MULTISCALE ATTENTION AGGREGATION NETWORK FOR 2D VESSEL SEGMENTATION

Wentao Liu[1]*,Huihua Yang*[1,2]*,Tong Tian*[3]*, Xipeng Pan*[2,4]*, Weijin Xu*[1]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China
[3]School of Aeronautics and Astronautics, Dalian University of Technology, Dalian, China
[4]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

ABSTRACT

Vessel segmentation is essential for clinical diagnosis and surgical planning. However, it is quite challenging for automatic blood vessel segmentation due to low contrast, complex structure, and variable scale, especially when the annotated data is scarce. In this paper, we propose a novel multiscale attention aggregation network (MAA-Net) for vessel segmentation. In MAA-Net, based on a U-shaped encoder-decoder architecture, the dual attention module with scale factors is employed behind the decoder at each stage to generate multi-resolution feature maps adaptively weighted by channel and location attention. In this way, multiscale contextual information with long-range dependencies can be captured to tackle scale variations of vessels. Meanwhile, these attention feature maps are gradually integrated into multi-level aggregate supervision to assemble multi-scale context information for refining segmentation results. The proposed method was evaluated on the retinal vessel and coronary angiography dataset (DRIVE and DCA1). Results demonstrate that MAA-Net achieves state-of-the-art performance for vessel segmentation. The code will be available at: `https://github.com/lseventeen/MAA-Net-Vessel-Segmentation`.

***Index Terms*—** Vessel segmentation, Long-range dependencies, Dual attention mechanism, Multi-level aggregate supervision.

1. INTRODUCTION

Accurate segmentation of blood vessels (e.g., retinal vasculature from color fundus images, X-ray coronary angiography, digital subtraction angiography) is essential for many clinical applications. Manual annotating of the blood vessels is a subjective, tedious, and labor-intensive task for human operators. As a result, many researchers have devoted themselves to developing automated vessel segmentation technology, which has become a hot topic in computer-aided medical systems. However, it is quite challenging for automatic blood vessel segmentation due to low contrast, complex structure, and variable scale.

Many automated techniques for segmenting blood vessels have been presented in recent decades, such as handcrafted features[1], filtering-based models[2], and statistical models[3]. These methods aim to enhance boundary gradient, remove undesired background information, and filter image noise, thereby simplifying the segmentation problem into a mathematical optimization problem with a fixed pattern. Recently, deep learning (DL)-based methods have achieved excellent results in vessel segmentation due to their extraordinary representation learning ability. Extensive research[4]has presented that the performance of vessel segmentation based on DL is superior to other methods. In particular, after the milestone network, UNet[5], was proposed, various outstanding variant models[6, 7] emerged for vessel segmentation. Fu et al.[8] applied a multiscale convolutional neural network with a side-output layer for vessel segmentation and introduced conditional random field as post-processing to optimize results. Zhang et al.[9] proposed the pyramid U-Net with pyramid-scale aggregation blocks that aggregate features at higher, current, and lower levels for accurate retinal vessel segmentation. The design of multiscale structures can effectively aggregate richer contextual information and has been applied in the networks mentioned above.

In addition, the attention mechanism can model long-range dependencies and has been widely applied in many tasks, including vessel segmentation. Mou et al.[10] added a dual self-attention module, consisting of spatial attention and channel attention, between the encoder and the decoder to further combine local features with their global dependencies. Wu et al.[11] proposed SCS-Net to capture multiscale contextual information by modeling correlations among feature channels between two adjacent layers and promoting feature fusion at different levels to obtain more semantic rep-

resentations. While these DL-based methods have reported encouraging results, we hypothesize that the vessel segmentation performance can be further improved by more effective modeling of the multiscale long-range dependencies.

Motivated by the above methods, we propose a multiscale attention aggregation network for vessel segmentation given the advantages of multiscale structural and attention mechanisms in representation learning. We deploy the dual attention module to the U-shaped encoder-decoder network composed of residual blocks. In contrast to previous works (DANet, CSNet), which only use the attention module between the encoder and decoder at the lowest scale stage, we feed the feature maps into dual attention modules at all stages to produce multiscale attention feature maps capable of capturing long-range contextual information with multi-resolution.In addition, we gradually aggregate the multiscale attention maps from high to low level for deep supervision, which can learn more distinctive semantic representation for refining the vessel maps. The proposed method has been evaluated on the DRIVE[12] and DCA1[13] datasets. Experimental results show that the proposed MAA-Net has outperformed state-of-the-art approaches.

2. METHODOLOGY

2.1. Overall Architecture

The proposed MAA-Net includes Res-UNet, simplified dual attention module (DAM), and multilevel aggregated supervision (MAS), as shown in Fig. 1. The backbone of the network is Res-Net, which is composed of residual blocks[14], upsampling, and downsampling. The essence of them is the convolution (Conv) block containing a Conv layer, a batch normalization (BN) layer[15], and a LeakyReLU function with a negative slope of 0.1 in sequence. In the MAA-Net, the residual block is modified by adding a dropout layer after each BN layer to reduce overfitting. Additionally, we utilized a 1×1 convolution followed by a BN layer to halve the number of channels after skip connections in the decoder, aiming to apply residuals. We feed the output maps from the decoder into DAM and generate attention maps with long-range contextual information. Finally, MAS gradually aggregates these multiscale attention maps to produce the pixel-level segmentation map. The details of DAM and MAS are described below.

2.2. Dual Attention Module

Long-range dependencies are critical for blood vessels with tree-like structures that are distributed throughout the entire image. It can only be captured when convolutional operations process a local neighborhood repeatedly, which has several limitations (e.g., computationally inefficient, optimization difficulties)[16]. Therefore, we add a dual attention module behind each encoder to directly encode a wider range of contextual information into local features of position and

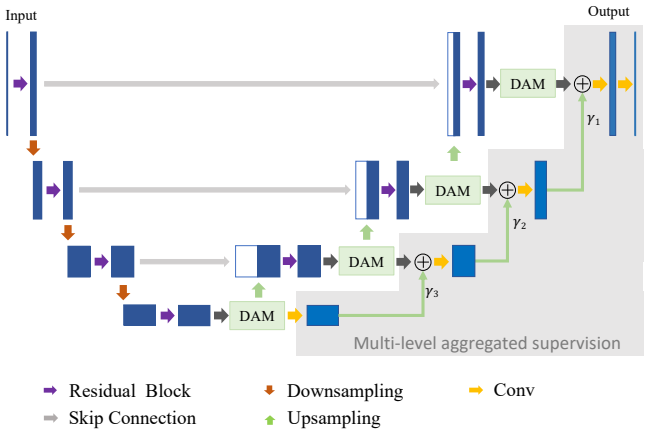

Fig. 1. The architecture of the proposed MAA-Net, which is implemented based on a Res-UNet and consists of dual attention modules and multiscale aggregated supervision.

channel, enhancing their representation capability. In contrast to previous works[17, 10], we apply the DAM at each stage, whereby the shallow-stage features can provide more refined attention information and the features at deep stages supplement high-level attention information. Meanwhile, to reduce the calculated amount of the dual attention module, we introduced the scale factors of the channel and space.

As illustrated in Fig. 2, given an input feature $A \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H and W denote the height and width of A, respectively. We first feed it into the Conv blocks to generate two new feature maps B and D in position attention. We reduced the number of output channels by β_1 to decrease the computational burden. They are then reshaped to $\mathbb{R}^{C/\beta_1\times(H\times W)}$. After that, we perform a matrix multiplication between the transpose of B and D, and apply a softmax layer to calculate the position attention S:

$$S_{j,i} = \frac{\exp(B_i^T D_j)}{\sum_{x=1}^{H\times W}\exp(B_i^T D_j)} \tag{1}$$

where $S_{j,i}$ indicates the i^{th} position's response on j^{th} position. Similarly, we feed input feature A into the Conv blocks to generate two new feature maps F and G in channel attention. We reduced the number of output sizes by β_2 to decrease the computational burden. They are then reshaped to $\mathbb{R}^{C\times(H\times W)/\beta_2^2}$, and the channel attention K is calculated as:

$$K_{j,i} = \frac{\exp(F_i G_j^T)}{\sum_{x=1}^{C}\exp(F_i K_j^T)} \tag{2}$$

where $K_{j,i}$ indicates the i^{th} channel's response on j^{th} channel. Meanwhile, we feed input feature A into the Conv blocks to generate two new feature maps E and L and reshape them to $\mathbb{R}^{C\times(H\times W)}$. Then we perform a matrix multiplication between E and the transpose of S. Similarly, a matrix multiplication is performed between K and L, we reshape them

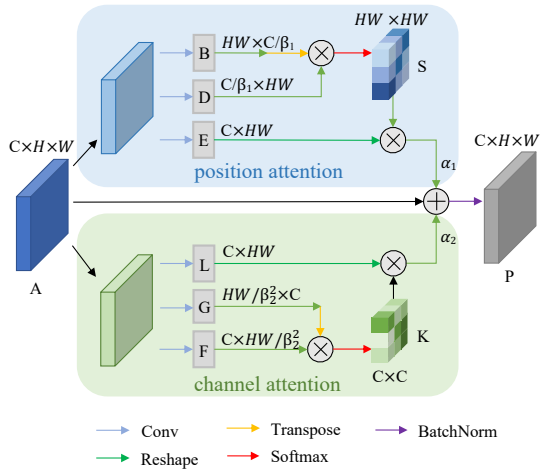

Fig. 2. The details of the simplified Dual Attention Mouble comprise of position attention and channel attention. (Best viewed in color)

to $\mathbb{R}^{C\times H\times W}$. Finally, we multiply the position and channel attention maps by the learnable scale parameters α_1 and α_2, respectively, and perform an element-wise sum operation with A to obtain the final output P as follows:

$$P_j = \alpha_1 \sum_{i=1}^{H\times W} (E_j S_{j,i}^T) + \alpha_2 \sum_{i=1}^{C} (K_{j,i} L_j) + A_j \quad (3)$$

As a result, the feature map P at each position and channel captures long-range dependencies by weighting the sum of the features across all positions and channels, which models global contextual and improves feature discriminability.

2.3. Multi-level Aggregated Supervision

The attention feature is up-sampled as the input of the next encoder. Meanwhile, the attention feature maps at all stages are gradually aggregated for deep supervision. The resulting multi-level aggregated supervision captures rich contextual relationships for better feature representations with multi resolution and obtains more semantic details to refine the segmentation results. As shown in Fig. 1, the operation of aggregated supervision is described as:

$$X_i = Conv(\gamma_i Up(X_{i-1}) + A_i) \quad (4)$$

where γ_i and A_i denote learnable scale parameters and attention maps at the i^{th} stage, respectively, and X_{i-1} indicates the result of aggregated supervision at the $i-1^{th}$ stage. Furthermore, we set the deepest stage as the initial stage (i.e., the 0^{th} stage), where the attention map is fed to a 3×3 Conv block to generate the initial state X_0. There are four stages in the proposed MAA-Net. Therefore, the aggregated supervision is performed three times, and the final segmentation map is obtained using 1×1 Conv. The MAS achieves deep supervision by gradually aggregating attention maps from deep to shallow stages, rather than directly upsampling the output maps of each scale to image size[11]or downsampling the supervision information to the corresponding scale[9], which effectively alleviates the segmentation performance degradation caused by scale difference.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We evaluated the proposed methods on the public retinal vessel and coronary angiography datasets: DRIVE and DCA1. The DRIVE dataset has 40 retinal images of 565×584 pixels with segmentation annotations officially divided into a training set and a test set, both containing 20 images. The DCA1 dataset consists of 134 X-ray coronary angiograms with 300×300 pixels in which the training set consists of 100 images, and the remaining 34 angiograms compose the test set. All images in the datasets are preprocessed by data normalization. Moreover, to increase the data diversity and reduce overfitting, we randomly extract 48×48 patches and perform data augmentation by horizontal flip, vertical flip, and rotation in the training phase.

We implemented the proposed MAA-Net under PyTorch and conducted experiments with a single GeForce RTX 3090. The Adam algorithm is used to minimize the joint loss consisting of cross entropy and dice loss with a weight decay of 1e-4 and an initial learning rate of 1e-3. The learning rate is gradually reduced by the cosine annealing algorithm in 40 iterations. The scale factors of channel and space (β_1 and β_2) are set to 8 and 2 in DAM. To evaluate the segmentation performance, we compared the segmentation results predicted by the network with the corresponding ground truths to calculate the area under the receiver operating characteristic curve (AUC) and evaluated the accuracy (Acc), sensitivity (Sen), specificity (Spe), F1 score (F1), and intersection (IOU) of the binary segmentation maps obtained by the threshold of 0.5.

3.2. Results

Ablation Study: To validate the effectiveness of DAM and MAS, we conducted ablation studies to evaluate how each component affects the results, using the DRIVE dataset as an example, which is shown in Table 1. The Res-UNet served as the baseline of the experiment. The results indicate that either DAM or MAS could improve the performance of vessel segmentation. DAM tends to generate high Sen because multi-level dual attention can capture long-range contextual information from low to high resolution. On the other hand, MAS has achieved a higher Acc due to aggregating hierarchical representations from all stages. Therefore, we embed both DAM and MAS blocks jointly in MAA-net, which leverages their advantages and achieves the best overall performance.

Table 1. Ablation studies on DRIVE. w/o means no corresponding modules.

Methods	Acc	Sen	Spe	AUC	F1
Res-Net	0.9688	0.8021	**0.9848**	0.9828	0.8184
MAA-Net w/o DAM	**0.9707**	0.815	0.9847	0.9876	0.8266
MAA-Net w/o MAS	0.9703	0.8376	0.9831	0.9880	0.8314
MAA-Net	0.9703	**0.8415**	0.9827	**0.9881**	**0.8325**

Table 2. Compared with state-of-the-art methods on DRIVE dataset.

Methods	Year	Acc	Sen	Spe	AUC	F1
Unet	2015	0.9678	0.8057	0.9833	0.9825	0.8141
UNet++	2018	0.9679	0.7891	0.985	0.9825	0.8114
ATT-UNet	2018	0.9662	0.7906	0.9831	0.9774	0.8039
CSNet	2019	0.9632	0.8170	**0.9854**	0.9798	0.8039
VSSC Net	2021	0.9627	0.7827	0.9821	0.9789	-
SCS-Net	2021	0.9697	0.8289	0.9838	0.9837	-
PYR-UNet	2021	0.9615	0.8213	0.9807	0.9815	-
MAA-Net	2021	**0.9703**	**0.8415**	0.9827	**0.9881**	**0.8325**

Comparing with State-of-the-art: We carried out vessel segmentation experiments on the most popular networks, such as UNet[5], UNet++[6], Attention U-Net[7], CS-Net[10]. In addition, we also compared the state-of-the-art vessel segmentation methods in the literature, including SCS-Net[11], VSSC Net[18] and Pyramid Unet[9] and Fernando[13]. Table 2, 3 provide qualitative results of vessel segmentation on the DRIVE and DCA1 datasets, respectively. It is obviously apparent from these tables that MAA-Net is competitive with other advanced methods by achieving the highest Acc, Sen, AUC and F1 for both datasets. Only the Spe is a bit lower.

Additionally, it is worth mentioning that the Sen of our method is much higher than the others, which shows that our approach yields fewer false negatives and can extract more vessel pixels accurately. We further visualize vessel segmentation results, including CSNet, MAA-Net, and corresponding original images and ground truths, as shown in Fig. 3. Compared to CSNet, which also uses a dual attention module, MAA-Net's segmented results are the closest to the ground truth due to retaining most of the spatial information of vessels. In particular, MAA-Net detects the vessel segment indicated by the arrow in the figure, but CSNet does not. The above comparisons demonstrate the strong capacity of the MAA-Net to tackle vessel segmentation.

4. CONCLUSIONS

In this paper, we propose a novel multiscale attention aggregation network for vessel segmentation. The network applies simplified dual attention modules at all stages to generate multi resolution feature maps weighted by channel and location attention. Furthermore, the multi-level aggregation supervision gradually integrated these feature maps for refining segmentation results. We have conducted comprehensive experiments on the vessel segmentation datasets DRIVE and DCA1. The ablation studies showed that both the DAM and the MAS could effectively improve segmentation performance. Compared with the state-of-the-art method, the proposed MAA-Net achieves better segmentation results.

Table 3. Compared with state-of-the-art methods on DCA1 dataset.

Methods	Year	Acc	Sen	Spe	AUC	F1
Unet	2015	0.9758	0.7816	0.9866	0.9879	0.7735
Unet++	2018	0.9761	0.7954	0.9862	0.9884	0.7786
ATT-UNet	2018	0.9755	0.7986	0.9853	0.9855	0.7748
CSNet	2019	0.9763	0.7895	0.9867	0.9889	0.779
Fernando	2019	0.9698	0.6364	**0.9880**	0.9775	-
VSSC Net	2021	0.97	0.7728	0.9809	0.9831	-
MAA-Net	2021	**0.9785**	**0.8545**	0.9854	**0.9929**	**0.8077**

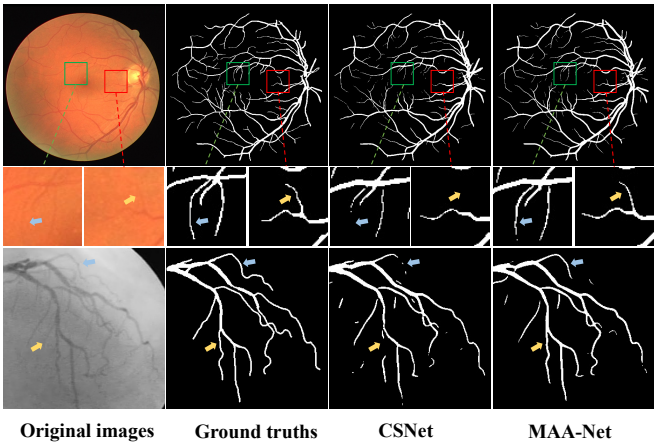

Fig. 3. Visualization of vessel segmentation results. The images in the first and second row are from the DRIVE and DCA1 datasets. The first column to the last column is the original images, ground truths, and the segmentation results of CSNet and MAA-Net, respectively. The retinal images contain many thin vessels, and we zoom in on the image details for clearer visualization, as shown in the highlighted rectangular regions.

Acknowledgment

This work was supported in part by the National Key R&D Program of China (No.2018AAA0102600) and National Natural Science Foundation of China (No.62002082).

5. REFERENCES

[1] Malihe Javidi, Hamid-Reza Pourreza, and Ahad Harati, "Vessel segmentation and microaneurysm detection using discriminative dictionary learning and sparse representation," *Computer Methods and Programs in Biomedicine*, vol. 139, pp. 93 – 108, 2017.

[2] Chenglong Wang, Masahiro Oda, Yuichiro Hayashi, Yasushi Yoshino, Tokunori Yamamoto, Alejandro F. Frangi, and Kensaku Mori, "Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation," *Medical Image Anal.*, vol. 60, 2020.

[3] Soodeh Kalaie and Ali Gooya, "Vascular tree tracking and bifurcation points detection in retinal images using a hierarchical probabilistic model," *Computer Methods and Programs in Biomedicine*, vol. 151, pp. 139 – 149, 2017.

[4] Dengqiang Jia and Xiahai Zhuang, "Learning-based algorithms for vessel tracking: A review," *Computerized Medical Imaging and Graphics*, vol. 89, pp. 101840, 2021.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[6] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, vol. 11045, pp. 3–11.

[7] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018.

[8] Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, and Jiang Liu, "DeepVessel: Retinal Vessel Segmentation via Deep Learning and Conditional Random Field," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 132–139.

[9] Jiawei Zhang, Yanchun Zhang, and Xiaowei Xu, "Pyramid u-net for retinal vessel segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 1125–1129, IEEE.

[10] Lei Mou, Yitian Zhao, Li Chen, Jun Cheng, Zaiwang Gu, Huaying Hao, Hong Qi, Yalin Zheng, Alejandro Frangi, and Jiang Liu, "CS-Net: Channel and Spatial Attention Network for Curvilinear Structure Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019, pp. 721–730.

[11] Huisi Wu, Wei Wang, Jiafu Zhong, Baiying Lei, Zhenkun Wen, and Jing Qin, "Scs-net: A scale and context sensitive network for retinal vessel segmentation," *Medical Image Anal.*, vol. 70, pp. 102025, 2021.

[12] Joes Staal, Michael D. Abràmoff, Meindert Niemeijer, Max A. Viergever, and Bram van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[13] Fernando Cervantes-Sanchez, Ivan Cruz-Aceves, Arturo Hernandez-Aguirre, Martha Alicia Hernandez-Gonzalez, and Sergio Eduardo Solorio-Meza, "Automatic Segmentation of Coronary Arteries in X-ray Angiograms using Multiscale Analysis and Artificial Neural Networks," *Applied Sciences*, vol. 9, no. 24, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, vol. 37, pp. 448–456.

[16] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[18] Pearl Mary Samuel and Thanikaiselvan Veeramalai, "VSSC net: Vessel specific skip chain convolutional network for blood vessel segmentation," *Comput. Methods Programs Biomed.*, vol. 198, pp. 105769, 2021.