electronics

MDPI

Article

Self-Paced Dual-Axis Attention Fusion Network for Retinal Vessel Segmentation

Yueting Shi [1,2], Weijiang Wang [1], Minzhi Yuan [1,2] and Xiaohua Wang [1,*]

[1] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China; shiyueting@bit.edu.cn (Y.S.)
[2] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing 314019, China
* Correspondence: xh_wong@bit.edu.cn

Abstract: The segmentation of retinal vessels plays an essential role in the early recognition of ophthalmic diseases in clinics. Increasingly, approaches based on deep learning have been pushing vessel segmentation performance, yet it is still a challenging problem due to the complex structure of retinal vessels and the lack of precisely labeled samples. In this paper, we propose a self-paced dual-axis attention fusion network (SPDAA-Net). Firstly, a self-paced learning mechanism using a query-by-committee algorithm is designed to guide the model to learn from easy to hard, which makes model training more intelligent. Secondly, during fusing of multi-scale features, a dual-axis attention mechanism composed of height and width attention is developed to perceive the object, which brings in long-range dependencies while reducing computation complexity. Furthermore, CutMix data augmentation is applied to increase the generalization of the model, enhance the recognition ability of global and local features, and ultimately boost accuracy. We implement comprehensive experiments validating that our SPDAA-Net obtains remarkable performance on both the public DRIVE and CHASE-DB1 datasets.

Keywords: retinal vessel segmentation; dual-axis attention; self-paced learning

Citation: Shi, Y.; Wang, W.; Yuan, M.; Wang, X. Self-Paced Dual-Axis Attention Fusion Network for Retinal Vessel Segmentation. *Electronics* **2023**, *12*, 2107. https://doi.org/10.3390/electronics12092107

Academic Editors: Cecilia Di Ruberto, Alessandro Stefano, Albert Comelli, Lorenzo Putzu and Andrea Loddo

Received: 22 March 2023
Revised: 30 April 2023
Accepted: 30 April 2023
Published: 5 May 2023

1. Introduction

Retinal vessel segmentation is a process of extracting the blood vessels from retinal fundus images that can help with quantitative analysis and diagnosis of retinal diseases caused by stroke, diabetes, hypertension, and glaucoma. Morphological properties of retinal vessels, such as vessel diameter, tortuosity, bifurcation pattern, and so on are important biomarkers for these diseases [1]. Thus, it is required to segment retinal vessels with high accuracy. On the other hand, with the development of computer-aid diagnose methods, the automatic segmentation technology of retinal vessels can ease the workload of ophthalmologists and assist the inexperienced ones. Therefore, the automatic segmentation of retinal vessels is important for the clinical diagnosis and treatment of eye diseases.

However, retinal vessel segmentation is not an easy task due to various challenges that affect the image quality and segmentation performance. First, anatomical variability between subjects and thin retinal vessels are difficult for identification. Second, the number of training samples accurately labeled by doctors is limited. Third, there can be low contrast, noise, and uneven illumination in blood vessel images. Additionally, vessels can be disturbed or covered by other components in the retina, such as the optic disc or lesions.

Efforts have been dedicated for the above challenges by proposing deep-neural-network-based methods [1–6]. Many of them have attempted to improve U-Net [7] by incorporating some carefully designed specific network modules. Wang et al. [4] proposed the Dual-Encoding U-Net (Dual-E U-Net) that remarkably enhanced the network's capability to segment retinal vessels in an end-to-end and pixel-to-pixel way. Gu et al. [6] proposed a context encoder network (CE-Net) to capture more high-level information and

preserve spatial information for 2D medical image segmentation. Wang et al. [1] proposed the hard attention net (HAnet) that enhanced the ability to segment hard and easy regions independently from three decoder branches. U-Net structures obtain high-level feature representations by stacked convolution and down-sampling and utilize mirrored operations with skip connections to recover the original-resolution feature maps for pixel-level supervised learning.

Although existing methods have continued to improve the segmentation accuracy and sensitivity, some limitations still exist. First, they lack the ability to learn the semantic information of tiny-scale vessels since valuable spatial information is diminished after downsampling, and they lack the ability to model long-range dependencies that are present in an image. Secondly, retinal vessel segmentation lacks large-scale labeled datasets. How to take full advantage of the data to guide the model to train at an appropriate speed and level of difficulty needs to be further explored.

In this paper, we propose a self-paced dual-axis attention fusion network for retinal vessel segmentation. Firstly, inspired by the learning process of humans, a self-paced learning method based on a query-by-committee (QBC) algorithm is designed that calculates the difficulty of samples through our designed committee blocks and selects input samples for training in each epoch with an adaptive threshold from easy to hard, hence improving model performance with limited data. Secondly, a dual-axis attention module is introduced to the encoder of the subsequent U-Net. The height-axis and width-axis attention enrich the capability of extracting precise vascular structural features while reducing computational complexity more than 2D attention. Thirdly, CutMix data augmentation is applied for vessel segmentation to improve model robustness. CutMix prevents overfitting and boosts classification accuracy by cutting out parts of one image and pasting them over another image while also mixing their labels according to the area ratio.

The contributions of this paper are summarized as follows:

- An end-to-end retinal vessel segmentation network with dual-axis attention mechanism and self-paced learning method is proposed to improve model sensitivity and accuracy.
- A width and height dual-axis attention module is employed to fuse the features with different scales and to integrate local and global semantics for better segmentation of various vessels.
- A human-inspired self-paced learning method that leverages a query-by-committee algorithm to assign the easiness weight of training sample is proposed for vessel segmentation and guides the model to learn from easy to hard, avoid local optima, and boost accuracy. To the best of our knowledge, this is the first work that applies self-paced learning for this task.
- We evaluate our framework on the DRIVE and CHASE-DB1 datasets. The designed self-paced learning loss is used as the loss function, and CutMix data augmentation is applied to increase the diversity of the training data. The experimental results demonstrate that our method achieves remarkable performance on both datasets.

2. Related Work

2.1. Retinal Vessel Segmentation

The structural information of retinal blood vessels has important guiding significance for the diagnosis of ophthalmic diseases, Accurate segmentation of retinal blood vessels has become an urgent clinical need. To improve the segmentation accuracy a lot of work has been done by numerous researchers.

To solve low vessel edge visibility and high vessel complexity, Wang et al. [1] proposed the hard attention net (HAnet), which can pay more attention to the vessels with high complexity. He et al. [8] proposed an evolvable adversarial framework that has fewer mis-segmented regions and more accurate boundaries. Liu et al. [9] proposed the OCTA retinal vessel segmentation method (ARP-Net) based on the adaptive gated axial transformer

(AGAT), Residual and point repair modules guide the network to focus on low vessel edge visibility.

To solve the problem that some features related to detailed structures are not discriminative enough, Zhang et al. [10] proposed a boundary enhancement and feature denoising (BEFD) module to prevent loss of high-frequency information. Xu et al. [11] proposed a local-region and cross-dataset contrastive learning method to enhance the tiny-feature embedding ability. Wang et al. [12] proposed a context-spatial U-Net with a two-channel encoder that can capture spatial information and multi-scale context information, respectively.

2.2. Self-Paced Learning

Self-paced learning (SPL) is a method that adaptively prioritizes easy and reliable samples before complex ones. It helps models learn better and avoid local optima.

Kumar et al. [13] introduced a self-paced regular function to the supervised algorithm by optimizing the improved objective function; the SPL not only optimizes the parameters of the supervised algorithm but also continuously improves the course. They used binary weights to select simple samples for classification and segmentation.

Jiang et al. [14] used self-paced learning to detect movements and events; they proposed a soft weight rule that assigns floating-point weights to measure the difficulty and diversity of samples. They proposed a self-paced learning-with-diversity (SPLD) method that extracts easy samples from diverse sets for fast access to complex knowledge and to obtain better performance.

Li et al. [15] used multi-objective optimization to design a self-paced learning method (MOSPL). They split the objective function into two parts and optimized them separately. They obtained a compromise value as the final result.

Since self-paced learning has evolved to the present state, the main difficulty is how to let the model distinguish the difficulty of training samples and how to control the learning speed, specifically, the transition from learning easy samples to difficult samples. Therefore, further methods that can better distinguish the difficulty of samples and adaptively optimize the thresholds of input samples need to be developed.

From the perspective of vessel segmentation, the major challenge is the complex structure of blood vessels that makes it difficult to separate blood vessels from the surrounding tissues. The step-by-step learning mode of self-paced learning from easy to hard may contribute to precise vessels segmentation.

2.3. Active Learning and Query by Committee

Active learning is a machine learning algorithm that finds the most valuable training samples to add to the training set by actively finding them, and if those samples are unlabeled, it automatically asks for manual labeling before using them for model training. In simple terms, it means training a model with as high performance as possible with fewer training samples.

The most widely used current active learning algorithms are the informativeness-based methods, which select as training data the samples with the highest uncertainty about the current model. Another popular class of methods is query by committee, which select as training data the samples with the most uncertainty for different models. In the following, we focus on the analysis of query by committee.

The basic idea of query by committee (QBC) is to view the optimized machine learning model as a version space search, and QBC finds the best machine learning model by compressing the search scope of the version space [16]. The steps of active learning based on QBC are to train multiple models with the same architecture on the same training set. Uncertain samples are identified by voting among the models, and then they are labeled and added to the training set. This process is repeated until satisfactory performance is achieved. It has been found that a committee size of two or three models is sufficient, and that diversity among the models is essential for effective integration.

The formula for calculating the voting entropy of the QBC method is as follows:

$$V_{qbc} = -\arg\max \sum_{i} \frac{Vot(y_i)}{C} \log \frac{Vot(y_i)}{C} \quad (1)$$

where C is the number of categories in the classification and $Vot(y_i)$ is the number of votes received by category y_i.

In QBC, each sample is sent to a different member of the committee, and then each member votes on the sample, judging the difficulty of the sample based on the consistency of the voting results: if the voting results are similar, it is an easy sample; conversely, if the voting results have great difference, it is a hard sample. This implies that the QBC method is an effective and reliable technique for assessing sample complexity.

3. Our Proposed Method

3.1. Framework Overview

As illustrated in Figure 1, our self-paced dual-axis attention fusion network (SPDAA-Net) consists of a series of committee blocks for the self-paced learning mechanism and a subsequent dual-axis attention U-like-Net for precise vessel segmentation. First, the input image is cropped into patches and CutMix data augmentation is applied to increase the robustness of the model. Then, a self-paced learning mechanism is adopted by using committee blocks to measure the difficulty of the input. The subsequent U-like-net is designed with dual-axis attention for height and width axes, which contributes to higher segmentation accuracy for vessel segmentation while reducing computational complexity. Finally, the network is trained with samples from easy to hard according to the self-paced learning, and the predicted segmentation result is obtained after passing through the fusion network.

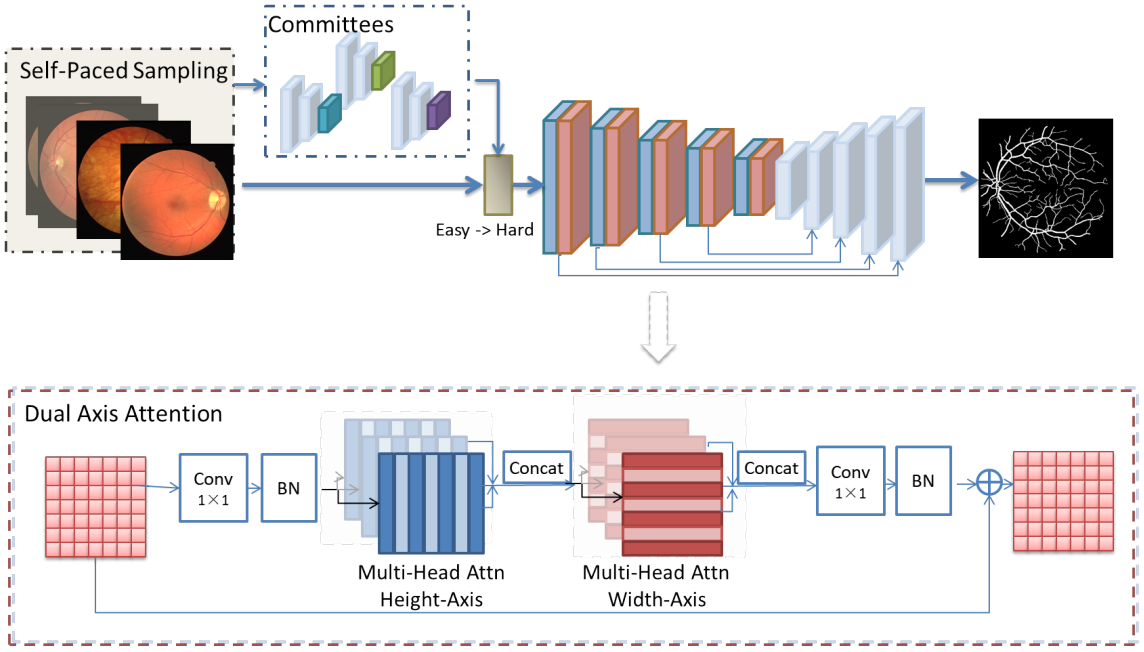

Figure 1. Scheme of the proposed self-paced dual-axis attention fusion network (SPDAA-Net).

Specifically, the committee block is comprised of a three-Conv-block encoder to extract multi-scale feature maps, and the output of the last layer of the encoder is used in the QBC algorithm for SPL by calculating the similarity and sorting and sampling the training data by a designed adaptive threshold.

The bottom part of Figure 1 shows the architecture diagram of the dual-axis attention module. A traditional axial attention layer propagates information along one particular

axis. To capture global information, dual-axial attention layers for the height axis and width axis sequentially are employed. Both of the axial attention layers adopt the multi-head attention mechanism.

3.2. The Axis Attention Module

For providing accurate and efficient segmentation, it is essential to understand which pixels belong to the mask and which belong to the background. As the background of the image is scattered, learning long-range dependencies among the background pixels can help the network avoid misclassifying a pixel as part of the mask and can reduce false positives (assuming 0 is the background and 1 is the segmentation mask). Likewise, when the segmentation mask is large, learning long-range dependencies among the mask pixels can also help make efficient predictions.

At present, transformers have been shown to be able to encode long-range dependencies. Parmar et al. [17] proposed adding positional bias while computing affinities through a self-attention mechanism, which helped make the affinities sensitive to the positional information. Relative positional encodings are often employed as this bias term. These encodings are trainable parameters that can represent the spatial features of the image. For any given input feature map x, the 2D self-attention mechanism with positional encodings can be written as [17]:

$$y = \sum_{p \in \mathcal{N}_{m \times m}(o)} \text{softmax}_p \left(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k \right) \left(v_p + r_{p-o}^v \right) \tag{2}$$

While transformer methods present strong capabilities to model the global context, their computational complexity grows quadratically, limiting their ability to scale up to high-resolution scenarios. To overcome the computational complexity of calculating the affinities, referred to as axial attention [18], self-attention is decomposed into two self-attention modules. The first module performs self-attention on the feature map's height axis, and the second one operates on the width axis. The axial attention consequently applied on the height and width axes effectively models the original self-attention mechanism with much better computational efficiency. Wang et al. [19] proposed an attention-based model for image segmentation that leverages both the axial attention mechanism and the positional encodings. Moreover, unlike previous attention models that only used relative positional encodings for queries [20], they suggested applying them to all queries, keys, and values. This additional position bias in queries, keys, and values can enhance long-range interaction with accurate positional information.

The self-attention mechanism with positional encodings along the dual-axis for the input feature map x can be expressed as:

$$y_{i-h} = \sum_{p \in \mathcal{N}_{H \times 1}} \text{softmax}_p \left(q^T k_{hj} + q^T r_{hj}^q + k^T r_{hj}^k \right) \left(v + r_{hj}^v \right) \tag{3}$$

$$y_{w-o} = \sum_{p \in \mathcal{N}_{1 \times W}} \text{softmax}_p \left(q^T k_{iw} + q^T r_{iw}^q + k^T r_{iw}^k \right) (v + r_{iw}^v) \tag{4}$$

where the formulation in Equations (3) and (4) refer to the attention model proposed in [17] and $r^q, r^k, r^v \in \mathbb{R}^{W \times W}$ for the width-wise axial attention model. Equation (3) describes the axial attention applied along the height axis of the tensor. Subsequently, Equation (4) is also used to apply axial attention along the height axis, and together they form a 1D-axis-based self-attention fusion model that is computationally efficient.

The dual-axis attention module is then adopted to the encoder of the U-Net structure. Optional striding is performed on each axis after the corresponding axial attention layer. The two 1×1 convolutions are kept to shuffle the features. The dual-axis attention module is illustrated in the bottom part of Figures 1 and 2. In the model, a 1×1 convolution is not applied between the two axial attention layers, and the stem from the original U-Net is

retained. This leads to a Conv–stem model where the first layer uses convolution and the rest of the block uses attention layers.

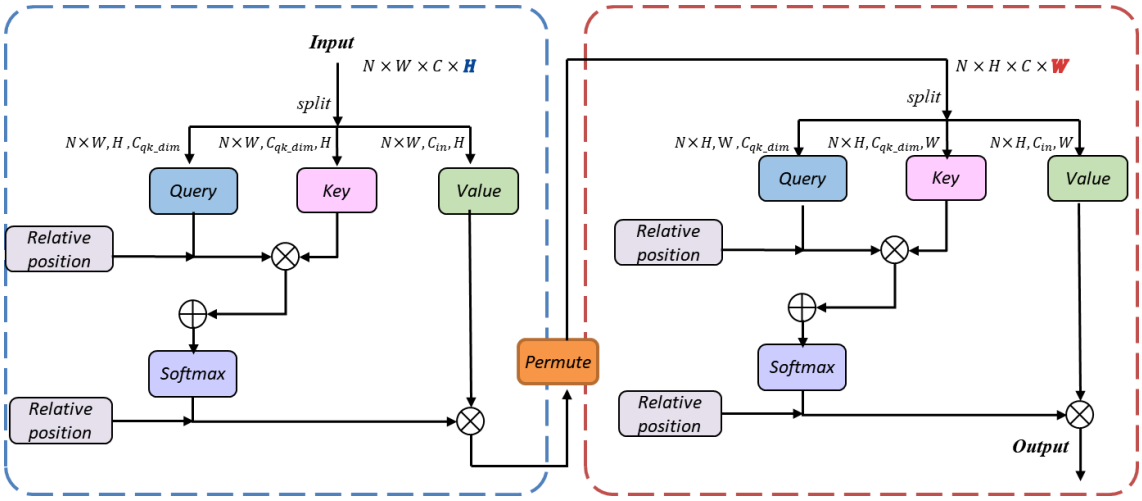

Figure 2. Scheme of the height-axis attention and width-axis attention.

3.3. Self-Paced Learning for Vessel Segmentation

Self-paced learning is inspired by the learning process of humans, who learn with easier concepts at first and then gradually involve more complex ones into training. The objective function of SPL with the self-paced regularity $g(v, \lambda)$ can be defined as:

$$\min E(w, v, \lambda) = \sum_{i=1}^{n} v_i L(y_i, f(x_i, w)) + g(v_i, \lambda) \tag{5}$$

$$g(v, \lambda) = -\lambda v, \mathrm{v} \in [0, 1] \tag{6}$$

where λ indicates the learning speed, f represents the machine learning model, x_i and y_i represent the training data and its corresponding labels, respectively, w represents the model parameters, v_i represents the weight variables that measures the difficulty of each sample, and L represents the loss function.

Learning robustness relies on sample selection to distinguish the reliable samples from the confusing ones. To train on samples organized in ascending order of learning difficulty, a query-by-committee algorithm is introduced in this paper to assign weights to the training data that reflect the easiness of the samples.

Supposing data $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ are a collection of image patches cropped from training images, $C = \{c_1, c_2, c_3\}$ is a committee of three members, and each member shares the same backbone, as shown in Figure 3. The encoder-based backbone used to construct the committee consists of three Conv blocks of different sizes. The multi-scale Conv blocks possess the ability to extract features of different receptive-field sizes, thereby enabling better feature extraction of samples. The feature map is extracted from the last layer of the encoder using the Global Average Pooling Layer (GAP) to transform the feature map into a vector; each of the three members of the committee generates a vector that represents the characteristics of the input image. To avoid algorithm failure due to having the same vector between outputs, we randomly initialize the model parameters with different seeds in each CNN. The vector generated from c_i is defined as $\vec{b}_i$; the similarity between two members is measured by the cosine distance, and the formula is as follows:

$$sim_{ij} = \frac{\vec{b}_i \bullet \vec{b}_j}{\left|\vec{b}_i\right| \times \left|\vec{b}_j\right|} \tag{7}$$

where $\left|\vec{b}_i\right|$ represents the length of the feature vector extracted by member c_i in committee C. We define the feature similarity of the kth input sample x_k between committees as:

$$sim(x_k) = \frac{1}{z} \sum_{i>j>0}^{z} sim_{ij}(x_k) \tag{8}$$

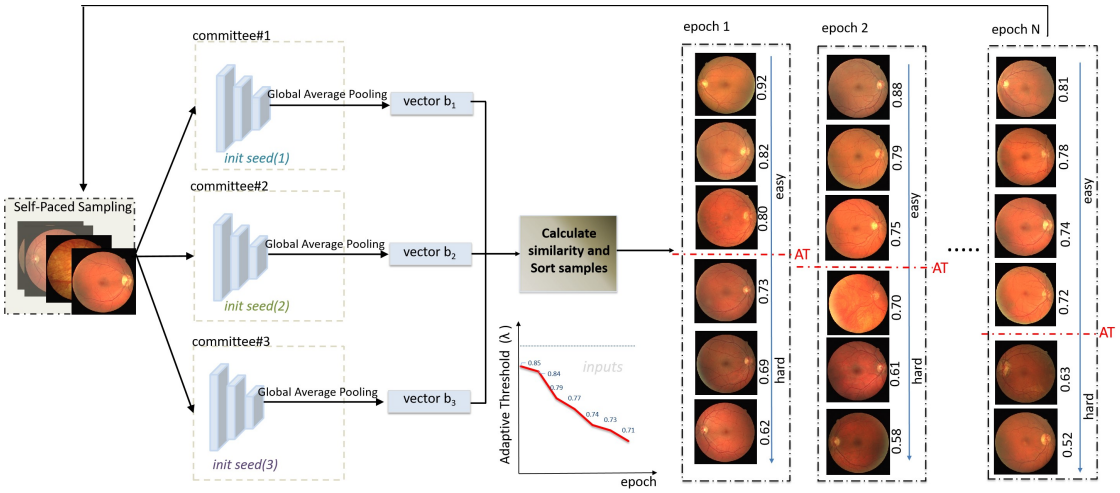

Figure 3. Scheme of proposed QBC-based self-paced learning for retinal vessel segmentation.

High similarity between vectors indicates that all z members of the committee have sufficiently learned the input data x_i (z is set to 3 in this paper). Therefore, if the vector similarity between members is high, it means that it is a simple sample. Conversely, if the similarity is low, it means that it is a hard sample that is difficult for the model to learn. After obtaining the similarity of all training samples they are normalized to the form of SPL regular terms. Then, the weights to reflect the easiness of the samples can be calculated as:

$$v_k = \frac{sim(x_k) - \min_{i\in[1,n]} sim(x_i)}{\max_{i\in[1,n]} sim(x_i) - \min_{i\in[1,n]} sim(x_i)} \tag{9}$$

To control the learning speed, a dynamic adaptive threshold λ is designed. When v_k exceeds the threshold, the sample x_k is fed into the network; λ adaptively updates every round, and its value is calculated from the maximum value of the previous round v_{k-1} according to the following formula:

$$\lambda = \alpha(1 - \frac{e}{m}) \max_{k\in[1,n]} v_k \tag{10}$$

where e represents the current training iteration and m represents the total number of training iterations; the hyperparameter α is set to 0.9 in our experiments.

For vessel segmentation, the self-paced guided loss function is designed as:

$$\min E(w, v, \lambda) = \sum_{i=1}^{n} v_i L_{BCE}(y_i, f(x_i, w)) - \lambda \sum_{i=1}^{n} v_i \tag{11}$$

4. Experiments

4.1. Datasets

We evaluated the proposed vessel segmentation framework on the DRIVE [21] and CHASE-DB1 datasets [22]. The DRIVE dataset has 40 color fundus images and ground-truth masks manually labeled by experts, from which 20 images are used for training and the remaining 20 images are used for testing. The image size of the DRIVE dataset is 584 by

565 pixels. The other common benchmark for retinal vessel segmentation is the CHASE-DB1 dataset, which contains 28 color retinal images, from which the first 20 images are used for training and the other 8 images are for testing. The image size of the CHASE-DB1 dataset is 999 by 960 pixels.

4.2. Implementation Details

During training, patches with a size of 128×128 are cropped as the input of the SPDAA-Net. All patches are randomly cropped, flipped, and zoomed in for data enhancement. We adopt the CutMix [23] data augmentation method for vessel segmentation, as shown in Figure 4. Patches are cut and pasted among training images wherein the ground-truth labels are also mixed proportionally to the area of the patches. The new combined images $\tilde{x}$ and labels $\tilde{y}$ can be defined as:

$$\begin{aligned} \tilde{x} &= \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \\ \tilde{y} &= \lambda y_A + (1 - \lambda) y_B \end{aligned} \tag{12}$$

where x_A and x_B represent two training samples, $\mathbf{M}$ denotes a binary mask indicating where to drop out and fill in from two images, $\mathbf{1}$ is a binary mask filled with ones, and $\odot$ is element-wise multiplication. In our experiments, the combination ratio is sampled from the uniform distribution $\lambda \sim \mathrm{U}(0, 1)$. Note that the major difference is that CutMix generates more locally natural images than MixUp does and also prevents the model from overfitting the training distribution and improves its generalization ability.

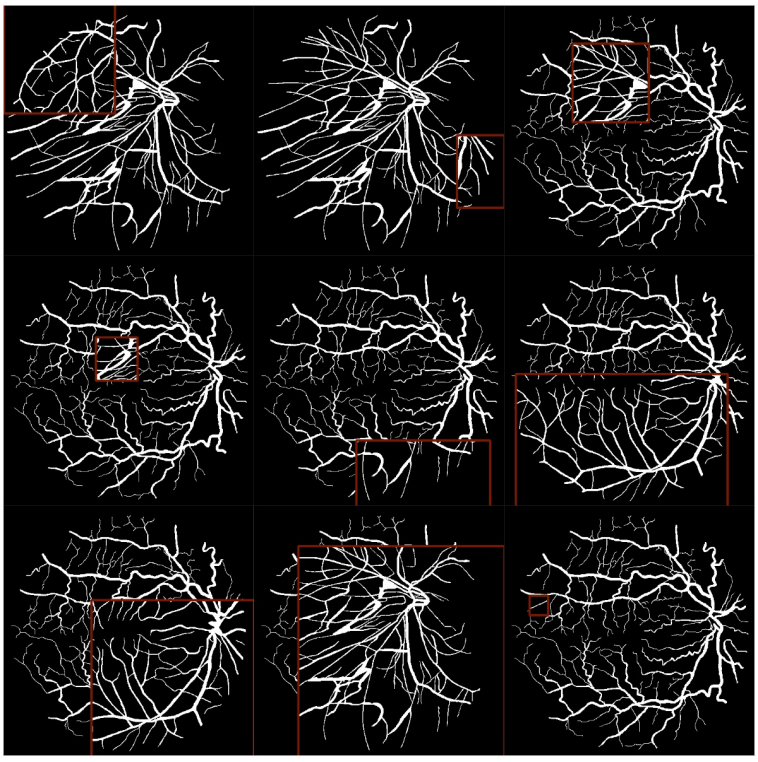

Figure 4. CutMix data augmentation for various retinal vessel segmentation. Red box indicates the combination of two different samples.

The binary mask $\mathbf{M}$ is sampled according to the bounding box coordinates $B = (r_x, r_y, r_w, r_h)$. In our experiments, we sample rectangular masks $\mathbf{M}$ whose aspect ratio is proportional to the original image. The box coordinates are uniformly sampled according to: $r_x \sim \mathrm{U}(0, W)$, $r_y \sim \mathrm{U}(0, H)$, $r_w = W\sqrt{1-\lambda}$, and $r_w = W\sqrt{1-\lambda}$. Within the cropping region, the binary mask $\mathbf{M}$ is decided by filling with 0 within the bounding box B and 1 otherwise.

In each training iteration, a CutMixed sample $(\tilde{x}, \tilde{y})$ is generated by combining two randomly selected training samples in a mini-batch according to Equation (12). The parameter settings of the self-paced learning was described in Section 3.3. We trained our network on two NVIDIA A40 GPUs with a batch size of 16 and used an Adam optimizer with a learning rate of 0.0005. The evaluation indicators are accuracy (ACC), sensitivity (SEN), and area under curve (AUC).

4.3. Experimental Results

4.3.1. Performance of Different Methods under Comparison

To verify the effectiveness of our proposed network, we conducted experiments on the Drive and CHASE-DB1 datasets and compared our model with advanced methods based on the U-Net structure and attention-based models. As shown in Table 1, our proposed network achieved an accuracy of 97.08% and an AUC of 98.93% on DRIVE, which outperforms the results obtained from the above end-to-end retinal vessel segmentation models. Our proposed network achieved an accuracy of 97.69% and an AUC of 99.13% on CHASE-DB1, which achieves good performance while reducing computational complexity.

Table 1. Comparison with advanced methods on the DRIVE and CHASE-DB1 datasets.

Dataset	Method	Year	ACC	SEN	AUC
DRIVE	Dual E-UNet [4]	2019	0.9567	0.794	0.9772
	CE-Net [6]	2019	0.9545	0.8309	0.9779
	HANet [1]	2020	0.9581	0.7991	0.9823
	SA-UNet [3]	2020	0.9698	0.8212	0.9864
	BEFD-UNet [10]	2020	0.9701	0.8215	0.9867
	CAR-UNet [2]	2021	0.9699	0.8135	0.9852
	SCL–Net [24]	2021	0.9678	0.8086	0.982
	LRCL-Net [11]	2022	0.9705	0.8441	0.989
	FR-UNet [5]	2022	0.9705	0.8556	0.9889
	Our proposed	2023	0.9708	0.8561	0.9893
CHASEDB1	Dual E-U-Net [4]	2019	0.9661	0.8074	0.9812
	HANet [1]	2020	0.9673	0.8186	0.9881
	SA-UNet [3]	2020	0.9755	0.8573	0.9905
	CAR-UNet [2]	2021	0.9751	0.8439	0.9898
	SCL-UNet [24]	2021	0.9751	0.8162	0.9879
	LRCL-Net [11]	2022	0.9771	0.8543	0.9919
	DA-Net [25]	2022	0.9766	0.8704	0.9913
	Our proposed	2023	0.9769	0.8793	0.9913

4.3.2. Ablation Studies

In order to verify the role of the self-paced module in enhancing accuracy and sensitivity in our proposed retinal vessel segmentation network, we compared the baseline model with and without SPL. Table 2 shows the results of the ablation study. The best segmentation performance is obtained after we apply all of the proposed components. The experimental result indicates that self-paced learning can significantly improve the method.

Table 2. Ablation study on DRIVE dataset.

Method	ACC	AUC	SP	DICE
Baseline w/o SPL	0.9691	0.9889	0.9826	0.8310
Baseline w/ SPL	0.9706	0.9891	0.9829	0.8315
Baseline + SPL + CutMix (our SPDAA-Net)	0.9708	0.9893	0.9830	0.8316

4.3.3. Visualization

We also give the visualization results in Figure 5 to show how these methods are visually different on probability maps and binary maps. The U-Net method can discriminate a

few of the hard pixels of detailed structures (capillaries and vessel boundaries), but breakpoints still exist on the binary segmentation results. Meanwhile, our proposed SPDAA-Net extracts more vessel pixels with insignificant feature differences, and the connectivity of the segmented vessels is greatly enhanced.

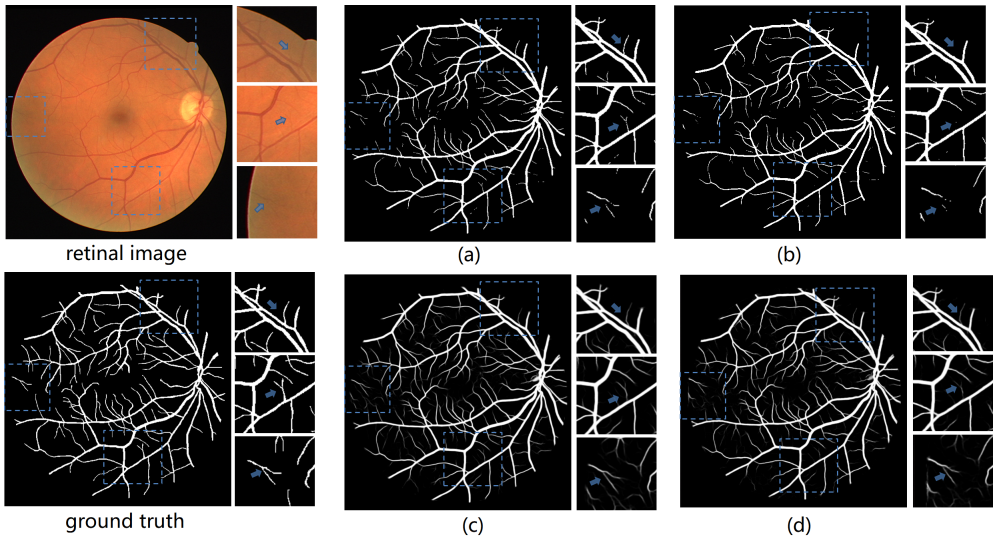

Figure 5. Visualization of segmentation results: (**a**,**b**) represent the binary maps; (**c**,**d**) represent the probability maps. From left to right: retina images and ground truths, proposed SPDAA-Net outputs, and U-Net outputs.

Figure 6 shows some visual segmentation results of our SPDAA-Net on the DRIVE and the CHASE-DB1 datasets. We can clearly see that our SPDAA-Net can obtain better segmentation visual results.

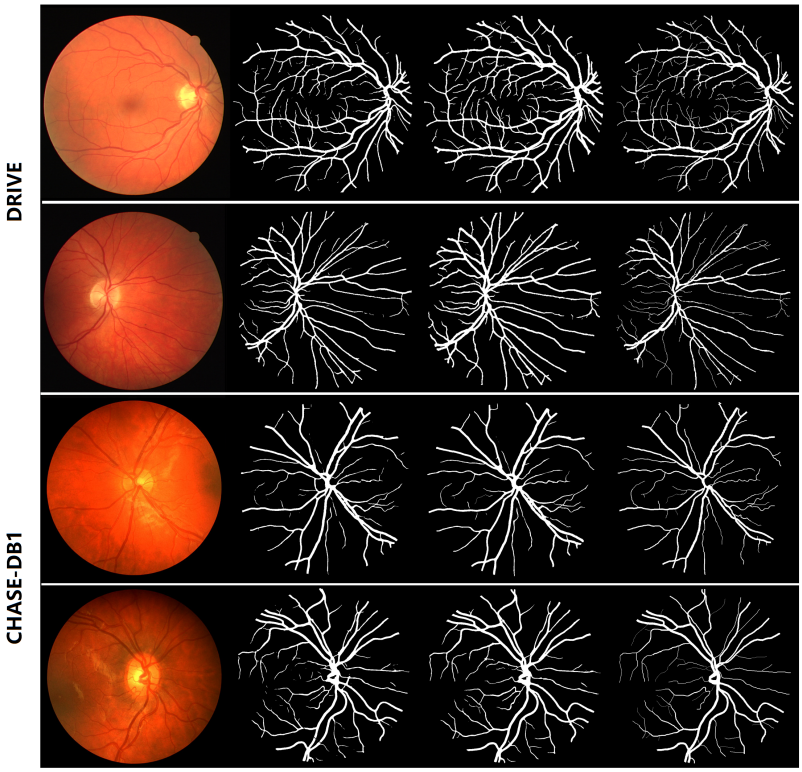

Figure 6. Visualization of segmentation results on DRIVE (**top**) and CHASE-DB1 (**bottom**) datasets. From left to right: retina images, ground truth, U-Net output, and proposed SPDAA-Net output.

5. Discussion

Deep learning has led to significant advancements in retinal blood vessel segmentation research. Despite the emergence of numerous deep learning-based methods in recent years, challenges remain in addressing the complexity of vessel segmentation and the limited availability of professionally labeled samples. In the face of these challenges, we developed a new method to enhance performance.

Due to the complexity of retinal blood vessels, network predictions may break or miss thin vessel segments. Some methods have attempted to solve this problem by increasing the field of view of the network. While some have increased the size of the convolution kernel [6,10], the receptive field remains limited. Others have added an attention module at the bottom of the network [1,12], but they still face the problem of inaccurate global context information. In this paper, we developed an axis-attention-based network bringing in long dependencies to extract global information to grasp the potential correlation and connectivity between blood vessels. Axial attention on a square image of size $N = S \times S$ performs attention on S sequences of length S: this is a total of $O(S \cdot S^2) = O(N\sqrt{N})$ computations, an $O(\sqrt{N})$ savings in computation over standard self-attention. The experimental results in Section 4.3 demonstrated that our proposed method outperforms the previous method in the retinal blood vessel segmentation task.

In addition, due to the high level of expertise required for medical annotation and the individual variations in medical images, the dataset size is much smaller compared to digital images. Developing methods to effectively utilize the available samples for network training is a pressing research challenge. In this paper, we set up feature extraction committees to calculate cosine similarity and use the similarity normalized by deviation to represent the weights of sample difficulty. Then, we designed an adaptive threshold that guides the network to take full use of available samples and learn from easy to difficult in a more intelligent manner, similar to human learning habits, by filtering the threshold during the training process. By designing the loss function based on self-paced learning and the segmentation task, the network can be trained more intelligently. Ablation experiments verified that our method based on self-learning has improved performance in retinal segmentation tasks. The combination of self-paced learning and an axis attention network enables our method to better solve the problems of vascular rupture or loss.

Although our proposed method effectively segments the complex retinal vascular structures, there is still some further work that could be researched. First, we can further improve the way the transformer interacts with the convolutions to make it more consistent with the retinal blood vessel segmentation task so as to further improve our network. Second, the calculation strategy for self-paced learning could be further improved, such as optimizing the settings of hyperparameters and the frequency of threshold updates. At present, the calculation of the self-learning module takes a relatively long time, with an average of less than two item per second in our training environment. 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. Conclusions

In this paper, a self-paced dual-axis attention fusion network for retinal vessel segmentation is proposed. In the network, the committee blocks extract features with a multi-scale receptive field and guide the model to learn samples from easy to hard based on difficultly to segment using self-paced learning. Moreover, to learn long-range dependency information and improve segmentation accuracy, a U-structure network with a width and height dual-axis attention module is applied to vessel segmentation. In addition, CutMix data augmentation is also applied to increase the diversity of the training data and to enhance model performance with various sizes of vessels and difficult samples. All of these demonstrate that our proposed method is accurate and efficient for retinal vessel segmentation. In future work, we will further improve accuracy through more efficient

attention mechanisms and will apply our proposed method to other few-shot medical imaging analyses.

Author Contributions: Conceptualization, Y.S. and X.W.; methodology, Y.S. and M.Y.; validation, W.W. and X.W.; formal analysis, M.Y.; investigation, Y.S. and X.W.; data curation, M.Y.; writing—original draft preparation, Y.S. and M.Y.; writing—review and editing, X.W.; supervision, W.W.; project administration, Y.S. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Wang, D.; Haytham, A.; Pottenburgh, J.; Saeedi, O.; Tao, Y. Hard attention net for automatic retinal vessel segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3384–3396. [CrossRef] [PubMed]
2. 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 ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 1185–1189.
3. Guo, C.; Szemenyei, M.; Yi, Y.; Wang, W.; Chen, B.; Fan, C. SA-U-Net: Spatial attention U-Net for retinal vessel segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1236–1242.
4. Wang, B.; Qiu, S.; He, H. Dual encoding U-Net for retinal vessel segmentation. *MICCAI* **2019**, *11764*, 84–92.
5. Liu, W.; Yang, H.; Tian, T.; Cao, Z.; Pan, X.; Xu, W.; Jin, Y.; Gao, F. Full-Resolution Network and Dual-Threshold Iteration for Retinal Vessel and Coronary Angiograph Segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4623–4634. [CrossRef] [PubMed]
6. 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. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI* **2015**, *9351*, 234–241.
8. He, J.; Zhu, Q.; Zhang, K.; Yu, P.; Tang, J. An evolvable adversarial network with gradient penalty for COVID-19 infection segmentation. *Appl. Soft Comput.* **2021**, *113*, 107947. [CrossRef] [PubMed]
9. Liu, X.; Zhang, D.; Yao, J.; Tang, J. Transformer and convolutional based dual branch network for retinal vessel segmentation in OCTA images. *Biomed. Signal Process. Control.* **2023**, *83*, 104604. [CrossRef]
10. Zhang, M.; Yu, F.; Zhao, J.; Zhang, L.; Li, Q. BEFD: Boundary enhancement and feature denoising for vessel segmentation. *MICCAI* **2020**, *12265*, 775–785.
11. Xu, R.; Zhao, J.; Ye, X.; Wu, P.; Wang, Z.; Li, H.; Chen, Y.W. Local-Region and Cross-Dataset Contrastive Learning for Retinal Vessel Segmentation. *MICCAI* **2022**, *13432*, 571–581.
12. 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 Inform.* **2021**, *25*, 1128–1138. [CrossRef] [PubMed]
13. Kumar, M.P.; Packer, B.; Koller, D. Self-paced learning for latent variable models. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1189–1197.
14. Jiang, L.; Meng, D.Y.; Yu, S.I.; Lan, Z.; Shan, S.; Hauptmann, A. Self-paced learning with diversity. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2078–2086.
15. Li, H.; Gong, M.; Meng, D.; Miao, Q. Multi-objective self-paced learning. *Proc. AAAI Conf. Artif. Intell.* **2016**, *30*, 1802–1808. [CrossRef]
16. Wang, H.; Jin, Y.; Doherty, J. Committee-Based Active Learning for Surrogate-Assisted Particle Swarm Optimization of Expensive Problems. *IEEE Trans. Cybern.* **2017**, *47*, 2664–2677. [CrossRef] [PubMed]
17. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 68–80.
18. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180.
19. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.C. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. *arXiv* **2020**, arXiv:2003.07853.
20. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 464–468.
21. Staal, J.; Abramoff, M.D.; Niemeijer, M.; Viergever, M.A.; van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **2004**, *23*, 501–509. [CrossRef] [PubMed]
22. Fraz, M.M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, A.R.; Owen, C.G.; Barman, S.A. An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 2538–2548. [CrossRef] [PubMed]

23. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
24. Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; Van Gool, L. Exploring cross image pixel contrast for semantic segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 7303–7313.
25. Wang, C.; Xu, R.; Xu, S.; Meng, W.; Zhang, X. DA-Net: Dual Branch Transformer and Adaptive Strip Upsampling for Retinal Vessels Segmentation. *MICCAI* **2022**, *13432*, 528–538.