Figure 8 that not adding the attention module will lead to incorrect segmentation, for example, identifying the background area as a blood vessel in the red box while adding the attention module can better suppress the feature response of background noise and reduce the false positive rate of segmentation.

4.3.3 *Ablation Experiments for MSAF*

As shown in Tables V and VI, we compared the quantitative results of using and not using the feature MSAF module on two datasets separately. It can be seen from this that adding the feature MSAF module has indeed improved segmentation performance, with DSC increasing from 83.53% and 80.78% to 84.57% and 84.18%, respectively. This indicates that the MSAF module can effectively compensate for the lost feature information during the down-sampling process, thereby solving the problem of discontinuous vessel segmentation. In addition, in order to more intuitively demonstrate the ability of the MSAF module in solving complex vessel curvature problems, it can be seen from the red arrows in Figure 9 that not using the MSAF module will result in discontinuous or even missing segmented vascular regions. Adding the MSAF module can accurately identify very small vascular feature information, improving the integrity of vascular segmentation.

5. DISCUSSION

The changes in the retinal vessel system are often closely related to various diseases. If not detected and treated in a timely manner, it may develop into more serious lesions and even lead to blindness. Therefore, accurate analysis of retinal vessel characteristics is crucial for the early diagnosis of ophthalmic diseases. In recent years, a large amount of work has been conducted on the issue of vessel segmentation.

However, due to the small contrast difference between vessel regions and background noise and the complex curvature of vessels in retinal vessel fundus images, current research methods have certain limitations. Therefore, establishing new effective vessel segmentation methods is crucial to accurately evaluating ophthalmic diseases. In this work, we propose the MATR-Net method for fully automatic vessel segmentation and fully demonstrate that the segmentation performance of the proposed model is superior to the state-of-the-art methods on the DRIVE and STARE datasets. From Tables III and IV, it can be seen that MATR-Net has improved the results of automatic vessel segmentation, with DSC and Precision being 84.57%, 80.78%, 84.18%, and 80.99%, respectively.

During the experimental process, this study found that the small contrast difference between the vessel region and background noise in fundus images, as well as the complex curvature of vessels, can pose bottlenecks to the training of DL models. Although the single encoder structure based on the CNN has been widely used in medical image segmentation, this structure cannot capture both local features and global contextual information features simultaneously. The use of a CNN-based single encoder structure to simultaneously extract local features and global long-range dependencies is limited in addressing the irregular changes in vessel shape and low contrast between vessels and background noise in complex fundus images. Currently, existing methods still cannot accurately distinguish the differences between targets and backgrounds. In view of this, we have designed a dual encoder input branch that combines Transformer and the CNN to simultaneously capture local features and global contextual information. At the same time, replacing the convolutional layers in the CNN with Residual ensures the avoidance of overfitting during model training. As shown in Tables V and VI, this paper uses four evaluation metrics to verify the impact of Transformer encoding branches on MATR-Net on the DRIVE and STARE datasets. In the case of removing the Transformer encoder branch, the DSC index of MATR-Net decreased by 1.33% and 2.15%, respectively, indicating that adding the Transformer encoder branch to MATR-Net can accurately identify the feature information of vessels in low contrast situations. In order to further obtain more vessel feature information, we added an Attention module before feature concatenation in each layer of the encoder and decoder to better improve the sensitivity of the model to foreground pixels and suppress background noise. As shown in Tables V and VI, we validated its impact on MATR-Net through four evaluation metrics, and the results showed that introducing the feature attention module enhanced the segmentation performance. Due to the complexity of vessel curvature, especially for very small vessels, a large amount of vessel detail information is lost, resulting in discontinuous vessel segmentation. Therefore, we propose an MSAF module to compensate for the loss of detailed information of vessels during the pooling process. This module captures different details and structural information of fundus images by fusing features extracted by different convolution kernels and combines attention mechanisms to explore deeper feature information and compensate for lost detail pixels. As shown in Fig. 9, after adding the MSAF module, the model significantly improves the continuity of vessel segmentation and can completely segment very small vessel regions.

Although the MATR-Net architecture can achieve good results in vessel segmentation, it has also certain limitations. Although MATR-Net has successfully segmented finer vessels, there are still cases of vessel rupture in the face of extremely low contrast. Therefore, we will continue to optimize the segmentation model to effectively solve the problem of vessel rupture in future work. In addition, we will attempt to validate the segmentation performance of the MATR-Net method in other retinal vessel fundus image segmentation tasks to further confirm its effectiveness and generalization.

6. CONCLUSIONS

In this paper, we propose a MATR-Net segmentation architecture for segmenting retinal vessels in fundus images. Unlike the segmentation model based on the CNN single encoder, we use a dual encoder branch combining Residual and Transformer for image segmentation, which enables our method to capture both local and global contextual information simultaneously. In addition, we also adopted a more effective MSAF module to solve the problem of complex vessel curvature, enabling the segmentation of complete and continuous large and small vessels. To evaluate our proposed retinal vessel segmentation method, this study conducted experiments on DRIVE and STARE datasets. Compared with state-of-the-art segmentation models, the effectiveness of MATR-Net has been demonstrated.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Basic Research Program of Shaanxi (Program No. 2024JC-ZDXM-47) and Xidian University Specially Funded Project for Interdisciplinary Exploration (No. TZJH2024029).

REFERENCES

[1] Z. Lin, Y. Wang, D. Li, L. Wen, G. Zhai, X. X. Ding, D. X. Zang, F. H. Wang, and Y. B. Liang, "Higher prevalence of diabetic retinopathy among female Chinese diabetic patients with metabolic syndrome," Japan. J. Ophthalmol. **66**, 102–109 (2022).

[2] T. R. P. Taylor, M. J. Menten, D. Rueckert, S. Sivaprasad, and A. J. Lotery, "The role of the retinal vasculature in age-related macular degeneration: a spotlight on OCTA," Eye **38**, 442–449 (2023).

[3] K. Zhang, L. F. Zhang, and R. N. Weinreb, "Ophthalmic drug discovery: novel targets and mechanisms for retinal diseases and glaucoma," Nat. Rev. Drug Discov. **11**, 541–559 (2012).

[4] U. T. V. Nguyen, A. Bhuiyan, L. A. F. Park, and K. Ramamohanarao, "An effective retinal blood vessel segmentation method using multi-scale line detection," Pattern Recognit. **46**, 703–715 (2013).

[5] S. Thangaraj, V. Periyasamy, and R. Balaji, "Retinal vessel segmentation using neural network," IET Image Process. **12**, 669–678 (2018).

[6] N. Memari, A. R. Ramli, M. I. Bin Saripan, S. Mashohor, and M. Moghbel, "Supervised retinal vessel segmentation from color fundus images based on matched filtering and AdaBoost classifier," PLoS One **12**, e0188939 (2017).

[7] Z. X. Jiang, J. Yepez, S. An, and S. Ko, "Fast, accurate and robust retinal vessel segmentation system," Biocybern. Biomed. Eng. **37**, 412–421 (2017).

[8] C. Wang, Z. Y. Zhao, Q. Q. Ren, Y. T. Xu, and Y. Yu, "Dense U-net based on patch-based learning for retinal vessel segmentation," Entropy **21**, 168 (2019).
[9] Y. Tang, Z. Y. Rui, C. F. Yan, J. J. Li, and J. P. Hu, "ResWnet for retinal small vessel segmentation," IEEE Access **8**, 198265–198274 (2020).
[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 640–651 (2017).
[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *The Processing of Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (Springer, Munich, Germany, 2015), pp. 234–241.
[12] Z. W. Qu, L. Zhuo, J. Cao, X. G. Li, H. X. Yin, and Z. C. Wang, "TP-Net: Two-path network for retinal vessel segmentation," IEEE J. Biomed. Health Inform. **27**, 1979–1990 (2023).
[13] J. H. Liu, D. X. Zhao, J. C. Shen, P. Geng, Y. Zhang, J. X. Yang, and Z. Q. Zhang, "HRD-Net: High resolution segmentation network with adaptive learning ability of retinal vessel features," Comput. Biol. Med. **173**, 108295 (2024).
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: transformers for image recognition at scale," Preprint, arXiv:2010.11929 (2020).
[15] K. Sun, Y. Chen, Y. Chao, J. M. Geng, and Y. S. Chen, "A retinal vessel segmentation method based improved U-Net model," Biomed. Signal Process. Control **82**, 104574 (2023).
[16] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retinal," IEEE Trans. Med. Imag. **23**, 501–509 (2004).
[17] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imag. **19**, 203–210 (2000).
[18] J. N. Chen, Y. Y. Lu, Q. H. Yu, X. D. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Y. Zhou, "TransUNet: transformers make strong encoders for medical image segmentation," Preprint arXiv:2102.04306 (2021).
[19] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," *The Processing of Medical Image Computing and Computer Assisted Intervention–MICCAI 2022* (Springer, Singapore, 2022), pp. 23–33.
[20] Z. X. Zhang, Q. J. Liu, and Y. H. Wang, "Road extraction by deep residual U-Net," IEEE Geosci. Remote Sens. Lett. **15**, 749–753 (2018).
[21] W. Zhou, W. Q. Bai, J. H. Ji, Y. G. Yi, N. Y. Zhang, and W. Cui, "Dual-path multi-scale context dense aggregation network for retinal vessel segmentation," Comput. Biol. Med. **164**, 107269 (2023).
[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: learning where to look for the pancreas," *The Processing of Medical Imaging with Deep Learning* (Springer, Amsterdam, Netherlands, 2018).
[23] M. T. Liu, Y. Y. Wang, L. Wang, S. B. Hu, X. Wang, and Q. M. Ge, "IMFF-Net: An integrated multi-scale feature fusion network for accurate retinal vessel segmentation from fundus images," Biomed. Signal Process. Control **91**, 105980 (2024).
[24] Z. H. Shang, C. H. Yu, H. Huang, and R. X. Li, "DCNet: A light-weight retinal vessel segmentation network," Digital Signal Process. **153**, 104651 (2024).
[25] Y. L. He, H. Sun, Y. G. Yi, W. H. Chen, J. Kong, and C. X. Zheng, "Curv-Net: Curvilinear structure segmentation network based on selective kernel and multi-Bi-ConvLSTM," Med. Phys. **49**, 3144–3158 (2022).
[26] L. H. Zhang, C. X. Xu, Y. Z. Li, T. Liu, and J. D. Sun, "MCSE-U-Net: multi-convolution blocks and squeeze and excitation blocks for vessel segmentation," Quant. Imag. Med. Surg. **14**, 2426–2440 (2024).