2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

AMF-NET: Attention-aware Multi-scale Fusion Network for Retinal Vessel Segmentation

Qi Yang, Bingqi Ma, Hui Cui and Jiquan Ma

Abstract— **Automatic retinal vessel segmentation in fundus image can assist effective and efficient diagnosis of retina disease. Microstructure estimation of capillaries is a prolonged challenging issue. To tackle this problem, we propose attention-aware multi-scale fusion network (AMF-Net). Our network is with dense convolutions to perceive microscopic capillaries. Additionally, multi-scale features are extracted and fused with adaptive weights by channel attention module to improve the segmentation performance. Finally, spatial attention is introduced by position attention modules to capture long-distance feature dependencies. The proposed model is evaluated using two public datasets including DRIVE and CHASE_DB1. Extensive experiments demonstrate that our model outperforms existing methods. Ablation study valid the effectiveness of the proposed components.**

Index Terms— **Retinal vessel segmentation, U-Net, attention mechanism, multi-scale fusion**

I. Introduction

Retinal vessel segmentation from fundus images has been proved of its contribution in quantitative analysis of ophthalmologic diseases, such as diabetic retinopathy (DR) and glaucoma [1]. DR is a complication caused by diabetes, which will induce structural changes of retinal vasculature [2]. Due to the complexity of vascular morphology, it is challenging for accurate segmentation of retinal vessels, especially for capillaries.

Traditional approaches for retinal vessel segmentation are based on the optimization of filtering. Various filtering methods have been proposed, such as Hessian matrix-based filters, symmetry filter, and tensor-based filter [3,4,5]. Recent years, deep learning (DL) has been introduced in fundus retinal vessel segmentation and achieved exciting results. In this way, Liskowski et al. proposed a retinal vessel segmentation method based on Convolutional Neural Network (CNN) [6]. Following this framework, it was further extended with conditional random field (CRF) was used for post-processing [7]. Inspired by U-Net [8], Laibacher et al. proposed M2UNet by adding pretrained components in the encoder part and contractive bottleneck blocks in the decoder part [9]. Zhuang et al. reported a chain of multi-path U-Nets (LadderNet), which has multiple pairs of encoder-decoder branches to extract semantic information [10]. Although existing methods have achieved exciting performance in vessel segmentation. However, it is still a challenging issue for capillary.

To improve the accuracy of retinal vessel segmentation, especially for capillary, we propose AMF-Net with multi-scale feature extraction and fusion. Firstly, channel attention module (CAM) is computed in dense block to weight the feature maps between channels and select useful feature map. Secondly, position attention module (PAM) is used to acquire long distance self-similarity in the feature maps for a better prediction. Furthermore, in order to prevent the loss of feature information, we design multi-scale fusion (MSF) are fused in the decoder part. The rest of this paper is organized as follows: Our retinal vessel segmentation methods and datasets are presented in Sect.II. The experiment results are shown in Sect.III. Finally, the conclusion is given in Sect.IV.

II. Data and Methods

A. Data: DRIVE and CHASE_DB1 Datasets

DRIVE is a colour fundus dataset, which is established from Dutch diabetic retinopathy screening project in 2004. There are 40 images acquired from 40 subjects age from 25 to 90, including 7 diabetic retinopathy and 33 normal cases. Image size is 565 × 584 vessel regions have been manually segmented by two experts. In our experiments, 20 images are randomly selected for training, and the rest for testing. CHASE_DB1 dataset is collected from the left and right eyes of 14 school-age children. There are 28 fundus images where the image size is 999 × 960 vessel regions have been manually segmented by two experts. In our experiments, 20 images are randomly selected for training, and the rest for testing.

B. Methods and Mathematical Frameworks

The architecture of network is visualize as Figure 1. In order to enlarge the receptive field and extract fine detailed capillaries' features. The encoder is composed of four dense blocks where atrous convolution layers with dilation rate of 1, 3 and 5 are cascaded in a dense manner. Followed the dense encoder, CAM is applied to weight the high level features by global average pooling. For intermediate features, PAM is used to exploit spatial correlation to reconstruct micro-structures in capillaries by a skip connection. Finally, multi-scale feature maps are fused by MSF, which fully utilize features at different levels guided by channel attention and spatial correlations.

For the capillaries at the ends of the retinal vessels and the branches, traditional convolution cannot achieve satisfied result due to the small receptive field in the feature maps. In order to solve this problem, we use atrous convolutions with

Qi Yang is with the Heilongjiang University, Harbin, Heilongjiang China (e-mail: 10197014366@qq.com).

Bingqi Ma is with the Heilongjiang University, Harbin, Heilongjiang China (e-mail: mabibgqi1@gmail.com).

Hui Cui is with the La Trobe University, Sydney Australia (e-mail: l.cui@latrobe.edu.au).

Jiquan Ma is with the Heilongjiang University, Harbin, Heilongjiang China (corresponding author phone: 18686825590; e-mail: majiquan@hlju.edu.cn).

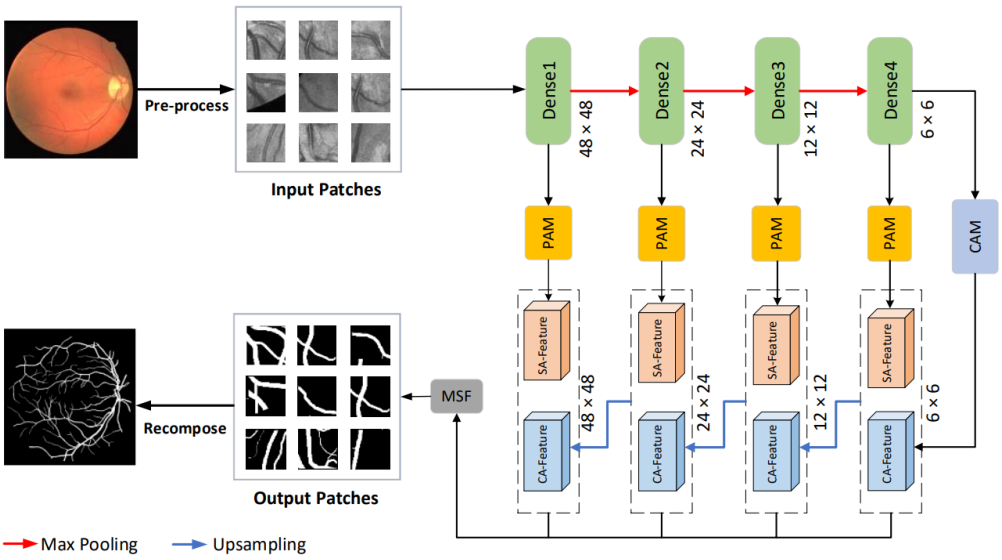

Fig 1. Overall architecture of the proposed network. Input image patches are fed into dense blocks and channel attention module (CAM) for channel attention enhanced feature (CA-Feature), position attention modules (PAM) are used to extract global information and spatial attention enhanced features (SA-Feature). Finally, multi-scale SA-Feature and CA-Feature are fused by multi-scale fusion (MSF) module to generate segmentation results.

different dilation rates to make feature extraction for micro-structure in fundus image through dense connections. The detailed architecture is shown in Figure 2(a). For each of the dense blocks, compared with the existing methods of extracting features through ordinary convolution, we use atrous convolutions with expansion rates of 1, 3 and 5 for dense feature extraction to expand the receptive field from 3 to 17. The 1 × 1 convolution is used at the last layer in dense block to compress channels and reduce the amount of calculation. This dense atrous convolution structure can extract more rich local information in the feature map, and in the meanwhile, mitigate gradient vanishing. Batch normalization, ReLU activation and dropout are performed after each atrous convolution. We use $A_{k,d}(x)$ to term an atrous convolution where k means kernel and d means dilation rate. Each dense block can be formulated as follows:

$$O = A_{1,1}(A_{3,5}(\mathrm{Cat}(A_{3,1}(I), A_{3,3}(A_{3,1}(I))))), \quad (1)$$

where I, O represent the input and output of dense block, and $Cat(\cdot)$ denotes the concatenation operation. In addition, we also add a CAM module to generate weight for feature maps in the channel direction. As shown in Figure 2(b), global average pooling is performed on the input feature map for output attention vector calculation. Two fully connected layers and Sigmoid activation function are applied to generate channel attention weight and make a element-wise product with the feature map. CAM is formulated as follows:

$$W_{CA} = \sigma(f_{C2}\delta(f_{C1}(P(I_C)))), \quad (2)$$

$$F_{CAM} = I_C \otimes W_{CA}, \quad (3)$$

where δ refers to the ReLU activation function, and σ refers to the Sigmoid activation function. F_{CAM}, I_C denote the output and input of CAM. The two weight matrices f_{C1}, f_{C2} are also the fully connected layers in the network. $P(\cdot)$ represents global average pooling.

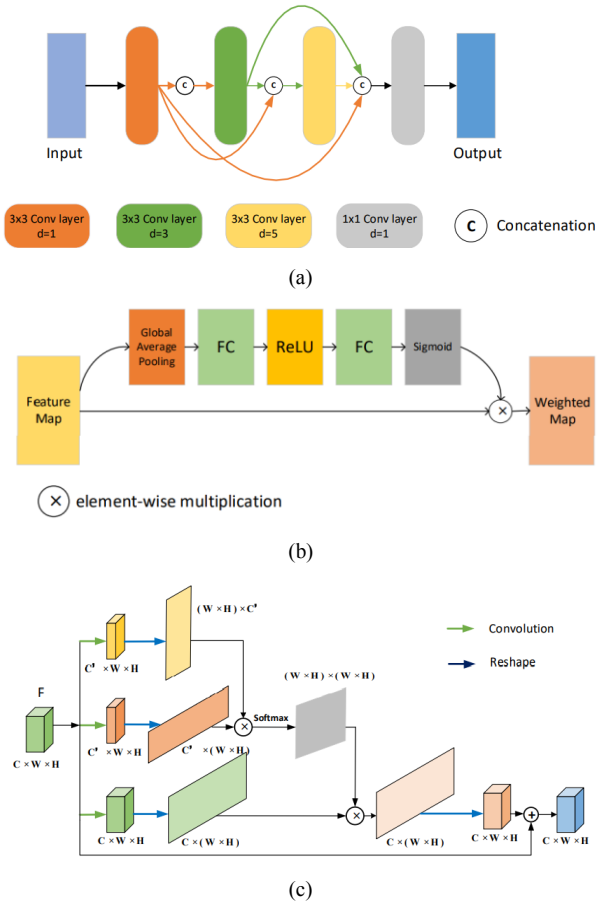

Fig 2. A architecture of (a) Dense Block, (b) Channel Attention Module (CAM) and (c) Position Attention Module (PAM)

Receptive fields in conventional deep models are limited to a local vicinity. As a result, it is hard to model broad and rich contextual representations. To address this issue, PAM is employed to our model. As shown in Figure 2(c), $F \in R^{C \times W \times H}$ denotes an input feature map to the attention

module, where C, W, H represent the channel, width and height dimensions, respectively. F is passed through three convolutional blocks, resulting in three feature maps $F_0 \in R^{C' \times W \times H}$, $F_1 \in R^{C' \times W \times H}$, and $F_2 \in R^{C \times W \times H}$ where C' is equal to C/16. Then, F_0 is reshaped to a feature map of shape $(W \times H) \times C'$ and F_1 is reshaped to $C' \times (W \times H)$. Both maps are multiplied, and Softmax is applied on the resulted matrix to generate the attention map $S \in R^{(W \times H) \times (W \times H)}$:

$$S_{i,j} = \frac{\exp(F_{0,i} \cdot F_{1,j})}{\sum_{i=1}^{W \times H} \exp(F_{0,i} \cdot F_{1,j})}, \quad (4)$$

where $S_{i,j}$ evaluates the impact of the i^{th} position on the j^{th} position. Then F_2 is multiplied by a permuted version of the attention map S, whose output is reshaped to a $C \times (W \times H)$. Thus, the attention feature map corresponding to the non-local attention module is formulated as follows:

$$F_{NLM,j} = \sum_{i=1}^{W \times H} S_{i,j} F_{2,j} + F_j. \quad (5)$$

As regions of small sizes may be neglected after multi-level feature extraction, we perform up-sampling of feature maps to increase the resolution by bilinear interpolation on four feature maps with different resolutions. Then these features are fused as a tensor before finally obtaining a multi-scale feature map through convolution. The MSF is formulated as follows:

$$F_{MSF} = Conv(Cat(B_3(f_1), B_2(f_2), B_1(f_3), f_4)). \quad (6)$$

In this setting, where f_1, f_2, f_3, f_4 indicates the level in the architecture and $Cat(\cdot)$ means concatenation operation. $B_i(x)$ is a bilinear interpolation, where i means scale size of up-sampling. $Conv$ denotes 1 × 1 convolution. Multi-scale feature fusion network performs feature integration of different levels, including shallow features and deep features.

III. Experiments and Results

A. Evaluation Metrics and Comparisons

Evaluation measures include accuracy (AC), sensitivity (SE), specificity (SP), and area under the receiver operating characteristics curve (AUC). We divide the pixels in the segmented vessel map into true positive (TP), false positive (FP), false negative (FN) and true negative (TN) by comparing them with the corresponding ground truth labels. The evaluation results using DRIVE and CHASE_DB1 are given in Table I and Table II. As shown, our network achieved the highest AC of 0.9581, and AUC of 0.9824 which is 0.0008 and 0.0008 higher than the second-best methods Li et al. [13] and LadderNet [11] respectively. For CHASE_DB1 dataset, our network also achieved the highest SE of 0.8344, SP of 0.9881, AC of 0.9729, and AUC of 0.9919 which is 0.0081, 0.0058, 0.0074 and 0.0051 higher than the second-best methods Li et al. [13] and Xu et al. [12] respectively.

The segmentation results of two retinal images from the DRIVE and CHASE_DB1 datasets by baseline U-Net and LadderNet, the proposed network, and GT are given in Figure 3. The results show that our network has the ability to segment the shape of capillaries better than other methods and preserve more retinal vascular spatial structures.

TABLE I

Comparison Results with Existing Methods on DRIVE Dataset

Method	Years	SE	SP	AC	AUC
Fu et al. [7]	2016	0.7603	-	0.9523	-
M2UNet [9]	2018	0.7863	0.9755	0.9511	0.9544
LadderNet [10]	2018	0.8081	0.9770	0.9553	0.9767
Bo et al. [13]	2019	0.9740	0.9816	0.9567	0.9772
Xu et al. [11]	2020	**0.9751**	0.9812	-	0.9805
Li et al. [12]	2020	0.7735	**0.9838**	0.9573	0.9816
Ours	**2021**	0.8139	0.9826	**0.9581**	**0.9824**

TABLE II

Comparison Results with Existing Methods on CHASE_DB1 Dataset

Method	Years	SE	SP	AC	AUC
M2UNet [9]	2018	0.7056	0.9873	0.9581	0.9754
LadderNet [10]	2018	0.7856	0.9799	0.9620	0.9772
Bo et al. [13]	2019	0.8074	0.9821	0.9661	0.9812
Xu et al. [11]	2020	0.8263	0.9775	-	0.9868
Li et al. [12]	2020	0.7970	0.9823	0.9655	0.9815
Ours	**2021**	**0.8344**	**0.9881**	**0.9729**	**0.9919**

B. Ablation Study

We further perform ablation studies using both databases to demonstrate the contributions of each of the new components. In our network, the experimental results of baseline U-Net, U-Net with dense convolution (U-Net + DB), U-Net using dense convolution and attention block (U-Net + DB + CAB), our network w/o PAM (U-Net + DB + CAM + MSF) and our network are given in Table 3. Using densely convolution resulted in an improved AUC of 0.43% and 1.39% over DRIVE and CHASE_DB1. Dense convolution and attention block contributed to improved AUC of 0.11% on DRIVE dataset respectively. Our network w/o PAM block achieved a further improved AUC gain of 0.03% on both datasets. Finally, our network achieved a further improved AUC gain of 0.16% and 0.12%.

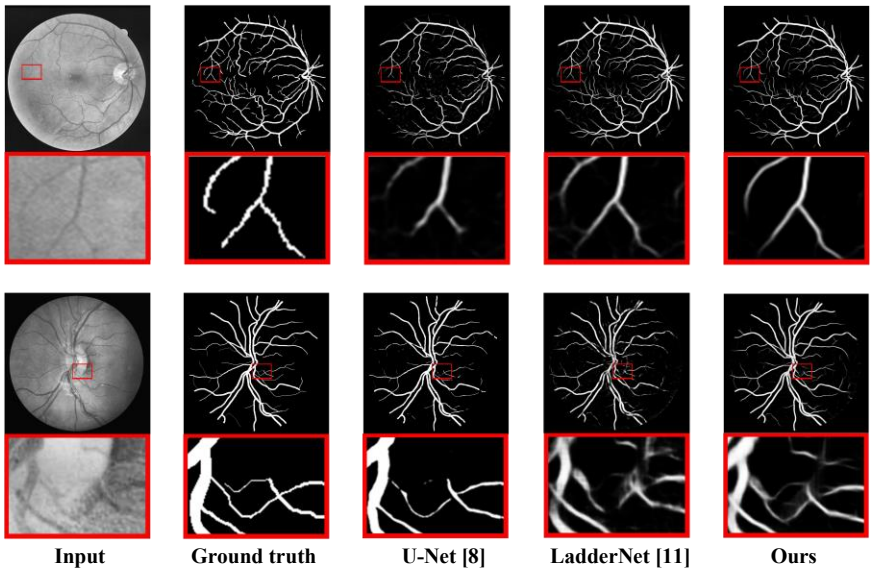

Fig. 3. Segmentation results by U-Net, LadderNet and ours network on DRIVE (first row) and CHASE_DB1 (second row) datasets. The red rectangle highlight improves fine detailed segmentation results.

TABLE III
ABLATION STUDIES USING THE SAME EXPERIMENT SETTINGS ON BOTH DATASETS

Method	DRIVE				CHASE_DB1			
	SE	***SP***	***AC***	***AUC***	***SE***	***SP***	***AC***	***AUC***
U-Net [8]	0.7598	0.9803	0.9499	0.9744	0.8288	0.9701	0.9578	0.9772
U-Net + DB	0.7855	0.9807	0.9558	0.9778	0.8331	**0.9884**	0.9713	0.9911
U-Net + DB + CAM	**0.8246**	0.9747	0.9556	0.9798	**0.8527**	0.9845	0.9710	0.9907
U-Net + DB + CAM + MSF	0.7904	0.9811	0.9565	0.9801	0.8352	0.9873	0.9714	0.9910
U-Net + DB + CAM + MSF + PAM (Ours)	0.8139	**0.9826**	**0.9581**	**0.9824**	0.8344	0.9881	**0.9729**	**0.9919**

IV. CONCLUSION

We proposed AMF-Net for retinal vessel segmentation especially for fine-detailed structures. The newly designed dense blocks and MSF demonstrated effectiveness in capturing rich multi-scale information. The channel attention module added attentional weights for details to the feature map. The PAM further improved the performance of capillary segmentation Evaluation results on two public datasets suggest the improved performance over state-of-the-art methods.

REFERENCES

[1] Fraz, M.M. et al., "Blood vessel segmentation methodologies in retinal images - a survey," Comput. Methods Progr. Biomed. 108(1), 407–433, 2012.

[2] Mohamed et al., "Management of diabetic retinopathy: a systematic review," JAMA 298(8), 902–916, 2007.

[3] Zhao, Y. et al., "Automatic 2D/3D vessel enhancement in multiple modality images using a weighted symmetry filter," IEEE Trans. Med. Imag. 37(2), 438–450, 2018.

[4] Frangi, A.F. et al., "Multiscale vessel enhancement filtering," MICCAI, vol. 1496, pp. 130–137, 1998.

[5] Cetin, S., Unal, G., "A higher-order tensor vessel tractography for segmentation of vascular structures," IEEE Trans. Med. Imag. 34, 2172–2185, 2015.

[6] Liskowski et al., "Segmenting retinal blood vessels with deep neural networks," IEEE Trans. Med. Imag. 35, 2369–2380, 2016.

[7] Fu, H. et al., "DeepVessel: retinal vessel segmentation via deep learning and conditional random field," MICCAI, vol. 9901, pp. 132–139, 2016.

[8] Ronneberger et al., "U-net: convolutional networks for biomedical image segmentation," MICCAI, vol. 9351, pp. 234–241, 2015.

[9] Laibacher, T. et al., "M2U-net: effective and efficient retinal vessel segmentation for resource-constrained environments," arXiv preprint arXiv:1811.07738, 2018.

[10] Zhuang, J., "Laddernet: multi-path networks based on U-Net for medical image segmentation," arXiv preprint arXiv:1810.07810, 2018.

[11] Xu, R. et al., "Semantics and multi-scale aggregation network for retinal vessel segmentation," ICASSP 2020.

[12] Li, Liangzhi, et al. "IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks." 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3656–3665,2020.

[13] Bo W., Shuang Q. et al, "Dual Encoding U-Net for Retinal Vessel Segmentation," MICCAI, vol. 11764, pp. 84–92, 2019.