mathematics

MDPI

Article

Dual Attention Multiscale Network for Vessel Segmentation in Fundus Photography

Pengshuai Yin [1,2,†], Yupeng Fang [3,4,†] and Qilin Wan [2,*]

[1] School of Future Technology, South China University of Technology, Guangzhou 510641, China
[2] Guangdong-Hong Kong-Macao Greater Bay Area Weather Research Center for Monitoring Warning and Forecasting, Shenzhen 518000, China
[3] School of Software Engineering, South China University of Technology, Guangzhou 510006, China
[4] Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, Guangzhou 510006, China
* Correspondence: wanqilin@gbamwf.com
† These authors contributed equally to this work.

Abstract: Automatic vessel structure segmentation is essential for an automatic disease diagnosis system. The task is challenging due to vessels' different shapes and sizes across populations. This paper proposes a multiscale network with dual attention to segment various retinal blood vessels. The network injects a spatial attention module and channel attention module on a feature map, whose size is one-eighth of the input size. The network also uses multiscale input to receive multi-level information, and the network uses the multiscale output to gain more supervision. The proposed method is tested on two publicly available datasets: DRIVE and CHASEDB1. The accuracy, AUC, sensitivity, and specificity on the DRIVE dataset are 0.9615, 0.9866, 0.7709, and 0.9847, respectively. On the CHASEDB1 dataset, the metrics are 0.9800, 0.9892, 0.8215, and 0.9877, respectively. The ablative study further shows effectiveness for each part of the network. Multiscale and dual attention mechanism both improve performance. The proposed architecture is simple and effective. The inference time is 12 ms on a GPU and has potential for real-world applications. The code will be made publicly available.

Keywords: vessel segmentation; Mmedical image analysis; deep learning

MSC: 68T01

Citation: Yin, P.; Fang, Y.; Wan, Q. Dual Attention Multiscale Network for Vessel Segmentation in Fundus Photography. *Mathematics* **2022**, *10*, 3687. https://doi.org/10.3390/math10193687

Academic Editor: Catalin Stoean

Received: 22 August 2022
Accepted: 30 September 2022
Published: 8 October 2022

1. Introduction

The segmentation of vasculature in retinal images is important in aiding the management of many diseases, such as diabetic retinopathy (DR) and hypertensive retinopathy (HR). DR is caused by high blood sugar levels and results in the swelling of the retinal vessels [1]. HR is caused by high blood pressure and results in the narrowing of vessels or increased vascular tortuosity [2]. The early diagnosis of pathological diseases often helps patients receive timely treatment. However, manually labeling vessel structures is time-consuming, tedious, and subject to human error. Automated segmentation of retinal vessels is in high demand and can release the intense burden of skilled staff.

Automatic retinal vessel segmentation faces many challenges. The retinal blood vessel structure is extremely complicated with high tortuosity and various shapes, such as angles, branching patterns, length, and width [3]. The high anatomical variability and varying vessel scales across populations increase the difficulty in segmentation. Furthermore, the noise and poor contrast accompanied by the low resolutions limit the segmentation performance. Traditional vessel segmentation methods often cannot robustly segment all vessels of interest.

Deep learning methods show impressive performance on image segmentation. The most widely used architecture is U-Net [4]. The coarse-to-fine feature representation learned

by U-Net is suitable to gain satisfactory performance on a small dataset. The attention U-Net can further improve the performance [5]. The attention module automatically learns to focus on interest vessels with varying shapes while preserving computational efficiency. DA-Net [6] proposes a spatial attention module and channel attention module for natural scene parsing. The spatial attention module and the channel attention module utilize a self-attention mechanism to capture feature dependencies in the spatial and channel dimensions, respectively. The spatial attention module aggregates features at all positions with weighted summation, and the channel attention module captures the channel dependencies between any two channel maps. This paper designs a dual attention multiscale network for vessel segmentation. The network collects different scale information from multiscale inputs and gains extra supervision from additional outputs. Furthermore, the proposed network can distinguish vessels from the background and select the feature map's most informative region and channels through a dual attention mechanism. In summary, the contributions of the paper are listed as follows:

- The paper proposes a simple and effective multiscale fully convolutional network for vessel segmentation including additional multiscale input paths and multiscale output paths. The multiscale input helps to detect vessels with different shapes and sizes, and the multiscale output provides the network with more supervision.
- The paper proposes a dual attention module on the multiscale architecture. The dual attention module learns the relationship between positions and channels. The dual attention module improves the discriminative power of the feature map.
- The paper conducts extensive experiments to verify the effectiveness of the proposed network. Experiments show that the multiscale architecture with a dual attention module is suitable for vessel segmentation.

2. Literature Review

In this section, we introduce the most commonly used fully convolutional neural network for retinal vessel segmentation. Then, we introduce the popular attention mechanisms and multiscale networks for vessel segmentation. Motivated by the existing successful network architecture, this paper proposes a fully convolutional neural network based on UNet [4]. The proposed network further incorporates multiscale and attention mechanisms to boost the discriminative power of the feature map.

2.1. Fully Convolutional Neural Network

The most widely used architecture for vessel segmentation is UNet [4]. The encoder–decoder structure of UNet combines low-level local features with high-level global features to produce high-resolution prediction. Many researchers build retinal segmentation networks based on UNet. For example, Jin et al. [7] improve UNet by injecting a deformable convolution block. The deformable convolution block adaptively adjusts the receptive fields to capture vessels with variance shape and scale. Wang et al. [8] reduce information loss caused by consecutive downsampling layers by introducing a feature refinement path. The feature refinement path improves the detail vessel information and boosts the discriminative power of the feature map. Our network is also built upon UNet to combine local and global contexts.

2.2. Attention Network

Attention mechanism is very successful and has already been adopted for vessel segmentation. Attention U-Net [5] captures a sufficiently large receptive field to collect semantic contextual information and integrates attention gates to reduce false-positive predictions for small objects that show large shape variability. Ni et al. [9] propose a global channel attention module for vessel segmentation that emphasizes the inter-relationship of the feature. CS-Net [10] integrates channel attention and spatial attention into U-Net for 2D and 3D vessel segmentation. Hao et al. [11] exploit contextual frames of sequential images in a sliding window centered at the current frame and equipped with a channel attention

mechanism in the decoder stage. Li et al. [12] propose an attention gate to highlight salient features that are passed through the skip connections. HANet [13] automatically focuses the network's attention on regions that are "hard" to segment. The vessel regions which are "hard" or "easy" are based on the coarse segmentation probabilistic map. The attention mechanism can let the network focus on the most informative region and discard irrelevant information. Our network also incorporates the attention mechanism to boost the discriminative power of the feature map.

2.3. Multiscale Network

Multiscale architecture can effectively detect vessels with various shapes and sizes. Yue et al. [14] utilize different scale image patches as inputs to learn richer multiscale information. Roberto et al. [15] propose a multiple-scale Hessian approach to enhance the vessels followed by thresholding. Wu et al. [16] generate multiscale feature maps by max-pooling layers and up-sampling layers. The first multiscale network converts an image patch into a probabilistic retinal vessel map, and the following multiscale network further refines the map. Yin et al. [17] proposed to utilize multiscale input to fuse multi-level information. Our network injects multiscale information by introducing additional multiscale input paths and multiscale output paths.

3. Materials and Methods

The architecture of the proposed method is shown in Figure 1. The network structure consists of multiscale input, dual attention module and multiscale output. The dual attention module contains the spatial positional attention module (PAM), and channel attention module (CAM).

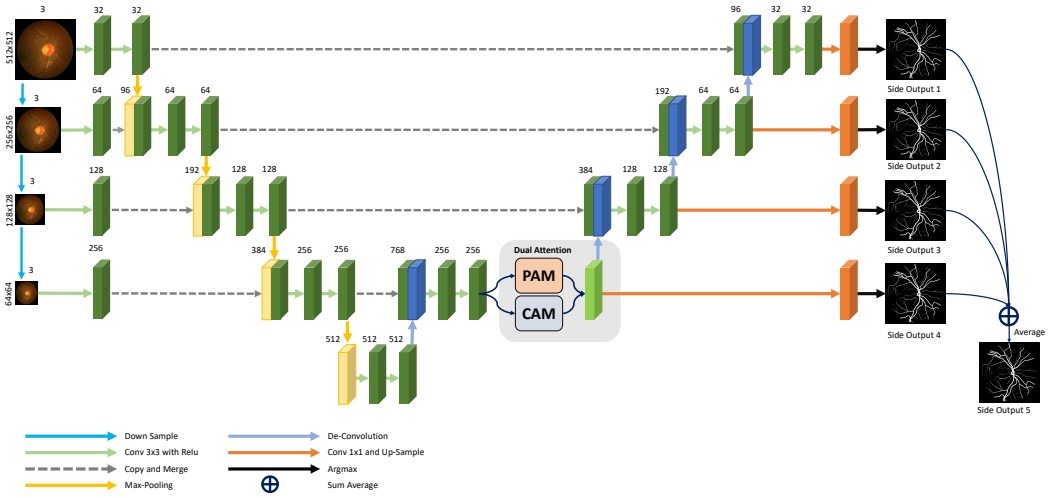

Figure 1. The proposed vessel segmentation framework.

3.1. Multiscale Input

Multiscale information helps the network to discriminate vessels with different sizes. The input image is downscaled to $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ of the original size. Each image is sent to the corresponding encoder path. The multiscale input can let the network gather different levels of information and improve the ability to detect multiscale vessels.

3.2. U-Shape Architecture

Our network is constructed based on U-Net, and the input of the encoder path is an image pyramid. Two 3×3 convolution layers are applied to the input image for each encoder path, followed by a 2×2 max-pooling operation with the element-wise rectified linear unit (ReLU) activation function to generate encoder feature maps. The top encoder feature map is then down-sampled and connected to the feature map of the bottom

encoder path with a smaller scale input image. The channel number of the feature maps is also doubled after down-sampling, enabling the architecture to learn complex structures efficiently. Like the encoder path, the decoder path produces a decoder feature map using two 3×3 convolution layers. The input of the decoder path is the combination of the up-sampled feature from the bottom decoder path and the feature map of the corresponding encoder path using skip connections. The channel number of the up-sampled feature map through a 2×2 up-sampling layer is also halved to preserve symmetry. Finally, the high-dimensional feature representation of the output of the last decoder layer is fed to the dual attention module to learn the relationship between position and channel. The feature representation with more discriminative power is sent to the multiscale output layer for final prediction.

3.3. Dual Attention Mechanism

The dual attention mechanism contains a spatial attention module and a channel attention module as shown in Figure 2.

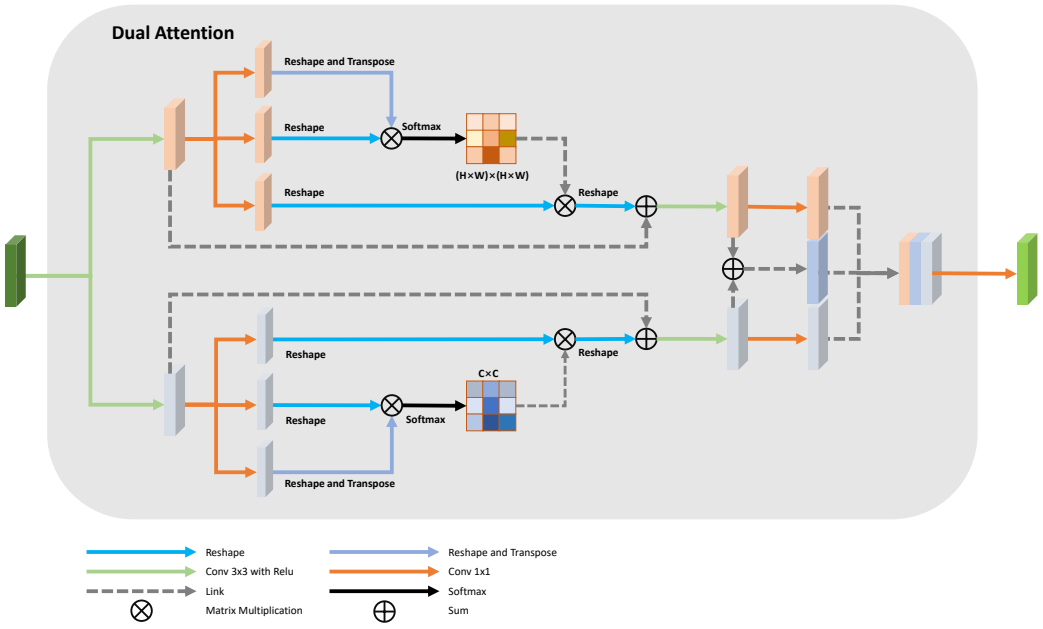

Figure 2. The proposed dual attention module.

Spatial Attention Module: The spatial attention module models rich contextual dependencies over feature maps by learning a spatial attention matrix, which represents the spatial relationships between the features of any two pixels. Different from DA-Net [6], we put the attention module in the branch with the feature map size equal to $\frac{1}{8}$ of the original image rather than directly resampling the attention map as the output. The design retains global context information without adding many parameters. Furthermore, the vessel segmentation requires skip-connection operations to fuse low-level information and recover the spatial information loss caused by down-sampling operations. The spatial attention module encourages the network to focus on vessel structure information to prevent spatial information loss. The input feature representation $S \in \mathbb{R}^{C \times H \times W}$ is fed into three convolution layers to generate three feature maps A, B, and C, where $A, B, C \in \mathbb{R}^{C \times H \times W}$. The three feature maps are reshaped into $C \times N$, where $N = H \times W$ is the total pixel number. After that, the transpose of A and B is multiplied and followed by a softmax layer to form the spatial attention $SA \in \mathbb{R}^{N \times N}$:

$$SA_{ij} = \frac{exp(A_i \cdot B_j)}{\sum_{i=1}^{N} exp(A_i \cdot B_j)}. \quad (1)$$

The spatial attention module SA represents the impact of position i on position j. A similar feature representation will have a greater correlation. The transposition of SA is multiplied by C to form a feature representation and reshaped to $\mathbb{R}^{C\times H\times W}$. The result is multiplied by a scale parameter α and followed by an element-wise summation with input feature S to generate the final spatial attention feature map SAO:

$$SAO_i = \alpha \sum_{j=1}^{N}(SA_{ij}C_j) + S_i. \tag{2}$$

Here, α is a learnable parameter, and it is initialized to 0. The spatial attention module calculates the weighted sum across all positions of the feature map. The relationship between vessel pixels at different locations will be fully learned. The similar vessel pixels promote each other, and the spatial attention module improves the semantic consistency. SAO_i is a linear combination of SA_{ij}, C_j, and S_i. The deviation of SA_{ij} can be written as:

$$\frac{\partial SA_{ij}}{\partial exp(A_i \cdot B_j)} = \frac{(exp(A_i \cdot B_j)')(\sum_{i=1}^{N} exp(A_i \cdot B_j)) - (\sum_{i=1}^{N} exp(A_i \cdot B_j))'(exp(A_i \cdot B_j))}{(\sum_{i=1}^{N} exp(A_i \cdot B_j))^2}. \tag{3}$$

Let $e^{m_{ij}} = exp(A_i \cdot B_j)$ represent the impact of position i of A on the position j of B. The deviation can be written as:

$$\frac{\partial SA_{ij}}{\partial e^{m_{ij}}} = \frac{(e^{m_{ij}})'(\sum_{i=1}^{N} e^{m_{ij}}) - (\sum_{i=1}^{N} e^{m_{ij}})'(e^{m_{ij}})}{(\sum_{i=1}^{N} e^{m_{ij}})^2}. \tag{4}$$

Since

$$(\sum_{i=1}^{N} e^{m_{ij}})' = \frac{\partial}{\partial m_{ij}}(\sum_{i=1}^{N} e^{m_{ij}}) = e^{m_{ij}}. \tag{5}$$

Then

$$\frac{\partial SA_{ij}}{\partial e^{m_{ij}}} = \frac{(e^{m_{ij}})(\sum_{i=1}^{N} e^{m_{ij}}) - (e^{m_{ij}}e^{m_{ij}})}{(\sum_{i=1}^{N} e^{m_{ij}})^2} = \frac{e^{m_{ij}}}{\sum_{i=1}^{N} e^{m_{ij}}} - (\frac{e^{m_{ij}}}{\sum_{i=1}^{N} e^{m_{ij}}})^2 = e^{m_{ij}}(1 - e^{m_{ij}}). \tag{6}$$

Equation (1) is derivable, and the derivation can be written as:

$$\frac{\partial SA_{ij}}{\partial exp(A_i \cdot B_j)} = exp(A_i \cdot B_j)(1 - exp(A_i \cdot B_j)). \tag{7}$$

The network with the spatial attention module can be trained end-to-end.

Channel Attention Module: The relationship between different feature map channels of high-level features can be learned by the channel attention module. The long-range contextual information in the channel dimension helps to improve the vessel segmentation performance since different vessel responses are associated with each other. The original feature representation $S \in \mathbb{R}^{C\times H\times W}$ is reshaped to $A' \in \mathbb{R}^{C\times N}$, where $N = H \times W$ is the total pixel number. A' and the transpose of A' is multiplied and followed by a softmax layer to form the channel-wise attention map:

$$CA_{ij} = \frac{exp(A'_i \cdot A'_j)}{\sum_{i=1}^{C} exp(A'_i \cdot A'_j)}. \tag{8}$$

The channel attention map CA calculates the impact of channel i on channel j. The transpose of CA and the input feature map S are multiplied, and the result is reshaped to

$\mathbb{R}^{C\times H\times W}$. The reshaped result is scaled by the parameter β and element-wise sum with S to form the final channel attention $CAO \in \mathbb{R}^{C\times H\times W}$:

$$CAO_i = \beta \sum_{j=1}^{C} (CA_{ij} A'_j) + A'_i. \tag{9}$$

Similar to the spatial attention module, the deviation of CA_{ij} can be written as:

$$\frac{\partial CA_{ij}}{\partial exp(A'_i \cdot A'_j)} = exp(A'_i \cdot A'_j)(1 - exp(A'_i \cdot A'_j)). \tag{10}$$

Different from other attention modules, we concatenate the spatial attention map, the channel attention map, and the summation of the two maps together to form a more capable feature representation.

3.4. Multiscale Output

Multiscale outputs provide more supervision in network training. There are M side-output layers in the network, and each side-output layer can be considered as a classifier to generate a matching local output map for the earlier layers. Here are the loss functions of all the side-output layers:

$$L_{side\text{-}output} = \frac{1}{M} \sum_{m=1}^{M} L_{cross\text{-}entropy}(y, y'), \tag{11}$$

$L_{cross\text{-}entropy}$ is the cross entropy loss for each side-output layer:

$$L_{cross\text{-}entropy} = -\sum_{i} (y'_i \log(y_i)), \tag{12}$$

y_i is the predicted probability value for class i, and y'_i is the true probability for that class.

We compute 4 side-output maps and an average layer to combine them all while the final optimization function is the sum of these 5 side-output losses. The side-output layer alleviates the gradient vanishing problem by back-propagating the side-output loss to the early layer in the decoder path, which is helpful for the training of the early layer. We use multiscale fusion because it has been proven to achieve high performance. The side-output layer also adds more supervision for each scale to improve the performance. The final layer which is considered a classifier treats the vessel segmentation as a pixel-wise classification to produce the probability map of each pixel.

4. Results

4.1. Data Preperation

We conduct experiments on two datasets: DRIVE and CHASEDB1.

DRIVE: The Digital Retinal Images for Vessel Extraction (DRIVE) [18] is a dataset for retinal vessel segmentation, which consists of 40 color fundus images of size 768 × 584 pixels, including 7 abnormal cases. It was equally divided into 20 images for training and 20 images for testing along with 2 manual segmentations of the vessels. The first segmentation was accepted as the ground truth for performance evaluation, while the second segmentation was accepted as a human observer reference for performance comparison. The images were captured in digital form from a Canon CR5 non-mydriatic 3CCD camera at 45° field of view (FOV).

CHASEDB1: The CHASEDB1 dataset [19] for retinal vessel segmentation which consists of 28 color retina images of size 960 × 999 pixels were collected from both left and right eyes of 14 school children. These images were captured by a handheld Nidek NM-200-D fundus camera at 30° field of view, and each image was annotated by two independent

human experts. We selected the first 20 images for training and the remaining 8 images for testing [20].

4.2. Evaluation Metrics

The vessel segmentation process is a pixel-based classification, with each pixel being classified as a vessel or surrounding tissue. We employed four indicators: Spe (Specificity), Sen (Sensitivity), Acc (Accuracy), and AUC (Area Under ROC) to measure model performance. Acc was measured by the ratio of the total number of correctly classified pixels (the sum of true positives and true negatives) to the number of pixels in the image FOV (field of view). Sen represents the ability to correctly detect real vessel pixels. Spe is the ability to detect non-vessel pixels. Acc, Sen, and Spe can be denoted as:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}, \tag{13}$$

$$Sen = \frac{TP}{TP + FN}, \tag{14}$$

$$Spe = \frac{TN}{TN + FP}, \tag{15}$$

here TP (true positive) is where a pixel is identified as the vessel in both the segmented image and ground truth; TN (true negative) is where a non-vessel pixel of the ground truth is correctly classified in the segmented image. FP (false positive) is the false positive where the non-vessels are incorrectly predicted as vessels. FN (false negative) is the false negative that the model wrongly predicts the negative class.

4.3. Implementation Details

We set the learning rate at 0.001 decayed by a factor of 10 every 50 epochs. The network was trained for 300 epochs from scratch on an NVIDIA GeForce RTX 3090 Ti GPU. The input images of the neural network were resized to 512×512. In order to improve the generalization ability of the network, we also used several data enhancement techniques, including random horizontal flip with a probability of 0.5, random rotation in $[-20°, 20°]$, and gamma contrast enhancement in $[0.5, 2]$.

4.4. Performance Evaluation

In this section, we compare our method with other state-of-the-art methods on DRIVE and CHASEDB1 datasets. The methods include U-Net [4], Zhang et al. [21], Liskowski et al. [22], DRIU [23], Yan et al. [24], CE-Net [25], LadderNet [26], DU-Net [27], Bo Liu et al. [28], VesselNet [29], Yue et al. [14], DA-Net [6] and Yin et al. [17]. Table 1 shows the performance on the DRIVE dataset. Figure 3 shows the prediction of the proposed method on the DRIVE dataset. DRIU [23] extracts side feature maps and designs specialized layers to perform blood vessel segmentation. DRIU does not take advantage of multiscale information. Liskowski et al. [22] design a convolutional neural network that contains three convolutional layers, one pooling layer, and two fully connected layers. Liskowski et al. train the network on image patches, and the improvement is mainly due to the elaborately designed image pre-processing method such as global contrast normalization and zero-phase whitening. Our network improves the result mainly due to the effective network architecture. Yan et al. [24] improve the performance by jointly adopting both the segment-level and the pixel-wise losses. LadderNet [26] has multiple pairs of encoder–decoder branches and can be viewed as a chain of multiple U-Nets. Our method further adds an attention mechanism to discard irrelevant information. DU-Net [27] contains two encoders: a spatial path with a large kernel to preserve the spatial information and a context path with a multiscale convolution block to capture more semantic information. Our method incorporates multiscale information by sending multiscale input to each encoder and adopting an attention mechanism to capture important information. VesselNet [29]

proposes a lightweight deep learning model by injecting the inception residual convolutional block inside a U-like encoder–decoder architecture for vessel segmentation. The method adopts multi-path supervision just like our network; however, VesselNet lacks multiscale information, and our network further utilizes an attention mechanism to highlight important regions. Sine-Net [30] applies up-sampling and then down-sampling to catch thin and thick vessel features. Guo et al. [31] propose a channel attention double residual block to enhance the discriminative ability of the network by considering the interdependence between feature maps. Guo et al. learn channel maps by 1D convolutions. Our network performs self-attention to learn channel maps. Gao et al. [32] utilize shuffle attention [33] multiple times to explore the feature dependencies in both spatial and channel dimensions. The authors also adopt ECA-Net [34] to reduce the model complexity while maintaining performance. Our method also utilizes spatial and channel attention to explore inter-relationship between locations and channels. The attention map of our method is calculated from the global image ,and Gao et al. extract attention from the local patch. Our method further adopts multiscale input to encode multiscale information and multiple side-outputs to receive more supervision. Jiang et al. [35] propose using conditional deep convolutional generative adversarial networks to segment the retinal vessels. Jiang et al. introduce residual modules to the generator for better representation learning ability. Li et al. [36] propose an attention module built on U-Net to capture global information and to enhance features by placing it in the process of feature fusion. The attention module proposed by Li et al. only considers the spatial locations; our network takes both channel and spatial information into consideration.

Table 1. Segmentation performance for DRIVE inside FOV.

Method	Year	Acc	AUC	Sen	Spe
Human Observer	-	**0.9578**	**N.A**	**0.8288**	**0.9701**
U-Net [4]	2015	0.9531	0.9755	0.7537	0.9820
Zhang et al. [21]	2016	0.9476	0.9636	0.7743	0.9725
Liskowski et al. [22]	2016	0.9542	0.9752	0.7653	0.9818
DRIU [23]	2016	0.9541	0.9801	0.8280	0.9728
Yan et al. [24]	2018	0.9542	0.9752	0.7653	0.9818
Wu et al. [16]	2018	0.9578	0.9821	0.8038	0.9802
CE-Net [25]	2019	0.9545	0.9779	0.8309	-
LadderNet [26]	2019	0.9561	0.9793	0.7856	0.9810
DU-Net [27]	2019	0.9567	0.9772	0.7940	0.9816
Bo Liu et al. [28]	2019	0.9559	0.9779	0.8072	0.9780
VesselNet [29]	2019	0.9578	0.9821	0.8038	0.9802
Yue et al. [14]	2019	0.9561	0.9796	0.8199	0.9762
DA-Net [6]	2019	0.9615	0.9808	0.8075	0.9841
Yin et al. [17]	2020	0.9604	0.9846	0.7614	0.9837
Jiang et al [35]	2021	0.9795	-	0.8258	0.9896
Li et al. [36]	2021	0.9568	0.9806	0.7921	0.9810
Sine-Net [30]	2021	0.9689	0.9851	0.7987	0.9854
Guo et al. [31]	2021	0.9699	0.9852	0.8135	0.9849
Gao et al. [32]	2022	0.9795	-	0.8258	0.9896
Ours	-	0.9615	0.9866	0.7709	0.9847

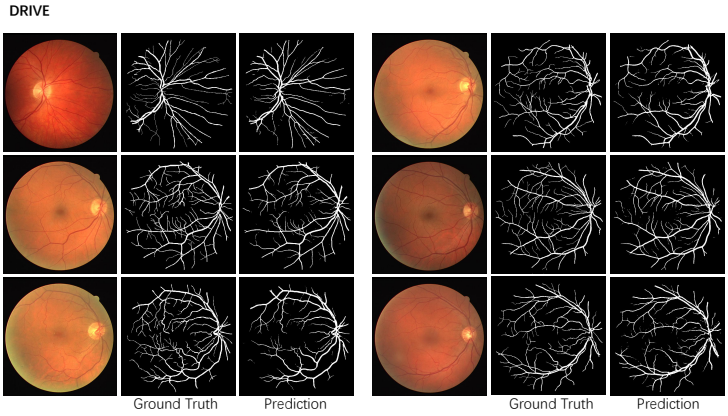

Figure 3. The segmentation example of the proposed method on the DRIVE dataset.

Our proposed method achieves the highest AUC compared to other methods. The proposed attention module gathers global context information from the feature map with one-eighth of the original image size, while the input image size of the proposed network is 512 × 512. DA-Net directly up-samples the attention map as the output and does not consider adopting skip-connections to recover the spatial information loss caused by down-sampling layers. The attention module used in our method is also different from DA-Net. We concatenate the spatial attention map, the channel attention map, and the sum of these two maps together to form a more discriminate feature representation. Yue et al. also aggregate multiscale context information. Our method not only takes advantage of multiscale context, but also provides the network with multiscale supervision through multiple side-outputs. Our method performs much better than Yue et al.

Table 2 shows the performance evaluation on the CHASEDB1 dataset. Figure 4 shows the corresponding prediction. Our method surpasses all the other methods. Yin et al. [17] provide the network with more edge information through a guided filter module. Our multiscale network adopts dual attention to aggregate the relationship between pixels and channels. Our network is built on U-Net and can detect vessels in various shapes and sizes. Our method improves Yin et al. for all the metrics.

Table 2. Segmentation performance of CHASEDB1 inside FOV.

Method	Year	Acc	AUC	Sen	Spe
Human Observer	-	**0.9545**	**N.A**	**0.8105**	**0.9711**
U-Net [4]	2015	0.9578	0.9772	0.8288	0.9701
DRIU [23]	2016	0.9657	0.9746	0.7651	0.9822
Liskowski et al. [22]	2016	0.9535	0.9823	0.7816	0.9836
Yan et al. [24]	2018	0.9610	0.9781	0.7633	0.9809
LadderNet [26]	2019	0.9656	0.9839	0.7978	0.9818
DU-Net [27]	2019	0.9661	0.9812	0.8074	0.9821
VesselNet [29]	2019	0.9661	0.9860	0.8132	0.9814
Yin et al. [17]	2020	0.9783	0.9869	0.7993	0.9868
Li et al. [36]	2021	0.9635	0.9819	0.7818	0.9819
Sine-Net [30]	2021	0.9678	0.9833	0.8011	0.9815
Ours	-	0.9800	0.9892	0.8215	0.9877

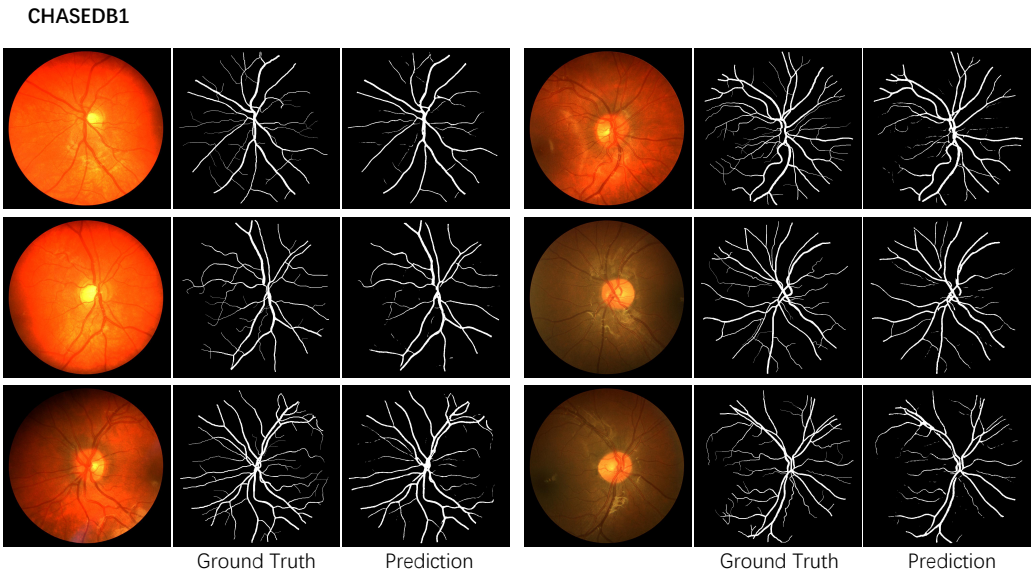

Figure 4. The segmentation example of the proposed method on the CHASEDB1 dataset.

4.5. Ablative Studies

This section evaluates the performance of each part of the network. Tables 3 and 4 show the performance of the proposed network using different modules on the DRIVE dataset and the CHASEDB1 dataset, respectively. The evaluation metrics include the mean IOU (mean intersection over union) commonly used in semantic segmentation. The multiscale architecture significantly improves the performance compared to our UNet backbone, and the attention mechanism further improves the performance. All the results are obtained on the testing set for fair comparison to other methods.

Table 3. The performance for each part of the DRIVE dataset inside FOV.

Method	Acc	AUC	Sen	Spe	MIOU
BackBone	0.9469 ± 0.0059	0.9604 ± 0.0083	0.6289 ± 0.1006	0.9823 ± 0.0082	0.5553 ± 0.0603
Multiscale Network	0.9614 ± 0.0044	0.9863 ± 0.0036	0.7668 ± 0.0528	0.9847 ± 0.0039	0.6790 ± 0.0310
Multiscale Network+Attention Module	0.9615 ± 0.0043	0.9866 ± 0.0034	0.7709 ± 0.0521	0.9847 ± 0.0039	0.6807 ± 0.0307

Table 4. The performance for each part of the CHASEDB1 dataset inside FOV.

Method	Acc	AUC	Sen	Spe	MIOU
BackBone	0.9693 ± 0.0055	0.9472 ± 0.0159	0.4892 ± 0.0066	0.9823 ± 0.0035	0.4230 ± 0.0480
Multiscale Network	0.9798 ± 0.0046	0.9880 ± 0.0031	0.8208 ± 0.0422	0.9875 ± 0.0044	0.6506 ± 0.0673
Multiscale Network+Attention Module	0.9800 ± 0.0043	0.9892 ± 0.0024	0.8215 ± 0.0381	0.9877 ± 0.0041	0.6548 ± 0.0661

For the DRIVE dataset, we first perform Shapiro–Wilk test to verify distribution normality. For our UNet backbone, the p for ACC, AUC, SPE, SEN, and MIOU are 0.56, 0.19, 0.36, 0.92, and 0.23, respectively. For the proposed network, the p for ACC, AUC, Spe, Sen, and MIOU are 0.98, 0.16, 0.67, 0.41, and 0.48, respectively. The result of the multiscale structure also satisfies the distribution normality. The p for ACC, AUC, Spe, Sen, and MIOU are 0.77, 0.15, 0.72, 0.74, and 0.78, respectively. We also conduct a paired sample T-Test and calculate Cohen's d to evaluate the significance of the performance improvement. The entire network significantly improves our UNet backbone in terms of Acc, AUC, Sen, Spe, and MIOU by 0.0146 ($p < 0.01$), 0.0262 ($p < 0.01$), 0.1420 ($p < 0.01$), 0.0024 ($p < 0.01$), and

0.0031 ($p < 0.01$), respectively. The effect size for Acc, AUC, Sen, Spe, and MIOU are 2.69, 3.99, 1.72, 0.4, and 2.55 respectively. The multiscale structure improves the performance of our UNet backbone in terms of Acc, AUC, Sen, Spe, and MIOU by 0.0145 ($p < 0.01$), 0.0259 ($p < 0.01$), 0.1379 ($p < 0.01$), 0.0024 ($p < 0.01$), and 0.1237 ($p < 0.01$), respectively. The corresponding effect size for Acc, AUC, Sen, Spe, and MIOU are 2.67, 3.90, 1.71, 0.32, and 2.52 respectively. The attention module further improves the performance compared to the multiscale architecture in terms AUC, Sen, and MIOU by 0.0003 ($p < 0.05$), 0.0041 ($p < 0.05$) and 0.0017 ($p < 0.05$) with effect size 0.3, 1.3, and 0.8, respectively. The attention module slightly improves ACC and SPE compared to the multiscale architecture.

For the CHASEDB1 dataset, we also perform Shapiro–Wilk test to verify distribution normality. For our UNet backbone, the p for ACC, AUC, Spe, Sen, and MIOU are 0.18, 0.72, 0.75, 0.40, and 0.70, respectively. For the proposed network, the p for ACC, AUC, Spe, Sen, and MIOU are 0.36, 0.70, 0.31, 0.71, and 0.43, respectively. The result of the multiscale structure also satisfies the distribution normality. The p for ACC, AUC, Spe, Sen, and MIOU are 0.84, 0.64, 0.76, 0.84, and 0.49, respectively. We also conduct a paired sample T-Test and calculate Cohen's d to evaluate the significance of the performance improvement. The entire network significantly improves the our UNet backbone in terms of Acc, AUC, Sen, Spe, and MIOU by 0.0107 ($p < 0.01$), 0.0420 ($p < 0.01$), 0.3323 ($p < 0.01$), 0.0054 ($p < 0.01$), and 0.2318 ($p < 0.01$), respectively. The effect size for Acc, AUC, Sen, and Spe are 2.01, 3.44, 5.71, 1.1, and 3.74, respectively. The multiscale input improves the performance of baseline U-Net in terms of Acc, AUC, Sen, Spe, and MIOU by 0.0105 ($p < 0.01$), 0.0408 ($p < 0.01$), 0.3316 ($p < 0.01$), 0.0054 ($p < 0.01$), and 0.2276 ($p < 0.01$), respectively. The corresponding effect size for Acc, AUC, Sen, Spe, and MIOU are 2.09, 3.35, 5.56, 1.30, and 3.74, respectively. The attention module significantly improves the performance compared to the multiscale architecture in terms of AUC, Sen, and MIOU by 0.0012 ($p < 0.05$), 0.007 ($p < 0.05$) ,and 0.0042 ($p < 0.05$) with effect size 0.5, 0.3, and 1.8, respectively. The attention module slightly improves ACC and Spe compared to the multiscale architecture.

Tables 5 and 6 show the performance of the proposed method for each sample. For the DRIVE dataset, the best case ACC, AUC, Sen, and Spe are 0.9719, 0.9930, 0.8809, and 0.9914, respectively, and the worst case measures are 0.9533, 0.9809, 0.6972, and 0.9778, respectively. For the CHASEDB1 dataset, the best case ACC, AUC, Sen, and Spe are 0.9861, 0.9930, 0.884, and 0.9932, respectively, and the worst case measures are 0.9735, 0.9856, 0.7665, and 0.9805, respectively. The AUC of our method is the highest for two datasets compared to other methods, The multiscale architecture lets the network learn a vessel feature from a different scale, and the attention module further lets the network discard irregular region and focus on the most discriminative region.

Table 5. The performance for each case of the DRIVE dataset inside FOV.

Image ID	Acc	AUC	Sen	Spe
01	0.9613	0.9899	0.8201	0.9788
02	0.9619	0.9897	0.7767	0.9893
03	0.9533	0.9851	0.7151	0.9865
04	0.9614	0.9826	0.7496	0.9914
05	0.9570	0.9827	0.7224	0.9893
06	0.9549	0.9809	0.7058	0.9886
07	0.9590	0.9835	0.6972	0.9907
08	0.9579	0.9834	0.7034	0.9870
09	0.9632	0.9844	0.7522	0.9862
10	0.9606	0.9839	0.7752	0.9814
11	0.9608	0.9840	0.7603	0.9840
12	0.9643	0.9886	0.7990	0.9828
13	0.9571	0.9837	0.7297	0.9860
14	0.9647	0.9889	0.7897	0.9831
15	0.9683	0.9901	0.8465	0.9809

Table 5. *Cont.*

Image ID	Acc	AUC	Sen	Spe
16	0.9616	0.9905	0.7714	0.9863
17	0.9587	0.9863	0.7421	0.9833
18	0.9652	0.9906	0.8338	0.9790
19	0.9719	0.9930	0.8809	0.9822
20	0.9666	0.9907	0.8473	0.9778
average	0.9615	0.9866	0.7709	0.9847

Table 6. The performance for each case of the CHASEDB1 dataset inside FOV.

Image ID	Acc	AUC	Sen	Spe
11R	0.9861	0.9930	0.8843	0.9904
11L	0.9825	0.9914	0.7695	0.9910
12L	0.9735	0.9868	0.8348	0.9820
12R	0.9805	0.9892	0.8043	0.9891
13L	0.9830	0.9892	0.8637	0.9892
13R	0.9754	0.9871	0.7665	0.9869
14L	0.9840	0.9918	0.8236	0.9932
14R	0.9756	0.9856	0.8261	0.9805
average	0.9800	0.9892	0.8215	0.9877

5. Discussion

The dual attention network combines spatial attention and channel attention with a multiscale network. In this section, we analyze the time efficiency and the parameters of the proposed model.

Time Efficiency: The inference time for AM-Net is 12 ms on a 3090Ti Nvidia GPU for one image with a size of 512 × 512. The simple and effective architecture can be easily applied to smart AI applications.

Image Distortions: Our method resizes the input image to 512x512, which distorts the image contents. We conducted an experiment to analyze such distortions. The performance evaluations are shown in Tables 7 and 8. We randomly cropped the original image to patch with size 512 × 512 and trained the network with these patches. The performance decreased for the two datasets, especially for the CHASEDB1 dataset. The global context is more important for vessel segmentation. The image resolution of 512 × 512 contains enough small vessel information. Removing the distortions of small vessels is not the main reason for performance improvement. Everyone's blood vessels are different in detail but similar from a global perspective. For the DRIVE dataset, using the original image increases the performance. For the CHASEDB1 dataset, the original image was too large to fit into our GPU memory.

Table 7. The performance of different settings in the DRIVE dataset.

Method	Acc	AUC	Sen	Spe
Scale to 512 × 512	0.9615	0.9866	0.7693	0.9851
Origin Image	0.9626	0.9869	0.7911	0.9842
Random Crop 512 × 512	0.9508	0.8510	0.7166	0.9854

Table 8. The performance of different settings in the CHASEDB1 dataset.

Method	Acc	AUC	Sen	Spe
Scale to 512 × 512	0.9797	0.9895	0.8432	0.9863
Random Crop 512 × 512	0.9439	0.9385	0.6932	0.8963

Model Parameters: We compared our method with other methods built on UNet as shown in Table 9, including Yan et al. [24], jiang et al. [35], and Ma et al. [37]. The proposed network has 9.95 M parameters and 75.438 G flops. The multiscale backbone needs 9.415 M parameter and 73.163 G flops. The multiscale input and side-outputs improve the performance without additional parameters. The dual-attention module also improves the performance with 0.54 M parameters. Our network improves the performance mainly by the multiscale architecture without many parameters compared to other networks.

Table 9. The network parameters comparison of the proposed method with other methods based on U-Net.

Method	Year	Parameters
Yan et al. [24]	2018	30.96 M
jiang et al. [35]	2018	58.31 M
Ma et al. [37]	2021	13.39 M
Ours	-	9.95 M

The proposed network can perform fast inference and does not have many parameters. The simple and effective multiscale dual attention network can effectively detect vessel structures.

6. Conclusions

Vessel segmentation is a vital and challenging problem. The variance of vessel size and the low contrast vessels often harm the performance. This paper proposes a dual attention multiscale network to solve the problem. The network contains multiscale input, multiscale output, and a dual attention module. The multiscale architecture helps to detect the vessels with different sizes, and the dual attention module helps the network focus on the discriminative area. The dual attention module contains spatial attention and channel attention. The spatial attention gathers information from different positions, and the channel attention module models the relationship between different channels. Extensive experiments verify the effectiveness of the multiscale structure for vessel segmentation.

Author Contributions: Conceptualization, P.Y.; methodology, P.Y.; writing—original draft preparation, P.Y. and Y.F.; writing—review and editing, P.Y. and Y.F.; supervision, Q.W.; funding acquisition, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Project of Shenzhen Science and Technology Innovation Commission (KCXFZ20201221173610028), National Natural Science Foundation of China (NSFC) 61876208 and 62272172, Tip-top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program (2019TQ05X200) and 2022 Tencent Wechat Rhino-Bird Focused Research Program Research (Tencent WeChat RBFR2022008).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

DR	Diabetic Retinopathy
HR	Hypertensive Retinopathy
ReLU	Rectified Linear Unit
PAM	Positional Attention Module
CAM	Channel Attention Module

Spe	Specificity
Sen	Sensitivity
Acc	Accuracy
AUC	Area Under ROC
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FN	False Negative
IOU	Intersection of Union
MIOU	Mean Intersection of Union

References

1. Smart, T.J.; Richards, C.J.; Bhatnagar, R.; Pavesio, C.; Agrawal, R.; Jones, P.H. A study of red blood cell deformability in diabetic retinopathy using optical tweezers. In Proceedings of the Optical Trapping and Optical Micromanipulation XII, International Society for Optics and Photonics, San Diego, CA, USA, 9–13 August 2015; Volume 9548, p. 954825.
2. Irshad, S.; Akram, M.U. Classification of retinal vessels into arteries and veins for detection of hypertensive retinopathy. In Proceedings of the 2014 Cairo International Biomedical Engineering Conference (CIBEC), Giza, Egypt, 11–13 December 2014; pp. 133–136.
3. Fraz, M.M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, 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] [PubMed]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
5. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.C.H.; Heinrich, M.P.; Misawa, K.; Mori, K.; McDonagh, S.G.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
6. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
7. Jin, Q.; Meng, Z.; Pham, T.D.; Chen, Q.; Wei, L.; Su, R. DUNet: A deformable network for retinal vessel segmentation. *Knowl.-Based Syst.* **2019**, *178*, 149–162. [CrossRef]
8. Wang, D.; Hu, G.; Lyu, C. Frnet: An end-to-end feature refinement neural network for medical image segmentation. *Vis. Comput.* **2021**, *37*, 1101–1112. [CrossRef]
9. Ni, J.; Wu, J.; Wang, H.; Tong, J.; Chen, Z.; Wong, K.K.; Abbott, D. Global channel attention networks for intracranial vessel segmentation. *Comput. Biol. Med.* **2020**, *118*, 103639. [CrossRef] [PubMed]
10. Mou, L.; Zhao, Y.; Fu, H.; Liu, Y.; Cheng, J.; Zheng, Y.; Su, P.; Yang, J.; Chen, L.; Frangi, A.F.; et al. CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging. *Med. Image Anal.* **2021**, *67*, 101874. [CrossRef] [PubMed]
11. Hao, D.; Ding, S.; Qiu, L.; Lv, Y.; Fei, B.; Zhu, Y.; Qin, B. Sequential vessel segmentation via deep channel attention network. *Neural Netw.* **2020**, *128*, 172–187. [CrossRef] [PubMed]
12. Li, R.; Li, M.; Li, J. Connection Sensitive Attention U-NET for Accurate Retinal Vessel Segmentation. *arXiv* **2019**, arXiv:1903.05558.
13. 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]
14. Yue, K.; Zou, B.; Chen, Z.; Liu, Q. Retinal vessel segmentation using dense U-net with multiscale inputs. *J. Med. Imaging* **2019**, *6*, 034004. [CrossRef] [PubMed]
15. Annunziata, R.; Garzelli, A.; Ballerini, L.; Mecocci, A.; Trucco, E. Leveraging multiscale hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation. *IEEE J. Biomed. Health Inf.* **2015**, *20*, 1129–1138. [CrossRef] [PubMed]
16. Wu, Y.; Xia, Y.; Song, Y.; Zhang, Y.; Cai, W. Multiscale network followed network model for retinal vessel segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 119–126.
17. Yin, P.; Yuan, R.; Cheng, Y.; Wu, Q. Deep guidance network for biomedical image segmentation. *IEEE Access* **2020**, *8*, 116106–116116. [CrossRef]
18. 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. Imaging* **2004**, *23*, 501–509. [CrossRef] [PubMed]
19. 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]
20. 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]
21. Zhang, J.; Dashtbozorg, B.; Bekkers, E.; Pluim, J.P.; Duits, R.; ter Haar Romeny, B.M. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Trans. Med. Imaging* **2016**, *35*, 2631–2644. [CrossRef]

22. Liskowski, P.; Krawiec, K. Segmenting retinal blood vessels with deep neural networks. *IEEE Trans. Med. Imaging* **2016**, *35*, 2369–2380. [CrossRef]
23. Maninis, K.K.; Pont-Tuset, J.; Arbeláez, P.; Van Gool, L. Deep retinal image understanding. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 140–148.
24. Yan, Z.; Yang, X.; Cheng, K.T. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 1912–1923. [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. Imaging* **2019**, *38*, 2281–2292. [CrossRef]
26. Zhuang, J. LadderNet: Multi-path networks based on U-Net for medical image segmentation. *arXiv* **2018**, arXiv:1810.07810.
27. Wang, B.; Qiu, S.; He, H. Dual Encoding U-Net for Retinal Vessel Segmentation. In Proceedings of the MICCAI, Shenzhen, China, 13–17 October 2019; pp. 84–92.
28. Liu, B.; Gu, L.; Lu, F. Unsupervised Ensemble Strategy for Retinal Vessel Segmentation. In Proceedings of the MICCAI, , Shenzhen, China, 13–17 October 2019; pp. 111–119.
29. Wu, Y.; Xia, Y.; Song, Y.; Zhang, D.; Liu, D.; Zhang, C.; Cai, W. Vessel-Net: Retinal vessel segmentation under multi-path supervision. In Proceedings of the MICCAI, Shenzhen, China, 13–17 October 2019; pp. 264–272.
30. Atli, I.; Gedik, O.S. Sine-Net: A fully convolutional deep learning architecture for retinal blood vessel segmentation. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 271–283. [CrossRef]
31. 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.
32. Gao, J.; Huang, Q.; Gao, Z.; Chen, S. Image Segmentation of Retinal Blood Vessels Based on Dual-Attention Multiscale Feature Fusion. *Comput. Math. Methods Med.* **2022**, *2022*, 8111883. [CrossRef] [PubMed]
33. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. 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. 2235–2239.
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
35. Jiang, Y.; Tan, N. Retinal Vessel Segmentation Based on Conditional Deep Convolutional Generative Adversarial Networks. *arXiv* **2018**, arXiv:1805.04224.
36. Li, X.; Jiang, Y.; Li, M.; Yin, S. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Trans. Ind. Inf.* **2020**, *17*, 1958–1967. [CrossRef]
37. Ma, Y.; Zhu, Z.; Dong, Z.; Shen, T.; Sun, M.; Kong, W. Multichannel retinal blood vessel segmentation based on the combination of matched filter and U-Net network. *Biomed Res. Int.* **2021**, *2021*, 5561125. [CrossRef] [PubMed]