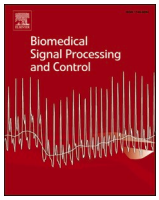
Biomedical Signal Processing and Control 83 (2023) 104604

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Transformer and convolutional based dual branch network for retinal vessel segmentation in OCTA images

Xiaoming Liu [a,b,*], Di Zhang [a,b], Junping Yao [c], Jinshan Tang [d]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China
[c] Tianyou Hospital Affiliate to Wuhan University of Science and Technology, Wuhan 430065, China
[d] Department of Health Administration and Policy, College of Health and Human Services, George Mason University, Fairfax, Virginia 22030, USA

ARTICLE INFO

Keywords:
OCTA
Retinal vessel segmentation
Convolutional neural network
Transformer

ABSTRACT

Optical coherence tomography angiography (OCTA) enables detailed visualization of the vascular system. OCTA is of great significance for the diagnosis and treatment of many vision-related diseases. However, accurate retinal vessel segmentation is a great challenge due to obstacles such as low vessel edge visibility and high vessel complexity. We propose a novel OCTA retinal vessel segmentation method (ARP-Net) based on the Adaptive gated axial transformer (AGAT), **R**esidual and Point repair modules. To reduce the impact of high vascular complexity on segmentation, we proposed a network composed of transformer and convolution branches to fuse the global and local information. Furthermore, considering the high computation of transformer, we propose an AGAT in the transformer branch. Finally, the low visibility of regions such as vessel edge in OCTA images makes the prediction of the network in these regions difficult. Therefore, we also propose a point repair module to re-predict these regions. We have performed experiments on two public OCTA vessel segmentation datasets and achieved better results than the latest state-of-the-art methods.

1. Introduction

Retinal blood vessels, which are the tissues that deliver nutrients to various tissues in the eye, can be observed non-invasively by physicians with different imaging modalities. The retinal vessel system will not change in its lifetime except for pathological changes. Retinal vascular characteristics, such as vascular thickness, reflectivity and curvature, can be used as important biomarkers for many retinal and hematological related diseases [1]. By observing the pathological changes of retinal vessels, it help doctors to diagnose early glaucomatous optic neuropathy (GON), diabetic retinopathy (DR) and other diseases [2]. For example, DR results in the proliferation of retinal new blood vessels on the optic disc, retina, and iris. This eventually leads to retinal detachment [3,4]. The diagnosis of DR is based on the presence and growth of new blood vessels on the retina [5]. Atherosclerosis (AS) associated with wet age-related macular degeneration (wAMD) causes conditions such as narrowing of retinal blood vessels and the growth of blood vessels in the choroid. It eventually leads to blindness in the patients [6]. wAMD can be diagnosed by neovascularization of the retina [5]. Therefore, it is necessary to obtain their various characteristics. It helps doctors diagnose retinal and hematology-related diseases to reduce the bad conseqences like blindness. The segmentation of blood vessels can obtain a variety of retinal vascular features. However, as shown in Fig. 1 (a), the structure of retinal vessels is complex. Manual marking the ground truth as shown in Fig. 1 (b) requires professional knowledge, and it is time-consuming and laborious. Therefore, an effective automatic retinal vessel segmentation method is essential. It needs to quickly and accurately identify and segment blood vessels to provide clinical diagnosis basis for doctors. This greatly reduces the burden on doctors and improves the efficiency of understanding patients' disease conditions and making decisions.

Among retinal vessel imaging techniques, color fundus is commonly used. However, it has difficulty in capturing microvasculature. Indocyanine green angiography and fluorescence angiography [7] can capture the retinal vessel system including capillaries. But they are invasive, non-immediate, and the injected substances may cause severe consequences [8]. In contrast, optical coherence tomography angiography (OCTA) [9] is an immediate, non-invasive imaging technique. As shown in Fig. 1, the OCTA image in Fig. 1 (a) has more details than that in Fig. 1 (c) of fundus image. More and clearer vessels are seen in OCTA images.

* Corresponding author.
E-mail address: lxmspace@gmail.com (X. Liu).

https://doi.org/10.1016/j.bspc.2023.104604
Received 14 September 2022; Received in revised form 8 December 2022; Accepted 14 January 2023
Available online 28 January 2023

OCTA can generate high-resolution images of the retinal vessel and present vessel detail at the capillary level. It allows quantitative assessment of the microvascular and morphological vessel structures of the retina [10]. In conclusion, OCTA images are more suitable for the diagnosis of diseases. It is essential for automatic on retinal vessel segmentation on OCTA images.

In recent years, deep learning methods have emerged in medical image segmentation [11–13]. OCTA images are characterized by a low signal-to-noise ratio. This makes the areas of vessel walls, vessel tips, and small vessels blurred in the image (we call these areas uncertain regions), resulting in poor segmentation. Additionally, variety in size, shadow artifacts, and pathology increase the difficulty of segmentation. It leads to the segmented vessels to be broken or missing [10]. To address these issues, various deep learning methods have been proposed for OCTA retinal vessel segmentation. Passas et al [14] proposed i-UNet. The network reduces the impact of low signal-to-noise ratio on segmentation by iteration. Convolutional networks cannot establish good long-distance dependencies and only consider local features. While small blood vessels are widespread in the vascular system. This makes the segmented vessels often fracture and missing. Ma et al [10] proposed OCTA-Net, which alleviates the discontinuous problem in segmented vessels by introducing centerline-level vessel segmentation in pixel-level retinal vessel segmentation. However, the network is not optimized for uncertain regions. This makes its predictions on uncertain regions inaccurate.

In summary, we observed that existing OCTA retinal vessel segmentation methods have some problems: 1) The structure of retinal vessels is connected, but most of the OCTA retinal vessel segmentation methods do not consider the correlation between vessels. They mostly use convolutional networks. However, the convolutional network is weak in capturing non-local features and lack global contextual information, which eventually leads to the problem of broken or missing vessel [15]; 2) Due to the relatively high noise of OCTA image, the uncertain regions are blurred and difficult to segment. However, most of the OCTA retinal vascular segmentation methods did not specifically deal with these uncertain regions. This affects the prediction, lead to vessel narrowing or thickening, and missing vessel tips or small vessels [14].

To solve these problems, we propose a novel OCTA pixel-level retinal vessel segmentation network, called **ARP-Net**. This network is a dual-branch network based on **A**daptive gated axial transformer and **R**esidual module, and a **P**oint repair module is incorporated into the network. For the first problem, we propose a network including transformer branch [16] and convolution branch. Transformer [16] calculates the similarity between any position in the feature map. Therefore, we use transformer to capture global context information. Based on this, the blood vessels in the feature map are used as a basis to calculate the potential correlation and connectivity between vessels at each location in the feature map. Then, the segmented vessels on the same vascular tree are aggregated together. The non-local features with strong semantic information captured by transformer-based branch are used to establish remote dependence on the target. This makes the vessel belonging to the same vessel tree closely connected, to alleviate the problems of missing vessels. In addition, considering transformer is computationally intensive, we use our proposed adaptive gated axial transformer module in the transformer branch to reduce the computational complexity. This module reduces the computational complexity through axial attention [17]. The relative position bias and adaptive gated mechanism are added to axial attention to alleviate the impact of the small OCTA retinal vessel segmentation dataset and the lack of location information for self-attention computation on predictions. To overcome the second problem, we propose a point repair module. This method achieves more accurate prediction results by filtering out the most uncertain pixels in the feature map to re-predict. It alleviates the problem of wrong vessels segmentation in the previous methods [18–20]. The ARP-Net we proposed has achieved good results in OCTA-6M and OCTA-3M subsets of OCTA-500 public dataset (https://ieee-dataport.org/open-access/octa-500).

The main contributions of this work include:

1) We proposed a novel OCTA retinal vessel segmentation method, which consists of a transformer-based branch and a convolution-based branch. Information is exchanged between inner layers through the feature interaction unit to fully utilize local and global information. This alleviates the problem of broken vessels in predictions. To the best of our knowledge, this is the first work using transformer and convolution to segment vessels in OCTA images.

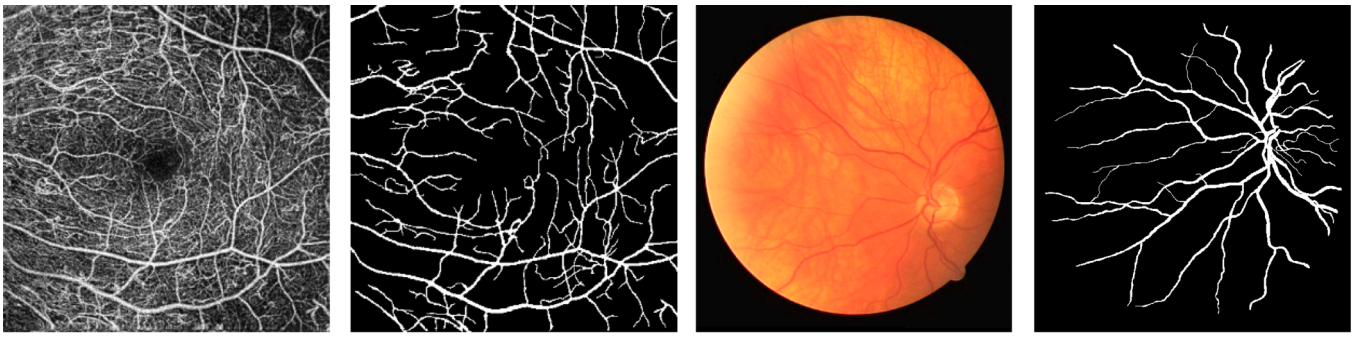
2) We proposed a point repair module. The module can correct the feature map to alleviate the problem of vessel missing in OCTA retinal vessel segmentation task. In addition, we also proposed adaptive gated axial transformer module. It forms a transformer branch to reduce the excessive computation of transformer.

3) Experiments on the OCTA-6M and OCTA-3M datasets show that our method outperforms state-of-the-art methods on the OCTA retinal vessel segmentation task.

2. Related work

2.1. Retinal vessel segmentation in OCTA image

Retinal vessel segmentation plays an important role in the diagnosis and treatment of many vision-related diseases. Due to the low signal-to-noise ratio of OCTA image, the edge of vessels and small vessels (we call



(a) OCTA retinal image (b) Vessel labeling of OCTA retinal image (c) Fundus image (d) Vessel labeling of fundus image

Fig. 1. Example of optical coherence tomography angiography (OCTA) images and fundus images sample, where (a) corresponds to (b) and (c) corresponds to (d).

these regions uncertain regions) in the image are blurred, so that it is difficult to segment [14]. To address this problem, some works [19,21] incorporated attention modules to convolutional networks. The attention mechanism can improve the recognizability of semantic features, thereby enhancing the segmentation of vessels in images. However, the mechanism cannot optimize for uncertain regions. This adds unnecessary overhead to the network. Another idea is to use a multi-stage approach. Some works [10,22] segment vessels in stages by introducing different types of images or labels, while other works [23] assist the segmentation of vessels by introducing different tasks. There are also works [8,14] to enhance the weight of vessels in the image through multi-stage segmentation. However, these methods increase the difficulty of completing the OCTA retinal vessel segmentation task. In this paper, we propose a point repair module to tackle the accurate segmentation of uncertain regions. By accurately finding the points with the highest uncertainty in the feature map, the module only re-predicts the values of these points, so as to improve the accuracy of those parts that are not easy to be segmented with a small amount of calculation [24].

2.2. Reduction of computational complexity in transformer

Transformer is popular in computer vision (CV). However, it must face the problem of high computation cost. To this end, extensive research has been carried out to alleviate the problem of high computational complexity of transformers. As a promising solution, the direction of structural modification of the transformer has attracted considerable attention. Because the transformer is originally a natural language processing (NLP) technology, its structure needs to be modified to adapt to solve problems in different CV fields. These methods try to bridge the difference between NLP and CV with different ideas, such as using the transformer only on patches that are chunked from the input image [16,25]. However, the transfer of these methods to other tasks requires extensive modifications to the model. Another way is to modify the self-attention module [26]. This way can be easily transferred to other tasks. Our method belongs to the late category, which reduces the computational complexity to a linear level by modifying the self-attention module.

3. Proposed method

In this section, we describe in detail the proposed ARP-Net. The total number of images for the training set is denoted with M. We leverage a dataset $D = \{(I_m, L_m)\}_{m=1}^{M}$ for training, where $I_m \in R^{H\times W}$ is the input image and $L_m \in \{0, 1\}^{H\times W}$ is label. H stands for the height and W for the width. L_m is a pixel-level label, which contains two types of labels: 0 represents the pixel of the background and 1 represents the pixel of the vessel. The overall framework of our method is shown in Fig. 2 (a). "Norm" in this method all means batch normalization operation. In the formulas in this section, we use X to represent the feature map that will be input to a certain module.

3.1. Network backbone – dual branch network

Considering that retinal vessels in OCTA images have wide coverage and variable thickness, it is necessary to utilize global information to guide local information extracted at different scales, which improve the accuracy of local information extracted by the convolutional module. As shown in Fig. 2 (a), the overall structure of the network refers to the idea of Conformer [15] for classification problem to the field of OCTA retinal vessel segmentation. The encoder component (as shown in Fig. 2 (b)) is composed of a residual module that extracts initial feature map, transformer branch based on adaptive gated axial transformer (AGAT) (shown in small blue boxes in Fig. 2(b)), convolution branch based on residual module [27] and feature interaction unit (FIU)(as shown in the orange box in Fig. 2 (b)), and the decoder is composed of up blocks (as shown in Fig. 2 (e)).

The convolution branch adopts the feature pyramid structure, and the resolution of its feature map decreases with the increase of network depth, while the number of channels increases. Transformer does not consider the spatial position relationship between each pixel when calculating the affinity, which leads to a large loss of local details. In convolution networks, the convolution kernel slides on overlapping

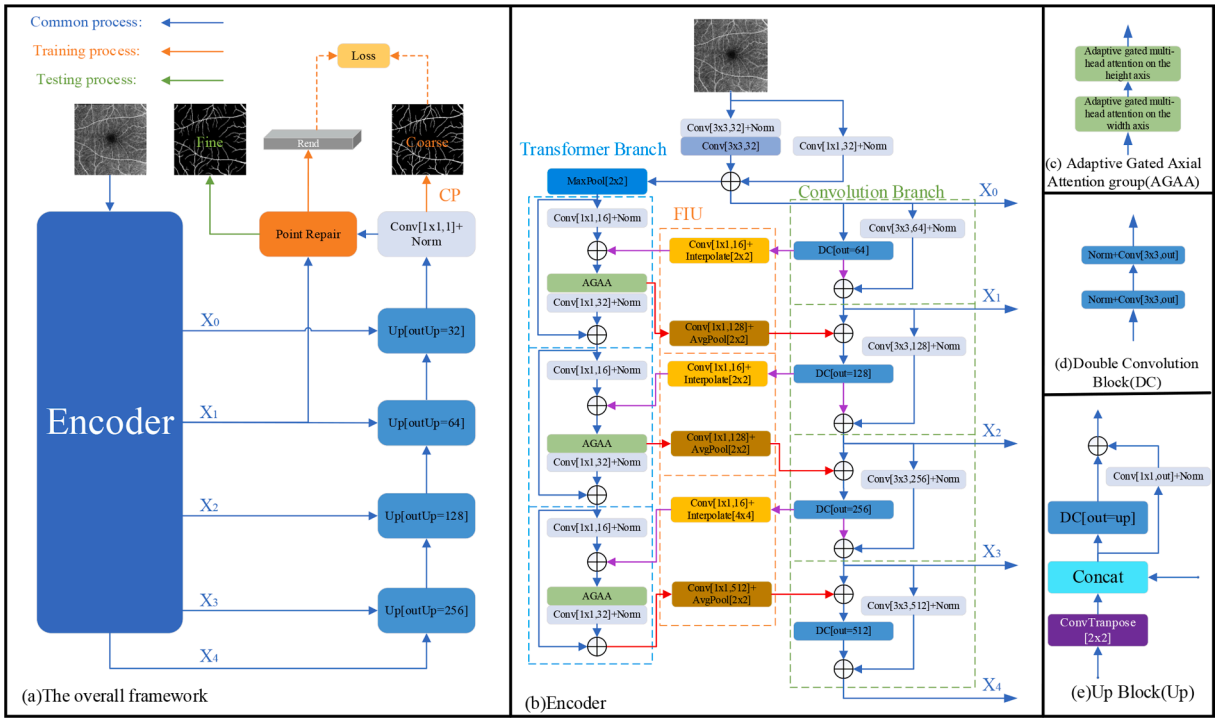


Fig. 2. Overall flow chart of our proposed method. (a) Overall framework, (b)Encoder, (c) Adaptive Gated Axial Attention Block (AGAA), (d) Double Convolution Block (DC), (e)Up Block (Up).

feature maps, which can preserve fine local features. Therefore, the convolution branch continuously provides the missing local information of transformer branch. Each convolution block (CB, as shown in the small green dashed box in Fig. 2(b)) in the convolution branch can be expressed as:

$$CB(X, F^c) = Norm(Cv_3(X)) + DC(X + F^c) \tag{1}$$

$$DC(X) = Cv_3(Norm(Cv_3(Norm(X)))) \tag{2}$$

Where F^c corresponds to the feature map indicated by the red arrows in Fig. 2(b). Cv_1 and Cv_3 represent 1×1 and 3×3 convolution operations, respectively. Note that in the first small green box, the input in Eq. (1) has only X and no F^c. In addition, the output of Eq. (2) will be input into the transformer branch through the FIU module.

The transformer branch contains three AGAT modules (as shown in the small blue box in Fig. 2(b)). In order to obtain more accurate global information, the resolution of the feature map of transformer branch of the network will not change with the increase of network depth. The AGAT block consists of 1×1 convolutions with decreasing number of channels, the Adaptive Gated Axial Attention (AGAA, as shown in Fig. 4) group, 1×1 convolutions with increasing number of channels, and residual connections. The AGAA group consists of adaptive gated multi-head attention block on the height axis and width axis (as shown in Fig. 2 (c)). The AGAT block reduces its computational load by first decreasing and then increasing the number of channels. It can be represented as:

$$AGAT(X, F^t) = X + Norm(Cv_1(AGAA(Norm(Cv_1(X)) + F^t))) \tag{3}$$

$$AGAA(X) = AGMA_H(AGMA_W(X)) \tag{4}$$

Where $AGMA_H$ and $AGMA_H$ represent adaptive gated multi-head attention on the height axis and width axis respectively. F^t corresponds to the feature map indicated by the purple arrows in Fig. 2(b). In addition, the output of Eq. (4) is input into the convolution branch through the FIU.

FIU is used to exchange information between the transformer branch and the convolution branch. The unit consists of from the transformer branch to the convolution branch (denoted as *trans*-conv branch, shown as deep-yellow box in Fig. 2(b)) and from the convolution branch to the transformer branch (denoted as conv-trans branch, shown as light-yellow box in Fig. 2(b)). In the *trans*-conv branch, we use convolution to change the number of channels and bilinear interpolation to change the resolution of the feature map. To obtain more complete global information, we input the feature maps obtained from the AGAA group into the FIU. The feature map obtained from the FIU are then added before the convolution operation in each block of the convolution branch to make the feature extracted by the convolution operations more accurate. The process can be written as:

$$FIU^c(X) = Interpolate(Cv_1(X)) \tag{5}$$

From the conv-trans branch, we use convolution to change the number of channels and average pooling to change the size of the feature map. We feed the feature map after double convolution in the convolution branch into FIU to obtain larger-scale, higher-level detailed features. After that, before adding the feature map obtained from FIU to the AGAA group, the feature map input into AGAA have richer and more detailed information. The process can be written as:

$$FIU^t(X) = AvgPool(Cv_1(X)) \tag{6}$$

In the process of training and testing, the image goes through the following process. When the image inputs the network, the feature map containing position prior information is first obtained through the residual module. The feature map is then input into transformer branch and convolution branch respectively, and feature interaction is performed between each layer using FIU. Convolution operation provide location prior information and local features, which preserve more details for transformer branch and inhibits the influence of noise. Transformer provide global and dynamic receptive fields, which provide global context information for convolution to help it to extract the required local information more accurately, and enable convolution branch to obtain more complete and accurate features [15]. They are compatible and complementary, but independent of each other, thus better integrating the advantages of convolution and transformer. After that, the feature map obtained by encoder is input into decoder, and coarse prediction CP of the same size as the label is obtained. Finally, coarse prediction CP was guided by feature map X_1 obtained from the first block of convolution branch in point repair module, and the results were obtained through re-prediction of uncertain areas such as vessel edge and vessel tip.

3.2. Point repair module

Due to low signal-to-noise ratio in OCTA images, it makes uncertain regions such as subtle and blurred retinal blood difficult to segment. The dual-branch network is not sensitive to these regions, and it is difficult to make accurate predictions. Therefore, we propose point repair module into the network to re-predict these hard points in order to achieve higher segmentation accuracy. The module is improved from point rend [24]. OCTA images are mostly single-channel grayscale images, so we modified the original point selection strategy. The module is divided into two stages of training and testing, and the specific steps are shown in Fig. 3.

Training phase: The module inputs feature map X_1 and coarse prediction CP into the module together. Considering that the coarse prediction is not so accurate during the training process, we add a specific random selection strategy to the point selection strategy, which makes the network easier to train and enhances its robustness. First, αP points ($\alpha > 1$) were randomly sampled on the coarse prediction CP (we set P to 256 empirically). Second, when selecting uncertain points among the αP points, we calculate the difference between these points and 0.5, and select the βP points ($0 < \beta < 1$) with the smallest difference. Third, $(1-\beta)P$ points are randomly sampled again on the coarse prediction CP, and combined with the βP points obtained earlier to build the final sampled P points. These points are the confusing points that the module needs to re-predict. Then bilinear interpolation is performed on the positions in feature map X_1 and the coarse prediction CP for these points, and their results are concatenated on the channel dimension. After that, the obtained results are input into a multi-layer perceptron (MLP) to obtain the final prediction of these uncertain points 'Rend'. Finally, we calculate the loss between the coarse prediction CP and label as well as the loss between predicted result 'Rend' and label.

Testing phase: The module inputs feature map X_1 and coarse prediction CP into the module together. First, calculate the difference between the coarse prediction CP and 0.5, and select the P points with the smallest difference. These points are the confusing points that the module needs to re-predict. Then bilinear interpolation is performed on the positions in feature map X_1 and the coarse prediction CP for these points, and their results are concatenated on the channel dimension. After that, the obtained results are input into a multi-layer perceptron (MLP) to obtain the final prediction results of these uncertain points 'Rend'. Finally, we assign 'Rend' to the point at the corresponding position in the coarse prediction CP to get the final prediction 'Fine'.

3.3. Adaptive gated axial transformer module

Given transformer's very high computational complexity for CV tasks, it is difficult to adopt it directly. As we know, transformer's high computational complexity is mainly due to its multi-head self-attention module. The self-attention is written as [16]:

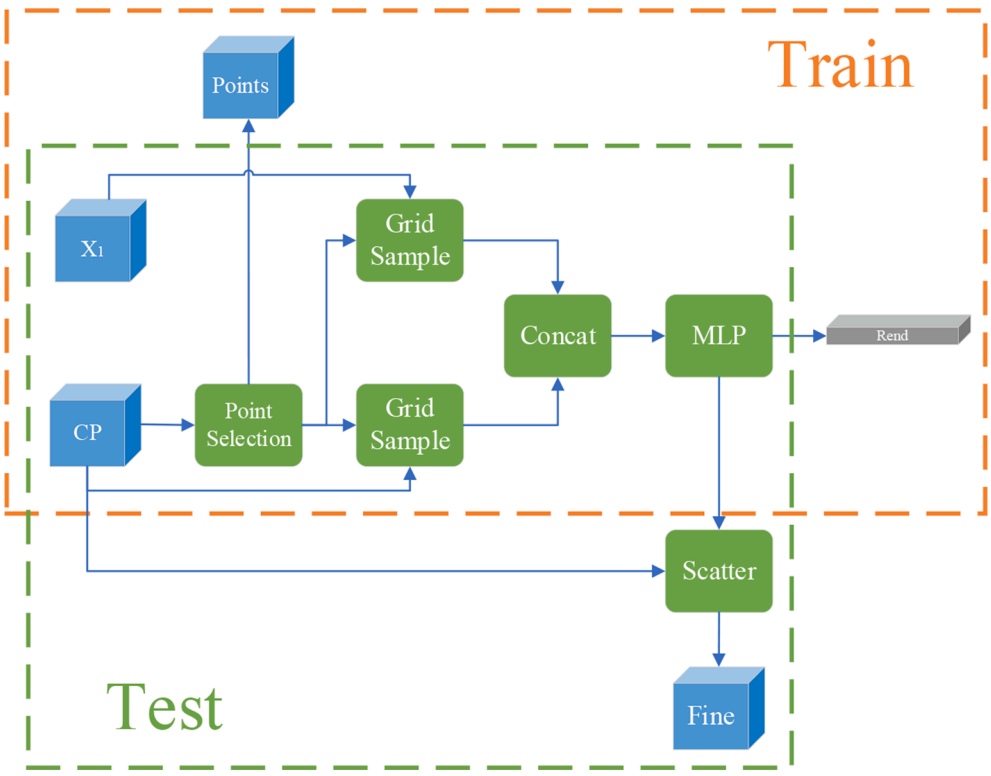


Fig. 3. Point repair module. The 'point selection' refers to the point selection strategy. The 'grid sample' refers to extracting feature vectors from the input feature maps according to position coordinate 'Point'. The 'concat' means concatenating different features. The 'MLP' refers to multilayer perceptron. The 'scatter' refers to replacing the value at 'points' in the feature map with the value of the feature 'rend'.

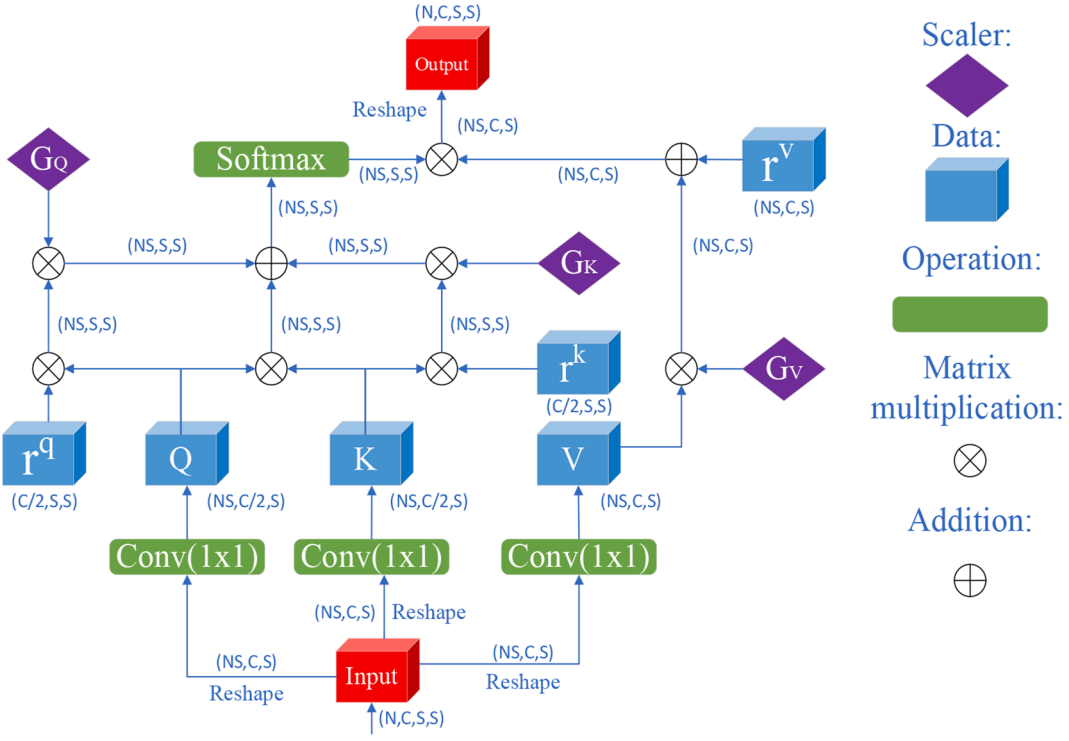


Fig. 4. Adaptive Gated Axial Attention. Where N is the batch number, C is the channel number, and S is the size of the feature map.

$$Y_{ij} = \sum_{h=1}^{H} \sum_{w=1}^{W} softmax(q_{ij}^{T} k_{hw}) v_{hw} \tag{7}$$

The projections $q = W_Q X$, $k = W_K X$ and $v = W_v X$ represent the query, key and value respectively, and all projections are calculated from the input feature map $X \in R^{H \times W}$. q_{ij}, k_{hw} and v_{hw} are query, key, and value at a position respectively $(i \in \{1, \cdots, H\}, \quad j \in \{1, \cdots, W\},$

$h \in \{1, \cdots, H\}$, and $w \in \{1, \cdots, W\}$). The projection matrices W_Q, W_K, $W_V \in R^{C_{in} \times C_{out}}$ are learnable. As shown Eq. (7), the global affinity is calculated based on $softmax(q^T k)$, and its result is multiplied with v. Hence, self-attention can capture non-local features from the entire feature map, which convolution operations cannot. However, this also brings the computational complexity of self-attention to a square level. Besides, location information is often used to capture the structure of objects. Self-attention does not utilize any location information to compute non-local information, which makes the model unable to capture the structure of vessels well.

To alleviate these problems, we draw on the idea of medical transformer [26] and propose an AGAT (as shown in the small blue box in Fig. 2(b)). The AGAT module consists of 1x1 convolutions with decreasing number of channels, the AGAA (as shown in Fig. 4) group, 1x1 convolutions with increasing number of channels, and residual connections. The AGAA group consists of an adaptive gated multi-head attention block on the height axis and an adaptive multi-head attention block on the width axis (as shown in Fig. 2(c)). AGAT reduces its computational complexity by splitting the self-attention into two axial self-attentions [17] that operate only on the height and width axes, respectively. The computational complexity is reduced to a linear level, which greatly improves the computational efficiency. In addition, position bias is introduced into axial self-attention to make it more sensitive to position information. On the other hand, considering the positional bias requires a large amount of data to train, and the dataset size of OCTA retinal vessel segmentation is small. This makes the learning of positional bias difficult, which affects the performance of the prediction. To this end, AGAT adds a learnable weight as a gate to the positional bias to reduce the influence of the learned inaccurate positional bias on the self-attention. The specific structure of AGAA is shown in Fig. 4. The equation is written as (AGAA on the width axis is similar to that on the height axis):

$$Y_{ij} = \sum_{h=1}^{H} softmax\left(q_{ij}^T k_{ih} + G_Q q_{ij}^T r_{ih}^q + G_K k_{ih}^T r_{ih}^k\right)\left(G_v v_{ih} + r_{ih}^v\right) \tag{8}$$

Where r^q, r^k, r^v respectively represent the position deviation corresponding to q, k, and v. G_Q, G_K, $G_V \in R$ are learnable weights that together create a gating mechanism that controls the degree to which the relative position encoding affects the non-local context. If the positional bias is accurately learned, the adaptive gated mechanism assigns a higher weight to it.

3.4. Loss function

ARP-Net training is a single-stage process. The model consists of pixel level dice loss function and binary cross entropy (BCE) loss. Its definition is expressed as:

$$Loss = Loss_{bce} + \lambda Loss_{dice} \tag{9}$$

Where $Loss_{dice}$ and $Loss_{bce}$ represent dice loss function and BCE loss function respectively. λ is the weight of the loss function. BCE loss considers the difference between the prediction and ground truth of all pixels in the image. However, the blood vessels are thin in OCTA image. In OCTA images, the pixels of blood vessels are less than one tenth, and more pixels belong to non-blood vessels. This indicates that there is a problem of category imbalance between blood vessels and background, which cannot be dealt with BCE loss. Therefore, we introduce dice loss into the total loss. Dice loss alleviates the problem of category imbalance. Finally, we use the combination of dice loss function and BCE loss function to optimize the model. Their definitions are expressed as:

$$Loss_{dice} = 1 - 2\sum_{i=1}^{n}(pd_i \times lb_i) / \sum_{i=1}^{n}(pd_i + lb_i) \tag{10}$$

$$Loss_{bce} = -\sum_{i=1}^{n}(lb_i log(pd_i) + (1 - lb_i) log(1 - pd_i)) \tag{11}$$

Where, pd represents the prediction of the network, and lb represents the corresponding ground truth. pd_i and lb_i represent pixels in network prediction map and ground truth respectively. n represents the number of pixels in the image.

4. Experimental results

4.1. Dataset

Our proposed method was tested on the OCTA-6M and OCTA-3M datasets. The two datasets both belong to the OCTA-500 dataset built by Li et al. [8], in which the OCTA-6M and OCTA-3M datasets represent the datasets composed of OCTA projected images with a field of view of $6mm \times 6mm$ and $3mm \times 3mm$, respectively. The OCTA-500 dataset used Iowa software (OCTExplorer 3.8) [28–30] and hierarchical segmentation method [31] to segment the inner limiting membrane (ILM), outer plexiform layer (OPL), and Bruch membrane (BM) of the retinal layers from 500 OCT volumes, and manually selected the better segmentation results to form the dataset. The pixel level label of the retinal vessels includes the vessels on the projection between ILM and OPL, the image was obtained by maximum projection. Five trained researchers annotated the retinal vessels, and three senior ophthalmologists reviewed the results to obtain the ground truth annotation [8]. OCTA-6M dataset contains the OCTA retinal image data of 300 subjects. The resolution of the OCTA projection image and the corresponding pixel-level label is $400px \times 400px$. OCTA-3M dataset contains the OCTA retinal image data of 200 subjects. The resolution of the projection image and the corresponding pixel-level label is $304px \times 304px$.

In the experiments on the two datasets, we randomly divided the dataset into five sub-datasets of the same size, and selected three of them as training set, one as test set, and one as validation set. The final experimental result is the average of the five experiments.

4.2. Comparing methods and metrics

In order to investigate the vascular segmentation performance of the proposed ARP-Net, we compared it with Weighted Res-UNet [32] (WR-UNet), CS-Net [21], i-UNet [14], IterNet [33], U-Net [34], Swin-UNet [35], Swin [25], trans-UNet [36] and Medical Transformer [26] (MedT) on two datasets. All methods are based on convolutional networks except the last four methods, they are based on the transformer. The first four are the segmentation techniques applied in the field of retinal vessel segmentation. U-Net has excellent performance in the medical image segmentation. WR-UNet adds the weighted attention mechanism based on Res-UNet, which allows the model to better learn the features of vessel. Unlike the U-Net based CNN, CS-Net includes a self-attention mechanism in the encoder and decoder [21]. Two types of attention modules are utilized-spatial attention and channel attention, to further integrate local features with their global dependencies adaptively. i-UNet is a method based on Res-UNet, which optimizes the results through iteration. IterNet assists vessel positioning by using FAZ segmentation task, and optimizes prediction results by introducing weight sharing and attention operations in vessel segmentation branches. Swin proposes a hierarchical transformer whose representation is computed with shifted windows. The shifted windowing scheme limits self-attention computation to non-overlapping local windows, and allows cross-window connections [25]. Swin-UNet makes it more suitable for medical image segmentation tasks by modifying the encoder of Swin-transformer and adding skip connections. trans-UNet combines the advantages of Transformer and U-Net. On the one hand, Transformer encodes tokenized image patches from CNN feature maps as input sequences for extracting global context. On the other hand, the decoder

upsamples the encoded features and combines them with high-resolution CNN feature maps for precise localization [36]. MedT extends the transformer architecture by introducing an additional control mechanism in the self-attention module. Furthermore, to efficiently train the model on medical images, MedT uses a local–global training strategy [26].

For the OCTA-6M dataset, we use the Adam optimizer to train the network with a batch size of 4 for 400 iterations. The learning rate is initially set to 0.0001 and then decreases with a rate of 0.9 as the iteration progresses. In our method, λ in the joint loss function and P, α, β in the point repair module are empirically set to 1.0, 256, 3, and 0.75, respectively. We rescaled the resolution of the two datasets to $256px \times 256px$ in the following experiment. Data augmentation used horizontal flip, vertical flip, and rotation. The augmented procedure was applied to all methods in the experiment. Our model is built using PyTorch. All the experiments in this paper were run on four NVIDIA GeForce 1080 Ti GPUs. On the OCTA-6M dataset, our method took 7 h to training. During the test, a single image takes 0.23 s.

In this paper, Dice coefficient (Dice), Balance Accuracy (BACC), Jaccard Index (JAC), Sensitivity (Sen), Specificity (Spe), G-mean score (G-mean) [37], False Discovery Rate (FDR) and False Negative Discovery Rate (FNDR) [21] were used as criteria. Where, DICE is used to measure the similarity between the prediction map and the target map, which is expressed as:

$$Dice = 2TP/(2TP + FN + FP) \tag{12}$$

Where TP, TN, FP and FN respectively represent the number of true positive, true negative, false positive and false negative in the prediction. BACC is used to measure the accuracy of prediction in unbalanced binary classification tasks, and its formula is:

$$BACC = (Spe + Sen)/2 \tag{13}$$

$$Sen = TP/(TP + FN) \tag{14}$$

$$Spe = TN/(TN + FP) \tag{15}$$

Where Sen represent the proportion of correctly predicted targets in the actual number of targets, Spe represents the proportion of correctly predicted backgrounds in the actual number of backgrounds. Jaccard index is used to compare the similarity and difference between the prediction map and the target map, and its formula is:

$$JAC = TP/(P + FN + FP) \tag{16}$$

G-mean score is often used for class imbalance problems. Regardless of positive and negative categories, it treats positive and negative categories equally [37]. It is used to compare the accuracy and balance between classes for each class. Its formula is:

$$G - mean = \sqrt{Sen \times Spe} \tag{17}$$

FDR means the ratio of the total number of pixels falsely detected as blood vessels to the total number of blood vessel pixels in the ground truth. Its formula is:

$$FDR = FP/(TP + FP) \tag{18}$$

FNDR means the proportion of undetected blood vessel pixels in the background. Its formula is:

$$FNDR = FN/(TN + FN) \tag{19}$$

4.3. Results on the OCTA-6M dataset

In this section, we present the results of qualitative and quantitative comparisons with several methods on OCTA-6M dataset. As can be seen from Fig. 5, compared with other methods, there are no broken blood vessels in the segmentation results of our method (as shown in column 12). This is due to the rich long-range information provided by the transformer branch of our method to the convolution branch, which enhance the robustness of our network. Row 3 in Fig. 5 show that our method is more accurate in segmenting small vessels than other methods. This is the point repair module that re-renders the small blood vessels with high uncertainty in the segmentation results, thereby improving the accuracy of the prediction of our network. As U-Net does not consider the global field of vision, it has many broken blood vessels (as shown in row 1). WR-UNet uses weighted and residual module segmentation to achieve higher accuracy of segmentation results, but it still suffers from vessel segmentation errors and vessel rupture (as shown in rows 2 and 3). i-UNet increases the weight of the blood vessels in the input image by an iterative method to reduce the impact of noise on the segmentation of the image. However, the convolutional network cannot establish a good long-distance dependence, and there are still obvious problems of fracture of segmented vessels (as shown in the rows 1 and 3). CS-Net adds a spatial and channel attention module at the bottom of the network to improve the recognition of vascular features and reduce the impact of noise and low contrast in OCTA images. It only uses the module at the bottom, so it lacks a lot of detail information, which reduces the accuracy of aggregation (as shown in column 3). MedT works by running the transformer on patches and the whole image separately, but only has the intersection at the output layer. This results in a lack of information interaction within the network, leading to the presence of broken blood vessels in the predictions (As shown in the row 3 and column 7). Swin is a hierarchical transformer whose representation is computed with shifted windows. The shifted window reduces

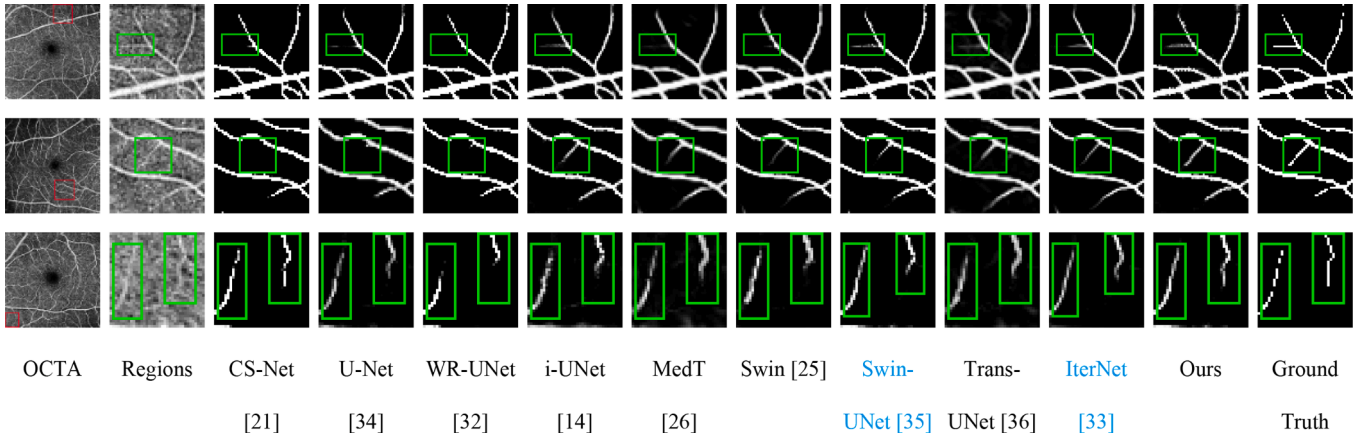


Fig. 5. Example of retinal vessel segmentation results for five methods on OCTA-6M. The image in column 2 is an enlarged view of the red box in the image in column 1. The images in other columns correspond to those in column 2.

computational complexity by restricting the self-attention computation to non-overlapping local windows. This leads to the incomplete global field of view of swin. As shown in the row 3, although the prediction of swin inhibited the segmentation of false branches, there were many broken blood vessels. Swin-UNet, as a network that has made improvements to Swin's decoder, obtained more accurate segmentation results than Swin (as shown in column 9). trans-UNet replaces the bottom layer of UNet with a transformer to obtain global information. However, trans-UNet only uses the transformer at the bottom layer, which makes the network lack a lot of detailed information, which affects the accuracy of the segmentation results (as shown in column 10). Because of the introduction of FAZ segmentation auxiliary task, weight sharing and attention technologies, IterNet obtained more accurate segmentation results than other convolutional networks, but it still does not achieve very good results at the tip and edge of blood vessels (as shown column 11). The proposed ARP-Net alleviated these problems. Non-local context information and local information are integrated in each layer, so that the utilization of information is greatly improved. Moreover, the introduction of point repair module makes the points with high uncertainty are predicted more accurately. The combination of these modules improved the effectiveness of the method.

We used eight indicators for quantitative comparison. Table 1 listed the quantitative results obtained from experiments on the OCTA-6M dataset. As can be seen from Table 1, our method has achieved the optimal results in all indicators. Compared with WR UNet, i-UNet and CS-Net, which are also based on residual modules, IterNet has achieved better results in all indicators, which is due to the introduction of auxiliary tasks, weight sharing and attention modules in the network. The results in the table show that the mechanism of this network is more effective than other networks based on residual modules. Compared with CS-Net, IterNet increases BACC by 1.06 % (from 92.98 % to 94.04 %), Dice by 1.73 % (from 87.52 % to 89.25 %), and G-mean by 1.11 % (from 92.80 % to 93.91 %). In order to be more suitable for medical image segmentation, Swin-UNet has improved the decoder part of swin. This has improved the network performance. Compared with the convolution-based network, the transformer-based network achieves better results, which proves the effectiveness of the transformer. Among them, trans-UNet obtained performance similar to IterNet which is specially designed for retinal blood vessel segmentation. Our method achieves better results than trans-UNet. Compared with trans-UNet, our method improved BACC by 0.81 % (from 94.06 % to 94.87 %), Dice by 0.97 % (from 88.94 % to 89.91 %), and G-mean by 0.83 % (from 93.95 % to 94.78 %). Paired *t*-test analysis of dice, JAC and G-mean shows that our method improves significantly ($p_{dice} < 0.03$, $p_{JAC} < 0.04$, $p_{G-mean} < 0.01$) compared to trans-UNet. This also shows the effectiveness of the dual-branch structure based on AGAT and convolution and the introduction of point repair module. In the dual-branch structure, transformer branch provides good global information, while convolution branch provides excellent local characteristics. They interact global and local information through FIU, to obtain better feature maps. The point repair module optimizes the results by screening and re-rendering the areas with high uncertainty such as the edge and tip of blood vessels. Additionally, our network has a parameter amount that are closer to convolutional networks than several other transformer-based networks. The number of parameters of our network is only 1/4 of that of trans-UNet (from 405.11 M to 100.66M). The introduction of these two methods ensures the accuracy of our model while requiring only a small number of parameters.

4.4. Ablation experiment

In this section, we investigated the impact of each module on the final network result to better understand their role in the overall model. All experiments were performed on the OCTA-6M dataset.

4.4.1. Ablation experiment of dual branch network

To investigate the influence of each branch on the segmentation results of the dual-branch network. We performed ablation experiments on the OCTA-6M dataset, and the experimental results are shown in Table 2. Among them, we control the parameters of the two single branch networks by setting the number of channels in each layer of the network, so that their parameters are about 50 MB. As the parameter number of the residual module-based single-branch network is 54.15 M. From the first two rows of Table 2, most indicators of the AGAT-based single-branch network with global features are higher than the residual module-based single-branch network. The residual module-based network lacks global information, which leads to inferior results. From the rows 2 and 3 of Table 2, compared with the residual module-based network, the dual-branch network improved Dice by 1.64 % (from 87.49 % to 89.13 %), BACC by 1.37 % (from 92.89 % to 94.26 %), and G-mean by 1.44 % (from 92.71 % to 94.15 %). The paired *t*-test results of dice, BACC and FNDR show that the improvement of the dual-branch network compared with the residual module-based single-branch network is significant ($p_{dice} < 0.01$, $p_{BACC} < 0.01$, $p_{FNDR} < 0.01$). The dual-branch network incorporates local information into the transformer branch, resulting in improved indicators.

As can be seen from Fig. 6, the dual-branch network works best. The residual module-based network has more broken blood vessels than AGAT-based network due to its lack of global information (as shown in the row 1). In the retinal vessel segmentation task of OCTA images, transformer is effective. Our dual-branch network utilizes FIU to exchange the information of the transformer branch and the convolution branch to overcoming the shortcomings of convolution and transformer. Dual branch network predicts fewer broken blood vessels than the single branch network (as shown in row 1), and it can segment smaller vessels (as shown in row 2). These results demonstrate the effectiveness of the dual-branch network in OCTA retinal vessel segmentation.

4.4.2. Analysis of feature interaction unit

To further investigate the effect of FIU on the dual-branch network, we also conduct ablation experiments on FIU. Since the feature map is small when the network depth exceeds 5 layers, the convolution branch

Table 1
On OCTA-6M, quantitative comparison between the proposed approach and the comparative method by BACC, Dice, Sen, JAC, Spe, G-mean, FDR and FNDR, the best results are bold.

Method	BACC (%)	Dice (%)	Sen (%)	JAC (%)	Spe (%)	G-mean (%)	FDR (%)	FNDR (%)	Params (M)
WR-UNet [32]	92.56 ± 1.52	87.23 ± 2.48	86.25 ± 2.93	77.35 ± 3.97	98.76 ± 0.22	92.29 ± 1.66	11.77 ± 1.92	1.48 ± 0.34	64.12
U-Net [34]	92.01 ± 1.68	86.60 ± 2.64	85.31 ± 3.09	76.37 ± 4.08	98.72 ± 0.23	91.77 ± 1.76	12.07 ± 2.14	1.60 ± 0.35	53.58
CS-Net [21]	92.98 ± 1.69	87.52 ± 2.71	87.24 ± 3.14	77.70 ± 4.42	98.71 ± 0.24	92.80 ± 1.37	12.34 ± 2.42	1.34 ± 0.32	33.77
i-UNet [14]	92.65 ± 1.81	87.19 ± 2.82	86.60 ± 3.34	77.29 ± 4.56	98.71 ± 0.23	92.46 ± 1.87	12.21 ± 2.28	1.43 ± 0.35	85.72
MedT [26]	93.22 ± 1.79	88.07 ± 2.79	87.66 ± 3.34	78.68 ± 4.51	98.78 ± 0.25	93.06 ± 1.99	11.52 ± 2.24	1.31 ± 0.39	32.16
Swin [25]	93.32 ± 2.13	88.24 ± 2.91	87.84 ± 4.07	78.95 ± 4.69	98.80 ± 0.19	93.16 ± 2.22	11.36 ± 1.76	1.29 ± 0.45	133.24
Swin-UNet [35]	93.85 ± 1.79	88.71 ± 2.73	88.92 ± 3.35	79.71 ± 4.52	98.78 ± 2.32	93.72 ± 1.87	11.49 ± 1.84	1.16 ± 0.36	156.82
Trans-UNet [36]	94.06 ± 1.56	88.94 ± 2.52	89.34 ± 3.16	80.08 ± 4.28	98.79 ± 0.23	93.95 ± 1.71	11.45 ± 2.17	1.11 ± 0.30	405.11
IterNet [33]	94.04 ± 1.65	89.06 ± 2.44	89.25 ± 3.14	80.28 ± 4.05	98.83 ± 0.19	93.91 ± 1.75	11.13 ± 1.76	1.13 ± 0.34	62.41
Ours	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33	100.66

Table 2
Results of ablation experiments on OCTA-6M were obtained by our method. The value in bold is the optimal result under this indicator. The unit is percentage (%).

Method	BACC	Dice	Sen	JAC	Spe	G-mean	FDR	FNDR
Only Res-UNet	92.89 ± 1.68	87.49 ± 2.58	87.06 ± 3.19	77.76 ± 4.24	98.73 ± 0.22	92.71 ± 1.86	12.08 ± 2.17	1.38 ± 0.33
Only AGAT	93.38 ± 1.84	87.86 ± 2.93	88.06 ± 3.42	78.35 ± 4.68	98.70 ± 0.26	93.23 ± 1.91	12.34 ± 2.46	1.25 ± 0.36
Dual branch	94.26 ± 1.74	89.13 ± 2.68	89.73 ± 3.67	80.39 ± 4.13	98.79 ± 0.18	94.15 ± 1.98	11.46 ± 1.82	1.07 ± 0.40
Ours	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33

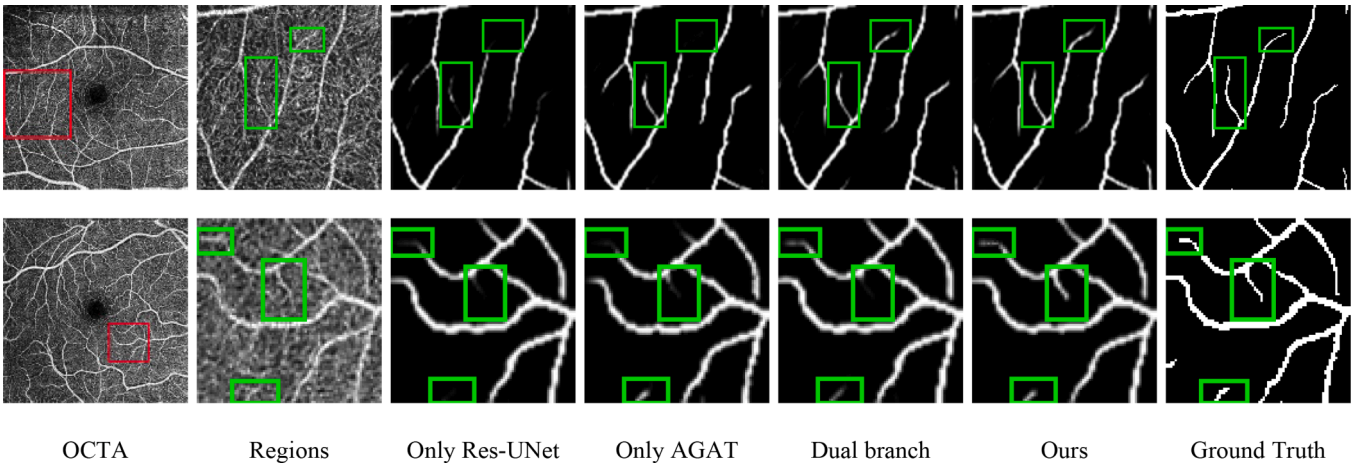


Fig. 6. Examples of retinal vessel segmentation results from ablation experiment with dual branch network on OCTA-6M. The image in column 2 is an enlarged view of the red box in the image in column 1. The images in other columns correspond to those in column 2.

cannot provide good details for the transformer branch, so our network depth only reaches 5 layers. Therefore, for the ablation experiment of FIU, we add a group of FIUs from bottom to top, until 3 groups (the number of FIU groups is limited to 3 when the convolution depth is 5). The experimental results are shown in Table 3. With the increase of the number of FIU groups, its indicators continue to rise. Compared with the network with only one FIU, the network with three FIUs achieves better results on all metrics. Its BACC improved by 0.74 % (from 94.13 % to 94.87 %), Dice by 0.99 % (from 88.92 % to 89.91 %) and G-mean by 0.76 % (from 94.02 % to 94.78 %). The paired *t*-test shows that the Dice of three FIUs has achieved a significant improvement compared with one FIU ($p < 0.03$). This shows that the introduction of FIU enhances the information interaction between the two branches and makes full use of the global and local information, thus achieving better results.

4.4.3. Analysis of point repair module

Considering that OCTA images have high noise, this will greatly affect the accuracy of the segmentation results. Especially in uncertain regions such as image edges, blood vessel walls and small blood vessels, noise greatly affect the segmentation results in these places. Therefore, we introduce the point repair module to re-predict the points in these regions.

To verify the effect of the point repair module on our network, we conduct ablation experiments on it. The results in Table 2 also show that the introduction of this module improves all indicators. Compared with the dual-branch network without the point repair module, ARP-Net improved Dice by 0.78 % (from 89.13 % to 89.91 %) and BACC by 0.61 % (from 94.26 % to 94.87 %). From the paired *t*-test results, compared with the dual branch network, ARP-Net is significant on BACC ($p < 0.05$), but not statistically significant on dice. From the green boxes in row 2 of Fig. 6, the introduction of point repair module improves the segmentation accuracy of small vessels in blurred images. The result indicates that point repair module improves OCTA retinal vessel segmentation.

In order to further investigate how many points should be set for re-predicting in point repair module in the OCTA retinal segmentation task. We conducted experiments on the OCTA-6M dataset, and the results are shown in Table 4. The number of points doubles from top to bottom. We investigated that for vessel segmentation on an OCTA image of size $256px \times 256px$, the number of misclassified points is roughly between 1000 and 1500 points. Therefore, we only carry out the experiment until the number of points is 2048. At 256 points, all of its indicators are better than others. Compared with 512 and 128 points, 256 points showed a small improvement in all indicators. Starting from 256 points, as the number of points increases, the corresponding indicators continue to decrease. When the number of points is 2048, since the number of re-predicted points is larger than the actual number of misclassified points, point repair module will misclassify the points that were originally classified correctly, making its indicators lower than the network without the point repair module. This resulted in some small blood vessels being classified as background in its prediction. The results lead us to choose the number of points to be 256.

4.4.4. Analysis of adaptive gated axial transformer

In this section, we investigated the effect of AGAT. We consider the 'No position bias' network in Table 5 as the baseline, which replaces

Table 3
Results of ablation experiments on OCTA-6M were obtained by FIU. The value in bold is the optimal result under this indicator. The unit is percentage (%).

Number	BACC	Dice	Sen	JAC	Spe	G-mean	FDR	FNDR
1	94.13 ± 1.69	88.92 ± 2.44	89.49 ± 3.21	80.05 ± 4.17	98.77 ± 0.19	94.02 ± 1.76	11.64 ± 1.71	1.09 ± 0.36
2	94.52 ± 1.71	89.46 ± 2.31	90.24 ± 3.15	80.93 ± 3.87	98.81 ± 0.17	94.43 ± 1.77	11.30 ± 1.42	1.01 ± 0.34
3 (Ours)	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33

Table 4
The number of points re-predicted in point repair module was experimented on OCTA-6M. The value in bold is the optimal result under this indicator. The unit is percentage (%).

Number	BACC	Dice	Sen	JAC	Spe	G-mean	FDR	FNDR
128	94.58 ± 1.84	89.45 ± 2.58	90.37 ± 3.51	80.91 ± 4.33	98.79 ± 0.19	94.49 ± 1.91	11.46 ± 1.72	1.00 ± 0.38
256 (Ours)	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33
512	94.62 ± 1.97	89.53 ± 2.50	90.43 ± 3.78	81.04 ± 4.21	98.80 ± 0.16	94.52 ± 2.04	11.36 ± 1.32	0.99 ± 0.41
1024	94.44 ± 1.65	89.06 ± 2.68	90.15 ± 2.97	80.28 ± 4.46	98.74 ± 0.25	94.35 ± 1.66	12.00 ± 2.40	1.01 ± 0.31
2048	94.04 ± 1.79	88.81 ± 2.71	89.30 ± 3.38	79.87 ± 4.51	98.77 ± 0.22	93.92 ± 1.86	11.68 ± 2.08	1.12 ± 0.37

Table 5
The results are obtained by performing ablation experiments on the adaptive gated axial transformer on OCTA-6M. The value in bold is the optimal result under this indicator. The unit is percentage (%).

Method	BACC	Dice	Sen	JAC	Spe	G-mean	FDR	FNDR
No position bias	94.07 ± 2.13	88.94 ± 2.94	89.37 ± 4.14	80.08 ± 4.89	98.79 ± 0.19	93.96 ± 2.25	11.49 ± 1.78	1.11 ± 0.45
No adaptive gate	94.38 ± 1.97	89.19 ± 2.76	89.99 ± 3.71	80.49 ± 4.61	98.78 ± 0.18	94.28 ± 2.03	11.59 ± 1.85	1.04 ± 0.41
Ours	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33

AGAA in ARP-Net with axial attention. 'No adaptive gate' means that only position bias is added to the baseline. ARP-Net refers to a network with position bias and adaptive gate added to the baseline. As observed from Table 5, although the convolution branch provides position information for the transformer branch, the introduction of additional position bias in the network also improves the performance of the network. This is because the transformer did not consider the position information when calculating the affinity, so it is necessary to add additional position bias in the process of calculating the affinity. To utilize the additional position bias further and more effectively, we add an adaptive gate on it, so as to increase the weight of useful information and reduce the weight of useless information in the position bias. By introducing these two components into the network, Dice, BACC and FDR of the network are increased by 0.97 % (from 88.94 % to 89.91 %), 0.80 % (from 94.07 % to 94.87 %), and 0.43 % (from 11.49 % to 11.06 %). From the paired *t*-test analysis, the introduction of the two components significantly improves the dice of our network ($p < 0.05$). The introduction of these components improves the indicators.

4.4.5. Analysis of weighted parameters in loss function

Our network is optimized by the joint loss function defined in Eq. (9). The joint loss function consists of two parts and affects the learning of the network by setting different weights λ. In order to study how much the weight λ is set, the joint loss function can better train the network. We conduct experiments on the joint loss function, which sets different $\lambda \in \{0.2, 0.6, 1.0, 2.0\}$ to train the network. The experimental results are shown in Table 6. When the weight λ is 1.0, the effect is better, so we set it to 1.0 in other experiments. But there is no significant improvement compared with other cases. We speculate that our joint loss function is insensitive to the weight λ.

4.4.6. Results on OCTA-3M dataset

The OCTA-3M is a dataset of $3mm \times 3mm$ OCTA images centered on the fovea region of the retina. The experimental results are shown in Fig. 7. The first four convolution-based network has a shortcoming of blood vessel rupture (as shown in row 3). The transformer-based network suppresses the vessel breakage with the global information obtained (as shown in row 3). Swin-UNet obtained more accurate segmentation results than swin (as shown in columns 8, 9). In addition, it can be seen from rows 1 and 2 that the transformer-based network can better segment the small vessels in the blurred image. This proves that the transformer can effectively alleviate the problem of blood vessel rupture and loss. Due to the introduction of the point repair module, our method has these advantages while being able to better differentiate between vessels and background (as shown in row 1). And the results in Table 7 also show that our method achieves better results. The IterNet obtained better prediction than other convolutional networks, but the segmentation results on small vessels are thinner (as shown in column 11). Compared with IterNet, our method improved on Dice, BACC and G-mean metrics with 1.32 % (from 91.74 % to 93.06 %), 0.82 % (from 95.34 % to 96.16 %) and 0.86 % (from 95.25 % to 96.11 %), respectively. From the paired *t*-test results, our network has a significant improvement over IterNet on dice, Sen and BACC ($p_{dice} < 0.01, p_{Sen} < 0.01, p_{BACC} < 0.01$). This demonstrates the superiority of our method.

5. Discussion

With the rise of deep learning, the research on retinal vascular segmentation has made great progress. In recent years, a large number of retinal blood vessel segmentation methods based on deep learning have emerged, but these methods still have difficulties to deals with the high complexity and low contrast problems [10,14] in blood vessels segmentation. In the face of these challenges, we proposed a novel method to improve the performance.

Because of the high complexity of retinal blood vessels, it will cause the network prediction to break or lose in the thin parts of blood vessels [10]. Some methods solve this problem by increasing the field of view of the network. Some of these works [38,39] increased the size of the convolution kernel, but the reception field was still limited. Other works [21,40] incorporated an attention module at the bottom of the network, but they cannot obtain accurate global context information due to the loss of many details. In this paper, we designed a transformer branch to extract global information to grasp the potential correlation and connectivity between blood vessels. In addition, we incorporated a

Table 6
The results are obtained by performing experiments on the joint loss function on OCTA-6M. The value in bold is the optimal result under this indicator. The unit is percentage (%).

λ	BACC	Dice	Sen	JAC	Spe	G-mean	FDR	FNDR
0.2	94.48 ± 1.96	89.57 ± 2.31	90.11 ± 3.76	81.11 ± 4.54	98.85 ± 0.18	94.38 ± 2.04	10.97 ± 2.28	1.03 ± 0.41
0.6	94.45 ± 1.71	89.61 ± 2.09	90.04 ± 3.31	81.18 ± 4.35	98.86 ± 0.21	94.35 ± 1.77	10.82 ± 1.99	1.04 ± 0.35
1.0 (Ours)	94.87 ± 1.62	89.91 ± 2.17	90.90 ± 3.11	81.67 ± 3.75	98.84 ± 0.14	94.78 ± 1.68	11.06 ± 1.58	0.94 ± 0.33
2.0	94.85 ± 1.68	89.86 ± 2.23	90.87 ± 3.61	81.59 ± 4.42	98.83 ± 0.23	94.77 ± 1.73	11.13 ± 2.18	0.94 ± 0.40

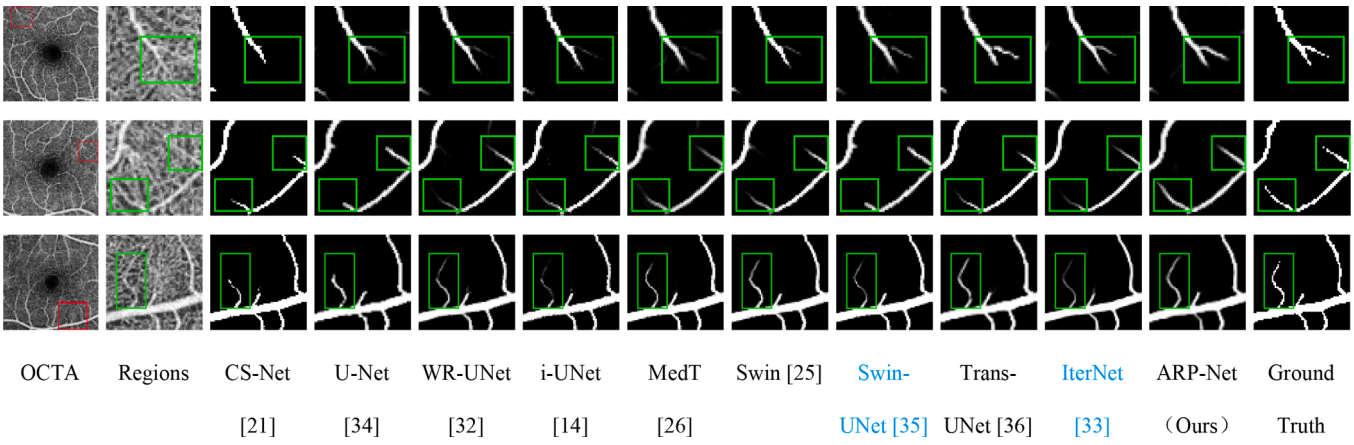


Fig. 7. Examples of retinal vessel segmentation results of the five methods on the OCTA-3M. The image in column 2 is an enlarged view of the red box in the image in column 1. The images in other columns correspond to those in column 2.

Table 7
The proposed method is quantitatively compared with other methods on OCTA-3M. The quantitative indexes are BACC, Dice, Sen, JAC, Spe, G-mean, FDR and FNDR. The best result is bold.

Method	BACC (%)	Dice (%)	Sen (%)	JAC (%)	Spe (%)	G-mean (%)	FDR (%)	FNDR (%)	Params (M)
WR-UNet [32]	93.44 ± 1.59	89.44 ± 1.94	87.50 ± 3.13	80.90 ± 3.21	99.38 ± 0.06	93.25 ± 1.68	8.53 ± 0.67	0.95 ± 0.27	64.12
U-Net [34]	93.17 ± 1.32	88.97 ± 2.06	87.01 ± 2.51	80.13 ± 3.44	99.34 ± 0.12	92.97 ± 1.39	8.98 ± 1.58	1.00 ± 0.21	53.58
CS-Net [21]	94.68 ± 1.48	91.03 ± 2.22	89.94 ± 2.95	83.54 ± 3.82	99.43 ± 0.09	94.57 ± 1.58	7.85 ± 1.46	0.75 ± 0.21	33.77
i-UNet [14]	94.40 ± 1.41	90.62 ± 2.31	89.39 ± 2.67	82.85 ± 3.91	99.41 ± 0.14	94.26 ± 1.49	8.12 ± 1.94	0.80 ± 0.21	85.72
MedT [26]	94.95 ± 1.62	91.25 ± 2.50	90.48 ± 3.11	83.91 ± 4.13	99.42 ± 0.14	94.85 ± 1.68	7.97 ± 1.88	0.71 ± 0.24	32.16
Swin [25]	95.21 ± 1.55	91.66 ± 2.43	90.98 ± 2.96	84.60 ± 4.20	99.44 ± 0.14	95.12 ± 1.60	7.65 ± 1.89	0.67 ± 0.23	133.24
Swin-UNet [35]	95.25 ± 1.49	92.03 ± 2.32	91.01 ± 2.84	85.24 ± 4.06	99.49 ± 0.13	95.16 ± 1.58	6.93 ± 1.78	0.67 ± 0.22	156.82
Trans-UNet [36]	95.27 ± 1.40	92.14 ± 2.13	91.04 ± 2.74	85.43 ± 3.79	99.50 ± 0.15	95.18 ± 1.74	6.73 ± 1.97	0.67 ± 0.17	405.11
IterNet [33]	95.34 ± 1.74	91.74 ± 2.21	91.24 ± 3.07	84.74 ± 3.85	99.43 ± 0.08	95.25 ± 1.82	7.76 ± 1.03	0.65 ± 0.26	62.41
Ours	96.16 ± 1.65	93.06 ± 2.07	92.82 ± 3.18	87.02 ± 4.24	99.51 ± 0.11	96.11 ± 1.69	6.70 ± 1.55	0.53 ± 0.24	100.66

convolution branch to the network to preserve high-quality local information for the transformer branch. The combination of local information and global information enables our network to better solve the problems of vascular rupture or loss. The experimental results in section 4 demonstrated that our proposed method outperformed the previous method in the OCTA retinal blood vessel segmentation task.

In addition, due to the low contrast of the OCTA image, it is difficult to accurately predict the uncertain areas such as the vessel tip and vessel edge in the image [14]. Some works [8,14] enhanced the weight of blood vessels in the image through multi-stage segmentation. Other works [10,22] segmented blood vessels in stages by introducing different types of images or labels. However, the requirements on other modality images limit the adoption of the methods and multi-stages make the methods complex. In this paper, we use the point repair module to deal with uncertain regions. This module improves the accuracy of those regions that are difficult to segment by finding the points with the highest uncertainty in the prediction and re-predicting them.

Although the proposed method effectively segments the complex retinal vascular structures, there are still some small and irregular vascular structures that have not been accurately classified. We can further improve the way transformer interacts with convolutions to make it more consistent with the retinal blood vessel segmentation task, so as to further improve our network. In addition, we can also introduce more powerful data enhancement technologies to expand the dataset, so that the network can learn more useful information.

6. Conclusion

Recently, deep learning-based segmentation techniques have been successfully applied to medical image segmentation [41–43]. In this paper, a novel OCTA retinal vascular segmentation network is proposed, which consists of a dual branch encoder based on adaptive gated axial transformer and residual module, and a decoder and a point repair module based on residual network. Encoder branch of the network exchange large amounts of global information and local information through the feature interaction unit module, to enhance the correlation between blood vessels and preserve a lot of detail information. The point repair module re-predicts the uncertain points to obtain more accurate segmentation. In the experiment of two OCTA datasets, the method achieves better results than the compared methods. In the future, we will further explore the characteristics of blood vessels in OCTA images and apply them to retinal blood vessel segmentation. Additionally, we will explore the feasibility of using a semi-supervised approach for the 3D OCTA retinal vessel segmentation task.

CRediT authorship contribution statement

Xiaoming Liu: Conceptualization, Methodology, Supervision. **Di Zhang:** Software, Methodology, Investigation, Writing – original draft. **Junping Yao:** Data curation. **Jinshan Tang:** Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data we used is publicly available.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176190.

References

[1] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, J. Qin, SCS-Net: A scale and context sensitive network for retinal vessel segmentation, Med. Image Anal. 70 (2021), 102025.
[2] Y. Tan, K.-F. Yang, S.-X. Zhao, Y.-J. Li, Retinal Vessel Segmentation with Skeletal Prior and Contrastive Loss, IEEE Trans. Med. Imaging (2022).
[3] L. Lahme, et al., Evaluation of ocular perfusion in Alzheimer's disease using optical coherence tomography angiography, J. Alzheimers Dis. 66 (4) (2018) 1745–1752.
[4] R.L. Engerman, Pathogenesis of diabetic retinopathy, Diabetes 38 (10) (1989) 1203–1206.
[5] J. Wei, et al., Genetic U-Net: Automatically Designed Deep Networks for Retinal Vessel Segmentation Using a Genetic Algorithm, IEEE Trans. Med. Imaging (2021).
[6] L.D. Hubbard, et al., Methods for evaluation of retinal microvascular abnormalities associated with hypertension/sclerosis in the Atherosclerosis Risk in Communities Study, Ophthalmology 106 (12) (1999) 2269–2280.
[7] R. Benson, H. Kues, Fluorescence properties of indocyanine green as related to angiography, Phys. Med. Biol. 23 (1) (1978) 159.
[8] M. Li, et al., Image projection network: 3D to 2D image segmentation in OCTA images, IEEE Trans. Med. Imaging 39 (11) (2020) 3343–3354.
[9] T.E. De Carlo, A. Romano, N.K. Waheed, J.S. Duker, A review of optical coherence tomography angiography (OCTA), Int. J. Retina Vitreous 1 (1) (2015) 1–15.
[10] Y. Ma, et al., ROSE: a retinal OCT-angiography vessel segmentation dataset and new model, IEEE Trans. Med. Imaging 40 (3) (2020) 928–939.
[11] X. Liu, et al., Weakly supervised segmentation of covid19 infection with scribble annotation on ct images, Pattern Recogn. 122 (2022), 108341.
[12] Y. Liu, J. Shen, L. Yang, G. Bian, H. Yu, ResDO-UNet: A deep residual network for accurate retinal vessel segmentation from fundus images, Biomed. Signal Process. Control 79 (2023), 104087.
[13] X. Liu, S. Wang, Y. Zhang, Q. Yuan, Scribble-Supervised Meibomian Glands Segmentation in Infrared Images, ACM Trans. Multimedia Comput., Commun., Appl. (TOMM) 18 (3) (2022) 1–23.
[14] T. Pissas, et al., Deep iterative vessel segmentation in OCT angiography, Biomed. Opt. Express 11 (5) (2020) 2490–2510.
[15] Z. Peng et al., Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 367–376.
[16] A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, 2020.
[17] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.-C. Chen, Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 108–126.
[18] Y. Yu, H. Zhu, M3U-CDVAE: Lightweight retinal vessel segmentation and refinement network, Biomed. Signal Process. Control 79 (2023), 104113.
[19] Z. Wu, et al., PAENet: A Progressive Attention-Enhanced Network for 3D to 2D Retinal Vessel Segmentation, in: 2021 IEEE International Conference on BioInformatics and BiomedicIne (BIBM), IEEE, 2021, pp. 1579–1584.
[20] X. Deng, J. Ye, A retinal blood vessel segmentation based on improved D-MNet and pulse-coupled neural network, Biomed. Signal Process. Control 73 (2022), 103467.
[21] L. Mou, et al., CS-Net: channel and spatial attention network for curvilinear structure segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 721–730.
[22] Y. Guo, et al., An end-to-end network for segmenting the vasculature of three retinal capillary plexuses from OCT angiographic volumes, Biomed. Opt. Express 12 (8) (2021) 4889–4900.
[23] Z. Chen, Y. Xiong, H. Wei, R. Zhao, X. Duan, H. Shen, Dual-consistency semi-supervision combined with self-supervision for vessel segmentation in retinal OCTA images, Biomed. Opt. Express 13 (5) (2022) 2824–2834.
[24] A. Kirillov, Y. Wu, K. He, R. Girshick, PointRend: Image Segmentation As Rendering, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2020, pp. 9796–9805.
[25] Z. Liu, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, IEEE Computer Society, 2021, pp. 10012–10022.
[26] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 36–46.
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778.
[28] B. Antony, et al., Automated 3-D method for the correction of axial artifacts in spectral-domain optical coherence tomography images, Biomed. Opt. Express 2 (8) (2011) 2403–2416.
[29] M.K. Garvin, M.D. Abramoff, X. Wu, S.R. Russell, T.L. Burns, M. Sonka, Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images, IEEE Trans. Med. Imaging 28 (9) (2009) 1436–1447.
[30] K. Li, X. Wu, D.Z. Chen, M. Sonka, Optimal surface segmentation in volumetric images-a graph-theoretic approach, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2005) 119–134.
[31] Y. Zhang, et al., Robust layer segmentation against complex retinal abnormalities for en face OCTA generation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 647–655.
[32] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th international Conference on information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 327–331.
[33] L. Peng, L. Lin, P. Cheng, Z. Wang, X. Tang, Fargo: A joint framework for faz and rv segmentation from octa images, in: International Workshop on Ophthalmic Medical Image Analysis, Springer, 2021, pp. 42–51.
[34] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[35] H. Cao et al., Swin-unet: Unet-like pure transformer for medical image segmentation, arXiv preprint arXiv:2105.05537, 2021.
[36] J. Chen et al., Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, 2021.
[37] J.-H. Ri, G. Tian, Y. Liu, W.-H. Xu, J.-G. Lou, Extreme learning machine with hybrid cost function of G-mean and probability for imbalance learning, Int. J. Mach. Learn. Cybern. 11 (9) (2020) 2007–2020.
[38] V. Sathananthavathi, G. Indumathi, Encoder enhanced atrous (EEA) unet architecture for retinal blood vessel segmentation, Cogn. Syst. Res. 67 (2021) 84–95.
[39] Z. Shi, et al., MD-Net: A multi-scale dense network for retinal vessel segmentation, Biomed. Signal Process. Control 70 (2021), 102977.
[40] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, Sa-unet: Spatial attention u-net for retinal vessel segmentation, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 1236–1242.
[41] X. Liu, J. Cao, S. Wang, Y. Zhang, M. Wang, Confidence-guided topology-preserving layer segmentation for optical coherence tomography images with focus-column module, IEEE Trans. Instrum. Meas. 70 (2020) 1–12.
[42] S. Guo, CSGNet: Cascade semantic guided net for retinal vessel segmentation, Biomed. Signal Process. Control 78 (2022), 103930.
[43] X. Liu, et al., Automated layer segmentation of retinal optical coherence tomography images using a deep feature enhanced structured random forests classifier, IEEE J. Biomed. Health Inform. 23 (4) (2018) 1404–1416.