Article

CPMF-Net: Multi-Feature Network Based on Collaborative Patches for Retinal Vessel Segmentation

Wentao Tang, Hongmin Deng* and Shuangcai Yin

School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
* Correspondence: hm_deng@scu.edu.cn

Abstract: As an important basis of clinical diagnosis, the morphology of retinal vessels is very useful for the early diagnosis of some eye diseases. In recent years, with the rapid development of deep learning technology, automatic segmentation methods based on it have made considerable progresses in the field of retinal blood vessel segmentation. However, due to the complexity of vessel structure and the poor quality of some images, retinal vessel segmentation, especially the segmentation of Capillaries, is still a challenging task. In this work, we propose a new retinal blood vessel segmentation method, called multi-feature segmentation, based on collaborative patches. First, we design a new collaborative patch training method which effectively compensates for the pixel information loss in the patch extraction through information transmission between collaborative patches. Additionally, the collaborative patch training strategy can simultaneously have the characteristics of low occupancy, easy structure and high accuracy. Then, we design a multi-feature network to gather a variety of information features. The hierarchical network structure, together with the integration of the adaptive coordinate attention module and the gated self-attention module, enables these rich information features to be used for segmentation. Finally, we evaluate the proposed method on two public datasets, namely DRIVE and STARE, and compare the results of our method with those of other nine advanced methods. The results show that our method outperforms other existing methods.

Keywords: fundus image; vessel segmentation; training methods; self-attention; channel attention

1. Introduction

As important sensory organs, more than 70% [1] of information to the brain is provided via eyes, which are most key channels for people to perceive the world. However, affected by some retinal diseases, such as diabetic retinopathy, glaucoma, cataract and so on, many people are at a moderate or above visual risk. For these patients, if they cannot get timely medical intervention, they might suffer from deterioration of vision and even blindness with the development of these diseases. Therefore, timely diagnosis and treatment are very significant for relevant patients. According to medical researches [2,3], the appearance of retinal diseases is often accompanied by the morphological changes in retinal vessels. Clinical doctors usually find diseases by extensive experience through observing these changes. Therefore, how to automatically segment retinal vessels from fundus images by computer is of significance in assisting clinical doctors to diagnosis these diseases. In recent years, with the gradual demonstration of its potential application, deep learning has been widely used in the field of computer vision, including the segmentation of retinal vessels. Under continuous explorations, considerable results have been achieved by deep-learning-based segmentation methods. However, due to the complex and diverse structures of retinal vessels, the low contrast between vessels and background in fundus images, it is often difficult to correctly segment the densely connected fine vessels in retinal vessels, which remains a challenge. This paper is focused on two difficult problems in the field of retinal vessel segmentation, one is how to better segment the fine blood vessels in the fundus image, and the other is how to choose the network training mode, end-to-end mode or patch-based
mode, so as to better reduce the occupation of computing resources and the loss of image information. In order to address the above two problems, this paper proposes a multi-feature collaborative segmentation network and collaborative patch training strategy for retinal vessel segmentation. On the one hand, the fusion structure of the pre-segmentation network and main-segmentation network can effectively improve the segmentation details of vessel structure, especially the fine vessel structure. On the other hand, the designed collaborative patch training strategy not only has the advantages of simple vessel structure and less computational resources of the patch-based training method but also effectively reduces the information loss caused by patch extraction. The contributions of this paper can be summarized as follows:

1. A multi-feature segmentation network is proposed for retinal vessel segmentation. The two-level sub-networks complete the pre-segmentation and main-segmentation tasks, respectively. The pre-extraction of basic information of blood vessels in pre-segmentation and the cooperation of multiple information features in main-segmentation provide a lot of effective information for blood vessel segmentation, which improves the segmentation accuracy of the network, especially in the face of difficult blood vessels.

2. A collaborative patch training strategy is designed to reduce the information loss in the patch-based method. On the basis of patches, the segmentation method combining one small patch with simple vessel structure and five large patches with global information not only effectively retains the advantages of patch-based method but also effectively reduces the information loss caused by patch extraction.

3. An adaptive coordinate attention module is designed to extract the direction information of blood vessels. This module provides the model with very helpful vessel orientation information for vessel structure segmentation and improves the vessel continuity in the model segmentation results.

4. A gated self-attention module suitable for retinal image segmentation task is designed. The self-attention module integrated into the main-segmentation network can alleviate the local dependence of convolution operation and help the network to obtain long-distance dependence.

2. Related Work

2.1. Network Structure for Retinal Vessel Segmentation

In the past decade, accompanied by the rapid development of deep learning, convolutional neural networks (CNNs) have also been widely carried out in the field of image segmentation, including retinal vessel segmentation. Compared with retinal vessel segmentation methods based on classical classifiers such as support vector machine (SVM) and k-nearest neighbor (KNN), the retinal vessel segmentation methods [4–6] based on convolutional neural network can more effectively extract image features from fundus images. Motivated by the great potential of convolutional neural networks on image segmentation, researchers did a lot of work in this area and proposed quite valuable structures, such as fully convolutional networks (FCNs) [7]. By extending the FCNs, a network with U-shaped encoder–decoder architecture (U-Net) was proposed by Ronneberger et al. [8]. The “skip-connection” structure between the downsampling encoder and the upsampling decoder in the U-Net could combine the semantic information from the high-level features with the spatial information from the low-level features to jointly complete the segmentation. The great potential of U-Net in image segmentation made it a basic backbone widely used in the field of medical image segmentation, including retinal vessel segmentation. Fu et al. [9] applied the framework of multi-scale convolutional neural network to the segmentation of retinal vessels. Through the rich hierarchical representation and the application of conditional random fields, the segmentation accuracy of retinal vessels was improved. Mo et al. [10] developed a deep supervised FCN, which improved the clinical applicability of the network by using the multi-level characteristics of the deep network. Guo et al. [11] designed a channel attention residual U-Net, which improved the extraction
ability of retinal vessel features by adding modified efficient channel attention module to the U-Net. The above methods have improved the segmentation results of retinal vessels to a certain extent, but there are still deficiencies in the complex but very important task of small vessel segmentation. A multi-feature segmentation network is proposed in this paper. Through a cascade of stage segmentations and the integration of multiple effective attention modules, the network was effectively improved in the segmentation of fine blood vessels.

2.2. Training Method of the Model

In general, the training methods of retinal vessel segmentation network based on a convolutional neural network can be roughly divided into end-to-end and patch-based methods, where the end-to-end method has attracted the attention of many researchers because of its simple and stable characteristics in training. However, due to the high resolution of fundus images and limited computing resources, the end-to-end method that requires the network to learn the whole image usually includes downsampling of the image, but this would undoubtedly lose the spatial information that is very important for segmentation. Therefore, patch-based training has become a more widely used method in recent years because of its advantages such as less computational resources and simpler vessel morphology. Liskowski et al. [12] used deep learning to train the network on the large dataset via data enhancement and complete the segmentation of blood vessels. Wu et al. [13] proposed a multi-scale network following network (MS-NFN) model to address the problem of accurate segmentation of blood vessels, in which different upper poolings and lower poolings were used to complete the segmentation of blood vessels. After that, they further proposed an NFN+ model [14] based on network following network, which obtained rich samples for network training through patch extraction and data enhancement of images. Nevertheless, most of the above patch-based methods did not pay enough attention to the information loss caused by patch extraction. In order to eliminate the negative impacts of information loss on blood vessel segmentation, in this paper, a collaborative patch training strategy is designed. By extracting information from the large patches based on the target region patch during segmentation, the information loss of the target patch is effectively alleviated.

3. Methods

The flow chart of the multi-feature segmentation method based on collaborative patches proposed in this paper is shown in Figure 1, where the collaborative patch training strategy is used to effectively reduce the information loss of patches through the transmission of associated information between collaborative patches and improve the segmentation accuracy, while maintaining low computational resources. Moreover, the application of a multi-feature segmentation network further improves the accuracy of the segmentation for retinal vessels, especially for capillaries, through hierarchical structure and aggregation of multiple information features.

3.1. Multi-Feature Segmentation Network

The architecture of the multi-feature segmentation network designed in this paper is shown in Figure 2. The network is composed of a pre-segmentation sub-network, main-segmentation sub-network and edge extraction branch. Among them, the pre-segmentation sub-network is responsible for the rough segmentation of the image, the edge extraction branch is responsible for the extraction of blood vessel edges, and the main-segmentation sub-network is responsible for the fine segmentation of the image. Both segmentation sub-networks are based on the framework of U-Net, but for the more fine segmentation of vessel structure, we integrate an adaptive coordinate attention module for perceiving the direction information and a gated self-attention module for reducing the local dependence of convolution in the second cascade network. In addition, considering the significance of vessel edge information in vessel segmentation, we also design a learnable edge extraction
branch to extract the edge information of vessel structure from fundus images, which is next sent to the main-segmentation sub-network to improve the segmentation results.

Figure 1. Flow chart of the proposed method. The gray dotted box shows the patch extraction process. The extracted small patch corresponds to the target area, and each of the large patches corresponds to the combination of the target area and some global information.

Figure 2. Our proposed multi-feature segmentation network (MF-Net). The $3 \times 3$ and numbers (1, 32, 64, 128 and 256) in the gray rounded rectangle correspond to the size of the convolution kernel and the number of output channels, respectively. The network consists of the pre-segmentation network (orange box), the main-segmentation network (pink box) and the edge extraction branch (yellow box).

3.2. Adaptive Coordinate Attention Module

As one of the ways to enhance the model, an attention module has been widely used in image segmentation tasks. The attention module integrated into the network can improve the extraction of effective information, as well as the interpretability of network prediction by observing the attention weight value. Inspired by research [15], we design an adaptive coordinate attention module, as shown in Figure 3. The design of this module is the application of classical channel attention in the retinal segmentation task. The main motivation of the module design is listed in the following three aspects. First, the direction information of blood vessels is very useful for segmentation because the retinal vessel structure has choroidal characteristics. In order to better obtain the direction information, we adopt one-dimensional pooling along the horizontal and vertical directions separately to encode the information, which makes the module very sensitive to the direction of blood
vessels and improves the module extraction of the direction information of blood vessels. Second, in the task of retinal vessel segmentation, the detection and segmentation of micro-vessels remains a very challenging task. However, many existing attention modules only use the average pooling method to encode information, which blurs the difference between the capillaries and the background, resulting in the capillaries being easily regarded as the background area. Therefore, we add the maximum pooling code on the basis of average pooling, making the boundaries between the capillary region and the background region more distinct and improving the module’s perception of the capillary region. Third, at present, in order to reduce the information loss caused by pooling operation, many channel attention modules introduce a variety of weights into different poolings. Even so, the simple addition of weights without considering the sample situation may aggravate the loss of information. Therefore, we use a learnable adaptive weight when combining the two pooling coding methods, so that the module can choose the best combination method through learning. Specifically, for an input \( T \in \mathbb{R}^{C \times H \times W} \), we use one-dimensional pooling to encode each channel information along the horizontal and vertical dimensions. Therefore, the output of the \( c \)-th channel at the width \( w \) can be defined as:

\[
k_{c}^{aw}(w) = \frac{1}{H} \sum_{0 \leq i \leq H-1} T_c(i, w)
\]

\[
k_{c}^{am}(w) = \max([T_c(0, w), ..., T_c(H - 1, w)])
\]

Similarly, the output of the \( c \)-th channel at the height \( h \) can be defined as:

\[
k_{c}^{ah}(h) = \frac{1}{W} \sum_{0 \leq j \leq W-1} T_c(h, j)
\]

\[
k_{c}^{am}(h) = \max([T_c(h, 0), ..., T_c(h, W - 1)])
\]

For the generated feature, we splice it into the two directional features in the two branches, respectively, and then feed them to the convolution transform function \( F_1 \) shared by the two branches to obtain the intermediate features \( z_a \) and \( z_m \). They can be expressed as:

\[
z_a = \delta(F_1([k_{a}^{aw}, k_{a}^{ah}]))
\]
where \( \left[ \cdot \right] \) represents the concatenation operation on the spatial dimension, and \( \delta \) represents a non-linear activation function; the two intermediate features \( z_a \in \mathbb{R}^{C/r \times (H+W)} \) and \( z_m \in \mathbb{R}^{C/r \times (H+W)} \) are the features of horizontal and vertical spatial information encoded by different coding methods in the two branches, respectively. Where \( r \) is the reduction ratio, and \( \left[ \right] \) represents rounding operation. Then, the tensors in the two branches are split into two independent tensors based on the corresponding dimension information. The resulting tensor is the merge of two tensors in the same direction from two branches by adaptively adjusting weights. The resulting horizontal and vertical tensors are transformed to the same number of channels as the input through 1 × 1 transformation functions \( F_h \) and \( F_w \), respectively. The output can be formulated as:

\[
\begin{align*}
    z_a &= \delta \left( F_h \left( \left[ k_h^a, k_m^a \right] \right) \right) \\
    z_m &= \delta \left( F_w \left( \left[ k_h^m, k_m^m \right] \right) \right)
\end{align*}
\]

(6)

3.3. Gated Self-Attention Module

In recent years, Transformer has been brilliant in many fields including retinal vessel segmentation. However, if we want to apply Transformer to the field of retinal vessel segmentation, we have to face two problems. One is the high resolution of fundus images, which means a huge amount of computation in the calculation of self-attention. The second is that the training of Transformer is more prone to be unsatisfactory, especially the training of relative position bias. Inspired by research [16], we design a gated self-attention module from the above two problems. By modifying the traditional self-attention in the above two aspects, the gated self-attention module is more suitable for retinal vessel segmentation tasks. On the one hand, we reduce the negative impact of non-ideal relative position bias on the segmentation results by adaptive control of the relative position bias. On the other hand, our self-attention calculation is carried out in the feature map after cutting, which greatly reduces the amount of calculation brought by calculating the autocorrelation coefficient. At the same time, through the cooperation of different segmentation methods, we reduce the loss of inter-patch dependence brought by segmentation. The structure of the gated self-attention module is shown in Figure 4. The module is composed of a patch self-attention block, LayerNorm (LN) layer, residual connection and 2-layer MLP. The output of the module can be formulated as:

\[
\begin{align*}
    \tilde{t}_n &= \text{PSA} \left( \text{LN} \left( t_{n-1} \right) \right) + t_{n-1} \\
    t_n &= \text{MLP} \left( \text{LN} \left( \tilde{t}_n \right) \right) + \tilde{t}_n \\
    \tilde{t}_{n+1} &= \text{PSA} \left( \text{LN} \left( \text{Shift} \left( t_n \right) \right) \right) + \text{Shift} \left( t_n \right) \\
    t_{n+1} &= \text{MLP} \left( \text{LN} \left( \tilde{t}_{n+1} \right) \right) + \tilde{t}_{n+1}
\end{align*}
\]

(12) (13) (14) (15)
where $t^{n-1}$ is the input of the module, $t^{n+1}$ is the output of the module and PSA is the patch self-attention we designed. Similar to other self-attention mechanisms, the detail of PSA is shown in Figure 4. The patch self-attention output is:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + S_a \ast B \right)$$

where $Q, K$ and $V \in \mathbb{R}^{M^2 \times d}$ are query, key and value matrices, respectively, and $M^2$ and $d$ represent the number of blocks and the dimension of query or key matrix, respectively. $B \in \mathbb{R}^{M^2 \times M^2}$ represents the relative position bias matrix, and its value comes from the deviation matrix $\tilde{B} \in \mathbb{R}^{(2M+1) \times (2M-1)}$ [17,18].

Figure 4. Details of the designed gated self-attention module. The calculation details of patch self-attention are shown in the orange dotted line box.

3.4. Collaborative Patch Training Strategy

Patch-based methods usually require less computing resources than end-to-end methods, because the training model uses the extracted patches rather than the whole image. However, the process of extracting patches from the whole image will inevitably lead to the loss of information due to the separation of pixels at the edge of the patch. The loss of this part of information is not beneficial for the segmentation of vessel structures. Based on the above analysis, in order to reduce the negative impact of information loss on vessel segmentation, we design a collaborative patch training strategy. In addition to one small patch that completely coincides with the target region, in this strategy, the features of five additional large patches composed of target regions and part of its neighborhoods are also extracted. Compared with the small patch, the vascular structure in each of the large patches is more complete but also more complex, so their segmentation features are more complete, although the segmentation details are poor. A more complete vascular structure means that their features contain the correlation between the blood vessels in the target region and the blood vessels in the neighborhood. Therefore, we can obtain the associated information between the target region and the neighborhood by extracting the structural relationship between the two regions. The obtained associated information will participate in the segmentation of the small patch and promote the segmentation of the small patch by making up the lost associated information of the small patch. Through patch collaborative extraction and associated information transmission, the lost information can be compensated for to help the segmentation of vessel structure in this paper.

3.4.1. Patch Collaborative Extraction

Different from other patch-based methods, as shown in Figure 1, we not only extract a small patch that completely coincides with the target area but also extract five additional large patches that contain the target area and its surrounding areas. Among them, a small
patch with simple and easily segmented vessel structure provides the vessel structure information, and five large patches with more complete information supplement the associated information lost due to pixel separation on the edge of the target area. The setting of five collaborative patches can ensure that there are enough associated information sources in the target area no matter where the target area is located in the image. Specifically, the small patch $S$ completely coincides with the target area and is used to extract the main vessel structure information for the network. Large patches $L_1, L_2, \ldots, L_5$ are five combinations of the target area and surrounding areas, respectively, which are used to supplement the associated information lost due to pixel separation.

3.4.2. Associated Information Transmission

In order to effectively extract and transmit the associated information in the large patches $L_1, L_2, \ldots, L_5$, we design an associated information fusion module and an associated information correction network to transmit the large patch information in two stages to help segment the vessel structure of the target area. The structure of the associated information fusion module and the associated information correction network is shown in the Figure 5. The associated information fusion module is integrated between the pre-segmentation sub-network and the main-segmentation sub-network. It will extract the lost associated information due to pixel separation from the pre-segmentation results of the five large patches and adaptively combine the associated information in the five large patches through the internal attention module. Then, the resulting associated information features combine with the pre-segmentation results of the small patch into the main-segmentation stage.

![Figure 5. Details of the associated information fusion module and the associated information correction network. (a) An associated information correction sub-network for correcting the global information following the semantic segmentation network. (b) The channel attention module used in the associated correction network and associated fusion module. (c) An associated information fusion module used to supplement global associated information between two-level networks.](image)

Specifically, for each of the pre-segmentation features $f_{L1}^p, f_{L2}^p, \ldots, f_{L5}^p$ in five large patches, the first is a $3 \times 3$ convolution to obtain the associated information. Then, the associated information is combined adaptively through the attention module after splicing. Finally, the associated information is transformed by a $1 \times 1$ convolution to extract the associated information features $f_{Lm}^p$ of the target region. $f_{Lm}^p$ can be formulated as:

$$f_{Lm}^p = F_{1 \times 1} \left( CA \left( Concat \left( f_{F3 \times 3}^p \left( f_{L1}^p, f_{L2}^p, \ldots, f_{L5}^p \right) \right) \right) \right)$$  \hspace{1cm} (17)

where, $CA$ represents the channel attention operation, $F_{1 \times 1}$ and $F_{3 \times 3}$ represent the convolution transformations of $1 \times 1$ and $3 \times 3$, respectively, and $Concat$ represents the concatenation operation. Similarly, for the features $f_{m1}^p, f_{m2}^p, \ldots, f_{m5}^p$ and $f_m^p$ generated by the large patches and a small patch in the main-segmentation stage, the associated information in $f_{m1}^p, f_{m2}^p, \ldots, f_{m5}^p$ will be extracted again through the associated information
correction network and used to correct and supplement the information of the small patch. Specifically, for each of $f_{m}^{1}, f_{m}^{2}, \ldots, f_{m}^{l}$, firstly, a $3 \times 3$ convolution transform is used to extract each of the information features $\tilde{f}_{m}^{1}, \tilde{f}_{m}^{2}, \ldots, \tilde{f}_{m}^{l}$, respectively. On the one hand, the extracted information features are used to generate large segmentation results through a $1 \times 1$ convolution transform. On the other hand, $\tilde{f}_{m}^{1}, \tilde{f}_{m}^{2}, \ldots, \tilde{f}_{m}^{l}$ will be cut into the same size as $f_{m}^{s}$ according to the relationship with the target area, and after splicing, the channel attention module will adaptively combine them to generate correction information feature $f_{m}^{l}$ required by the small patch. Finally, the small patch feature $f_{m}^{s}$ and the correction information feature $f_{m}^{l}$ are sent into two convolution layers to generate the vessel segmentation results of the small patch. The correction information feature $f_{m}^{l}$ and the segmentation result $S_{s}$ of the small patch can be formulated as:

$$f_{m}^{l} = CA\left(\text{cut}\left(F_{1}\left(f_{m}^{s}, \tilde{f}_{m}^{1}, \ldots, \tilde{f}_{m}^{l}\right)\right)\right)$$  \hspace{1cm} (18)$$

$$S_{s} = F_{2}\left(F_{1}\left(f_{m}^{s}, f_{m}^{l}\right)\right)$$ \hspace{1cm} (19)$$

where, $CA$ represents the channel attention operation, $F_{1}$ and $F_{2}$ represent $3 \times 3$ and $1 \times 1$ convolution transforms, respectively, and $\text{cut}$ represents the cutting operation.

4. Experiments

4.1. Dataset

Two public datasets are used to evaluate the proposed multi-feature collaborative segmentation network and collaborative patch training strategy. The information of the datasets is shown in Table 1.

| Dataset | DRIVE | STARE |
|---------|-------|-------|
| Number of Images | 40 | 20 |
| Original Size | 584 × 565 | 700 × 605 |
| Patch Size | 72 × 72 | 72 × 72 |
| Tran/Test Split | 20/20 | 16/4 |

The DRIVE [19] dataset consists of 40 RGB fundus images, in which 20 images are selected as training sets and 20 images are selected as test sets. The resolution of each image is 565 × 584. For the test set, we use the first manual annotation to evaluate the performance of segmentation.

The STARE [20] data set consists of 20 RGB fundus images, including 10 pathological images and 10 normal images. The resolution of each image is 700 × 605. Among them, 16 images are selected as the training set and 4 images are selected as the test set.

For DRIVE datasets, each image has a corresponding field of view (FOV) mask. However, there is no corresponding FOV mask for the image in the STARE dataset, so we use the method proposed by Marin et al. [21] to generate the mask for the image. All index calculations of our experiments in this paper only consider the pixels in the FOV mask.

4.2. Pre-Processing

Considering that the low quality of some original fundus images may hinder the segmentation of vessels, some pre-processing methods are adopted to improve the quality of the fundus images. As shown in Figure 6, after graying, contrast-limited adaptive histogram equalization (CLAHE) and gamma correction, the display quality of blood vessel structure in the fundus image has been significantly improved. In order to avoid overfitting and improve the generalization ability of the model, a series of data enhancement methods are also adopted to increase the amount of data before training. The enhancement
methods include random rotation operation, random erase operation, up and down flip operation, left and right flip operation, etc.

**Figure 6.** Details of the pre-processing process.

### 4.3. Evaluation Metrics

In this section, a series of indicators are used to quantitatively and comprehensively evaluate the proposed model, including accuracy (ACC) reflecting the correct classification of pixels, sensitivity (SE) reflecting the correct detection of blood vessels, specificity (SP) reflecting the correct detection of background, F1-score (F1) reflecting the recall and accuracy and area under receiver operating characteristic curve (AUC), reflecting the comprehensive performance of multiple aspects. These indicators can be formulated as:

\[
\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}
\]  
\[
\text{SE} = \frac{TP}{TP + FN}
\]  
\[
\text{SP} = \frac{TN}{TN + FP}
\]  
\[
\text{F}_1 = \frac{2TP}{2TP + FN + FP}
\]  

In the formula, \( TP \) and \( FP \) are true positive and false positive, respectively, which represents the correct classification and wrong classification of pixels in which the model detects blood vessels. Accordingly, \( TN \) and \( FN \) are true negative and false negative, respectively, representing being correctly and wrongly classified into the background by the model, respectively.

### 4.4. Experimental Settings

Our model is based on the Pytorch framework and trained on an RTX2060(6GB). In the training process, the Adam optimizer is used. The initial learning rates of the training are 0.0025 and 0.002 in the two datasets, respectively. The learning rates in the two datasets decay 0.8 times every 20 and 8 epochs, respectively. To avoid over-fitting of the model, we adopted L2 regularization and set the weight decay of the optimizer to 0.000007.

### 5. Results

#### 5.1. Experiment of Training Strategy

In this part, we use two experiments to verify the ability of our collaborative patch training strategy in improving segmentation accuracy and saving computing resources. In terms of improving the segmentation accuracy, we compared the performance of our model with and without the collaborative patch training strategy on the two datasets to verify the effectiveness of the collaborative patch training strategy. Model MF-Net and model CPMF-Net for comparison maintain the same settings in all aspects, except in the collaborative patch training strategy. In addition, in order to improve the objectivity of the verification results, we also add the contrast experiment of the U-Net model and the U-Net model with the collaborative patch training strategy (CPU-Net). The Figure 7 shows the
qualitative segmentation results of the four models. It can be found that the segmentation results of U-Net and MF-Net are inferior to others in the continuity of the vessel structure, and there are wrong segmentations of large areas. In contrast, the segmentation results of the CPU-Net model and CPMF-Net model with collaborative patch training strategy have both achieved good performances in the continuity of vessel structure and the accuracy of segmentation. The difference between the segmentation results of the same group of models reflects whether the information loss caused by patch extraction can be compensated. The segmentation results of the two baseline models after U-Net and MF-Net applying the collaborative patch training strategy show that this strategy can improve the accuracy of model segmentation by supplementing the lost information. The quantitative comparison result of the models show in Table 2. It can be found that the collaborative patch training strategy has achieved very good performance in improving the segmentation accuracy, whether in an MF-Net model or a U-Net model. Specifically, all the indicators are improved on two datasets by introducing the collaborative patch training strategy, except SP index of the STARE dataset.

![Figure 7. Model segmentation results using different training strategies.](image)

Figure 7. Model segmentation results using different training strategies. The top row is based on DRIVE datasets, and the bottom row is STARE datasets. (a) The original image, (b) the ground truth, (c) the segmentation result of U-Net, (d) the segmentation result of CPU-Net, (e) the segmentation result of MF-Net and (f) the segmentation result of our CPMF-Net.

Table 2. Quantitative comparison of the performance of different training methods in segmentation accuracy.

| Model       | DRIVE          | STARE          |
|-------------|----------------|----------------|
|             | F1 (%) SE (%) SP (%) ACC (%) AUC (%) | F1 (%) SE (%) SP (%) ACC (%) AUC (%) |
| U-Net       | 81.36 77.92 98.12 95.61 97.66 83.27 82.95 | 98.15 96.60 98.76 |
| CPU-Net     | 82.55 82.77 97.54 95.70 98.07 84.67 86.21 | 98.02 96.81 98.94 |
| MF-Net      | 82.10 82.19 97.50 95.59 97.84 83.64 80.77 | 98.60 96.78 98.84 |
| CPMF-Net    | 82.94 83.45 97.53 95.78 98.19 85.66 86.81 | 98.20 97.03 99.16 |

As for saving computing resources, we use floating-point operations per second (Flops) and total memory occupation (Memory) indicators to make a quantitative comparison among the models based on whole-graph training, patch training and collaborative patch strategy training, respectively. Table 3 shows the quantitative comparison results among the models. It can be found that the model based on the whole-graph training (WMF-Net) has the maximum amount of floating-point calculation and the maximum computing memory requirement. Although the model based on collaborative patch strategy training has a small increase in computation and memory usage compared with the model based on single-patch training, the overall computation and memory usage still remain at a low level. Moreover, the performance of the collaborative patch training strategy improved the segmentation accuracy of the model, which means the collaborative patch training strategy can adopt smaller patch size under the same segmentation accuracy, achieving
lower computational load and memory requirements, which is very helpful for the clinical application of the computer-aided diagnosis.

### Table 3. Quantitative comparison of the computational resource occupation among different training methods.

| Model    | Flops  | Memory   |
|----------|--------|----------|
| MF-Net   | 2.04 G | 38.43 M  |
| CPMF-Net | 2.05 G | 41.25 M  |
| WMF-Net  | 141.51 G | 2651 M |

#### 5.2. Experiment of Segmentation Model

In this sub-section, we investigate the roles of each of the two attention mechanisms, namely adaptive coordinate attention module and gated self-attention module, and their fusion in our model.

##### 5.2.1. Experiment of Adaptive Coordinate Attention Module

In this part, we first verify the effectiveness of the adaptive coordinate attention (ACA) module without using the gated self-attention (GSA) module. Then, we compare our module with other channel attention modules: (1) squeeze-and-excitation attention module (SE) and (2) classical coordinate attention module (CA). Figure 8 shows the qualitative segmentation results of the four models. It can be seen from the figure that the models with the attention mechanism such as SE, CA, or ACA have improved the segmentation results compared with Basenet, especially, our model integrated with the ACA module has achieved the best performance in the detection capability of small vessels and the continuity of vessel structure segmentation. Table 4 shows the quantitative comparison between models, where Basenet represents the model after using convolution to replace all the adaptive coordinate attention modules in the model, SE and CA represent the model after using corresponding attention modules, respectively. It can be seen from the table that compared with Basenet, the ACA module has improved in all indicators except the ACC indicator on the DRIVE dataset. Compared with the models with the SE or CA modules, the one with the ACA module has achieved the best performance in F1/ACC/AUC on two datasets, reaching 82.82%/95.67%/98.18% and 85.14%/97.07%/99.15%, respectively, which shows that our adaptive coordinate attention module is effective in improving segmentation accuracy.

![Figure 8](image)

**Figure 8.** The model segmentation results with different channel attention modules integrated. The top row is the result on the DRIVE dataset, and the bottom row is the one on the STARE dataset. (a) The original image, (b) the ground truth, (c) the segmentation result of Basenet, (d) the segmentation result integrated with SE module, (e) the segmentation result integrated with CA module and (f) the segmentation result integrated with ACA.
Table 4. Quantitative performance of different channel attention modules in retinal segmentation task.

| Model | DRIVE | STARE |
|-------|-------|-------|
|       | F1 (%) | ACC (%) | AUC (%) | F1 (%) | ACC (%) | AUC (%) |
| Basenet | 82.03 | 95.74 | 98.06 | 84.84 | 97.01 | 99.13 |
| SE | 82.49 | 95.60 | 98.11 | 84.95 | 96.72 | 99.12 |
| CA | 82.46 | 95.62 | 98.14 | 84.66 | 96.59 | 99.15 |
| ACA | 82.82 | 95.67 | 98.18 | 85.14 | 97.07 | 99.15 |

5.2.2. Experiment of Gated Self-Attention Module

In this part, we first verify the effectiveness of the gated self-attention module without adaptive coordinate attention. Then, in order to verify that our proposed gated self-attention module can alleviate the negative impact caused by insufficient training when there are few samples, we also compared the gated self-attention module with the Swin-Transformer (SWT) module. Figure 9 shows the comparison results of qualitative segmentation among the models. It can be seen that on the DRIVE dataset, the GSA model gives full play to the advantages of self-attention and has the best performance in the detection and segmentation of fine blood vessels. On the STARE dataset, due to the relatively small amount of data and large changes in image quality, the SW model has achieved the poorest segmentation results. In contrast, due to adaptive control, the segmentation results of the GSA model do not deteriorate significantly. The good segmentation results on the two datasets show the effectiveness of the self-attention module of our gated self-attention module. Table 5 shows the quantitative comparison results among three variants of our model, where Basenet and SW are the models where a convolutional block and Swin self-attention block are used in place of the gated self-attention module, respectively. The three models have the same settings as that of the Basenet, except the adding module to be substituted in different models. It can be seen from the Table 5 that the indicators of SW on both datasets have decreased compared with that of Basenet, which is caused by inadequate training of SWT on smaller datasets. In contrast, the GSA we designed has the ability to mitigate the negative impact of inadequate training, so it has achieved the best performance on both datasets.

Figure 9. The model segmentation results of different self-attention modules integrated. The top row is the based on DRIVE dataset, and the bottom row is the STARE dataset. (a) The original image, (b) the ground truth, (c) the segmentation result of Basenet, (d) the segmentation result of the integrated SW module and (e) the segmentation result of the our model integrated with the GSA module.
Table 5. Quantitative performances of the models based on different self-attention modules in retinal segmentation task.

| Model   | DRIVE          | STARE          |
|---------|----------------|----------------|
|         | F1 (%) | ACC (%) | AUC (%) | F1 (%) | ACC (%) | AUC (%) |
| Basenet | 82.03   | 95.74   | 98.06   | 84.84  | 97.01   | 99.13   |
| SW      | 81.54   | 95.73   | 98.11   | 84.86  | 96.97   | 99.07   |
| GSA     | 82.04   | 95.75   | 98.15   | 85.27  | 97.15   | 99.14   |

5.2.3. Ablation Experiment

In order to verify the further effectiveness of the combination of the two attention modules, we conducted ablation experiments. The quantitative results are shown in Table 6. It can be found that the integration of the two attention modules improves the ability of the model to obtain specific features, and when the two attention modules are integrated with each other at the same time, the model can combine the advantages of the two modules to implement better segmentation.

Table 6. Ablation experiments on DRIVE and STARE. Basenet represents the basic model using convolutional blocks and without the two additional attention modules.

| Model        | DRIVE          | STARE          |
|--------------|----------------|----------------|
|              | F1 (%) | ACC (%) | AUC (%) | F1 (%) | ACC (%) | AUC (%) |
| Basenet      | 82.03   | 95.74   | 98.06   | 84.84  | 97.01   | 99.13   |
| Basenet+GSA  | 82.04   | 95.75   | 98.15   | 85.27  | 97.15   | 99.14   |
| Basenet+ACA  | 82.82   | 95.67   | 98.18   | 85.14  | 97.07   | 99.15   |
| Basenet+GSA+ACA | 82.94 | 95.78   | 98.19   | 85.66  | 97.03   | 99.16   |

5.3. Comparison with the State-of-the-Art Methods

In this part, we compare our method with the other nine advanced methods. Figure 10 intuitively shows the segmentation results of our model and some advanced models. It can be seen from the figure that the segmentation results of our model are the closest to the ground truth marked by experts, and our model CPMF-Net has the best performance in terms of segmentation integrity and continuity. The results of quantitative comparison are shown in Table 7. It can be found that our method achieves the best performance in three of all five indicators, including F1, ACC and AUC on DRIVE dataset, reaching 82.94%, 95.78% and 98.19%, respectively. Meanwhile, our method also achieves the best performance in F1, SE, ACC and AUC indicators on the STARE dataset, and the SE indicator among them is 2.49% higher than that of CSU-Net, which ranks second. Figure 11 shows the ROC curves of our model and some advanced models, from which we can find that our model outperforms the other models on the ROC curve, reflecting the overall performance of segmentation.

In addition, considering the importance of generalization ability to the practical application of the method [22,23], we also evaluated the generalization ability of our method through cross-training. Table 8 shows the quantitative comparison of the generalization performance between our method and the other three existing methods. It can be found that our method has better generalization capability than other methods on both datasets. It has achieved the optimal performance of all indicators on the DRIVE dataset and the optimal performance on the STARE dataset, except for the SP indicator.
Figure 10. Comparison of segmentation results between our approach and some other advanced approaches. (a) The ground truth, (b) the segmentation result of UNet model, (c) the segmentation result of R2Unet model, (d) the segmentation result of CE-Net model and (e) the segmentation result of our model.

Table 7. Comparison with the most advanced methods on DRIVE and STARE.

| Model              | Year | DRIVE F1 (%) | SE (%) | SP (%) | ACC (%) | AUC (%) | STARE F1 (%) | SE (%) | SP (%) | ACC (%) | AUC (%) |
|--------------------|------|--------------|--------|--------|---------|---------|--------------|--------|--------|---------|---------|
| U-Net [8]          | 2015 | 81.36        | 77.92  | 98.12  | 95.61   | 97.66   | 82.95        | 98.15  | 96.60  | 98.76   |
| R2U-Net [24]       | 2018 | 78.07        | 83.05  | 95.86  | 95.95   | 95.95   | 77.58        | 79.62  | 97.08  | 95.30   | 97.17   |
| CE-Net [25]        | 2019 | 78.64        | 77.78  | 97.21  | 94.80   | 97.11   | 82.74        | 84.04  | 97.63  | 96.32   | 98.68   |
| Xu et al [26]      | 2020 | 82.52        | 79.53  | 98.07  | 95.57   | 98.04   | 83.08        | 83.78  | 97.41  | 95.90   | 98.17   |
| Zhou et al [27]    | 2020 | 80.35        | 74.73  | 98.35  | 95.35   | 97.13   | 81.32        | 77.76  | 96.05  | 97.40   |
| Li et al [28]      | 2021 | 79.21        | 98.10  | 95.68  | 98.06   | -       | 83.52        | 98.23  | 97.82  | 98.75   |
| CSU-Net [29]       | 2021 | 82.51        | 80.71  | 98.01  | 95.65   | 98.01   | 84.32        | 98.45  | 97.02  | 98.25   |
| Bridge-Net [30]    | 2022 | 82.03        | 78.53  | 95.18  | 95.65   | 98.34   | 82.89        | 80.02  | 98.64  | 96.68   | 99.01   |
| Li et al [31]      | 2022 | 82.88        | 83.59  | 97.31  | 95.71   | 98.10   | 83.63        | 83.52  | 98.23  | 96.71   | 98.75   |
| CPMF-Net(ours)     | 2022 | 82.94        | 83.54  | 97.53  | 95.78   | 98.19   | 85.66        | 86.81  | 98.20  | 97.03   | 99.16   |

Table 8. Cross-training results on DRIVE and STARE datasets.

| Test Set | Training Set | Model      | SE (%) | SP (%) | ACC (%) | AUC (%) |
|----------|--------------|------------|--------|--------|---------|---------|
| STARE    | DRIVE        | Fraz [22]  | 72.42  | 97.92  | 94.56   | 96.97   |
|          |              | Li [23]    | 72.73  | 98.10  | 94.86   | 96.77   |
|          |              | Yan [32]   | 72.92  | 98.15  | 94.94   | 95.99   |
|          |              | CPMF-Net(ours) | 75.93  | 98.15  | 95.39   | 97.53   |
| DRIVE    | STARE        | Fraz [22]  | 70.10  | 97.70  | 94.95   | 96.71   |
|          |              | Li [23]    | 70.27  | 98.28  | 95.45   | 96.71   |
|          |              | Yan [32]   | 72.11  | 98.40  | 95.69   | 97.08   |
|          |              | CPMF-Net(ours) | 80.24  | 98.12  | 96.04   | 98.51   |

The excellent performance in SE and AUC indicators means that our proposed method has stronger vessel perception ability while maintaining segmentation accuracy. This ability enables the model to better detect and segment fine vessel structures, which is very helpful for the computer-aided diagnosis in the clinical diagnosing application of early ocular diseases.
Figure 11. ROC curves of our model and some advanced models.

6. Conclusions

In this paper, according to the characteristics of the retinal blood vessel segmentation task, we propose a new retinal blood vessel segmentation method, namely the multi-feature segmentation method based on collaborative patches. By combining a powerful multi-feature network with an effective collaborative patch training strategy, high-precision segmentation without the extremely rigid hardware condition can be achieved. The experimental results on the DRIVE and STARE datasets show the effectiveness and great application potential of our approach. For the future work, we plan to optimize the feature extraction ability of the segmentation model and the associated information compensation ability of the collaborative patch training strategy.

Author Contributions: Conceptualization, methodology and writing, W.T.; editing and supervision, H.D.; visualization, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0553 and in part by the National Natural Science Foundation of China under Grant 62020106010.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this work are publicly available online.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Diao, Y.; Chen, Y.; Zhang, P.; Cui, L.; Zhang, J. Molecular guidance cues in the development of visual pathway. Protein Cell 2018, 9, 909–929. [CrossRef] [PubMed]
2. Kipli, K.; Hoque, M.E.; Lim, L.T.; Mahmood, M.H.; Sahari, S.K.; Sapawi, R.; Rajaee, N.; Joseph, A. A review on the extraction of quantitative retinal microvascular image feature. Comput. Math. Methods Med. 2018, 2018, 4019538. [CrossRef] [PubMed]
3. Cheung, C.Y.L.; Zheng, Y.; Hsu, W.; Lee, M.L.; Lau, Q.P.; Mitchell, P.; Wang, J.J.; Klein, R.; Wong, T.Y. Retinal vascular tortuosity, blood pressure, and cardiovascular risk factors. Ophthalmology 2011, 118, 812–818. [CrossRef] [PubMed]
4. Jiang, Y.; Liang, J.; Cheng, T.; Lin, X.; Zhang, Y.; Dong, J. MTPA_Unet: Multi-scale transformer-position attention retinal vessel segmentation network joint transformer and CNN. Sensors 2022, 22, 4592. [CrossRef] [PubMed]
5. Hu, X.; Wang, L.; Li, Y. HT-Net: A hybrid transformer network for fundus vessel segmentation. Sensors 2022, 22, 6782. [CrossRef] [PubMed]
6. Jiang, Y.; Yao, H.; Tao, S.; Liang, J. Gated skip-connection network with adaptive upsampling for retinal vessel segmentation. Sensors 2021, 21, 6177. [CrossRef] [PubMed]
7. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 640–651. [CrossRef] [PubMed]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention MICCAI 2015—18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.

9. Fu, H.; Xu, Y.; Lin, S.; Kee Wong, D.W.; Liu, J. DeepVessel: Retinal vessel segmentation via deep learning and conditional random field. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016—19th International Conference, Athens, Greece, 17–21 October 2016; pp. 132–139.

10. Mo, J.; Zhang, L. Multi-level deep supervised networks for retinal vessel segmentation. Int. J. Comput. Assist. Radiol. Surg. 2017, 12, 2181–2193. [CrossRef] [PubMed]

11. Guo, C.; Szemenyei, M.; Hu, Y.; Wang, W.; Zhou, W.; Yi, Y. Channel attention residual U-Net for retinal vessel segmentation. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–12 June 2021; pp. 1185–1189.

12. Liskowski, P.; Krawiec, K. Segmenting retinal blood vessels with deep neural networks. IEEE Trans. Med. Imag. 2016, 35, 2369–2380. [CrossRef]

13. Wu, Y.; Xia, Y.; Song, Y.; Zhang, Y.; Cai, W. Multiscale network followed network model for retinal vessel segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018—21st International Conference, Granada, Spain, 16–20 September 2018; pp. 119–126.

14. Wu, Y.; Xia, Y.; Song, Y.; Zhang, Y.; Cai, W. NFN+: A novel network followed network for retinal vessel segmentation. Neural Netw. 2020, 126, 153–162. [CrossRef] [PubMed]

15. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, Nashville, TN, USA, 19–25 June 2021; pp. 13708–13717.

16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.

17. Han, H.; Gu, J.; Zheng, Z.; Dai, J.; Wei, Y. Relation networks for object detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3588–3597.

18. Hu, H.; Zhang, Z.; Xie, Z.; Lin, S. Local relation networks for image recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3464–3473.

19. Staal, J.; Abràmoff, M.D.; Niemeijer, M.; Viergever, M.A.; van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imag. 2004, 23, 501–509. [CrossRef] [PubMed]

20. Hoover, A.D.; Koznetsova, V.; Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imag. 2000, 19, 203–210. [CrossRef] [PubMed]

21. Gégundez-Arias, M.E.; Aquino, A.; Bravo, J.M.; Marin, D. A function for quality evaluation of retinal vessel segmentations. IEEE Trans. Med. Imag. 2012, 31, 231–239. [CrossRef] [PubMed]

22. Fraz, M.M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnica, A.R.; Owen, C.G.; Barman, S.A. Blood vessel segmentation methodologies in retinal images—A survey. Comput. Methods Programs Biomed. 2012, 108, 407–433 [CrossRef]

23. Li, Q.; Feng, B.; Xie, L.; Liang, P.; Zhang, H.; Wang, T. A cross-modality learning approach for vessel segmentation in retinal images. IEEE Trans. Med. Imaging 2015, 35, 109–118. [CrossRef] [PubMed]

24. Alom, M.Z.; Yakopcic, C.; Hasan, M.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. J. Med. Imag. 2019, 6, 014006. [CrossRef]

25. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-Net: Context encoder network for 2D medical image segmentation. IEEE Trans. Med. Imag. 2019, 38, 2281–2292. [CrossRef]

26. Xu, R.; Ye, X.; Jiang, G.; Liu, T.; Li, L.; Tanaka, S. Retinal vessel segmentation via a semantics and multi-Scale aggregation network. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020, Barcelona, Spain, 4–8 May 2020; pp. 1085–1089.

27. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. U-Net++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imag. 2020, 39, 1856–1867. [CrossRef]

28. Li, X.; Jiang, Y.; Li, M.; Yin, S. Lightweight attention convolutional neural network for retinal vessel image segmentation. IEEE Trans. Ind. Inf. 2021, 17, 1958–1967. [CrossRef]

29. Wang, B.; Wang, S.; Qiu, S.; Wei, W.; Wang, H.; He, H. CSU-Net: A context spatial U-Net for accurate blood vessel segmentation in fundus images. IEEE J. Biomed. Health. Inf. 2021, 25, 1128–1138. [CrossRef] [PubMed]

30. Zhang, Y.; He, M.; Chen, Z.; Hu, K.; Li, X.; Gao, X. Bridge-Net: Context-involved U-net with patch-based loss weight mapping for retinal blood vessel segmentation. Expert Appl. 2022, 195, 116526. [CrossRef]

31. Li, Y.; Zhang, Y.; Cui, W.; Lei, B.; Kuang, X.; Zhang, T. Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation. IEEE Trans. Med. Imag. 2022, 41, 1975–1989. [CrossRef]

32. Yan, Z.; Yang, X.; Cheng, K. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. IEEE Trans. Biomed. Eng. 2018, 65, 1912–1923. [CrossRef] [PubMed]