ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414164

PYRAMID U-NET FOR RETINAL VESSEL SEGMENTATION

Jiawei Zhang[1,3]*, Yanchun Zhang*[2,3]*, Xiaowei Xu*[4]

[1] School of Computer Science, Fudan University, Shanghai, China
[2] College of Engineering and Science, Victoria University, Melbourne, Australia
[3] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
[4] Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China

ABSTRACT

Retinal blood vessel can assist doctors in diagnosis of eye-related diseases such as diabetes and hypertension, and its segmentation is particularly important for automatic retinal image analysis. However, it is challenging to segment these vessels structures, especially the thin capillaries from the color retinal image due to low contrast and ambiguousness. In this paper, we propose pyramid U-Net for accurate retinal vessel segmentation. In pyramid U-Net, the proposed pyramid-scale aggregation blocks (PSABs) are employed in both the encoder and decoder to aggregate features at higher, current and lower levels. In this way, coarse-to-fine context information is shared and aggregated in each block thus to improve the location of capillaries. To further improve performance, two optimizations including pyramid inputs enhancement and deep pyramid supervision are applied to PSABs in the encoder and decoder, respectively. For PSABs in the encoder, scaled input images are added as extra inputs. While for PSABs in the decoder, scaled intermediate outputs are supervised by the scaled segmentation labels. Extensive evaluations show that our pyramid U-Net outperforms the current state-of-the-art methods on the public DRIVE and CHASE-DB1 datasets.

Index Terms— Retinal Vessel Segmentation, U-Net, Pyramid Scale Aggregation, Deep Pyramid Supervision.

1. INTRODUCTION

Visible structure of retinal vascular may indicate many diseases. Accurate segmentation helps capture visible changes of retinal vascular structures, which assists doctors in diagnosis of eye related diseases [1, 2, 3]. Thus, it is particularly important in current retinal image analysis tasks. For example, hypertensive retinopathy is a retinal disease caused by hypertension, and patients with hypertension can find increased vascular curvature or stenosis [1]. Conventionally, manual segmentation is performed by experts, which is laborious, time-consuming, and suffers from subjectivity among experts. Automatic segmentation methods are highly demanded in clinical practice to improve efficiency as well as reliability and reduce the workload of doctors [4].

In practice, it is difficult to segment vessel structures such as thin capillaries well from the color retinal image, due to low contrast and ambiguousness. Profiting from the aggregation of multi-scale context information, a variety of deep neural networks [5, 6, 7, 8] have boosted medical image segmentation with a large margin, especially for small objects. For example, [9] scaled vessel images to both high and low resolutions, and employed two processes to extract and aggregate feature presentations at multiple scales. [10] integrated feature maps at all or most scales from the previous layers in the encoder sub-network to strengthen feature propagation and encourage feature reuse. However, current multi-scale dense connections can not ensure containing both higher-level and lower-level features, which may not fully explore and utilize the information with multiple scales. For example, the multi-scale dense connections can not supply the higher-level feature maps in the encoder sub-network for accurate localization, because all previous feature maps own a larger size than current layers.

Meanwhile, multi-scale inputs also attract more and more attention in medical image segmentation. For example, MIMO-Net [11] trained the network parameters using multiple resolutions of the input image connecting the intermediate layers for better localization. MILD-Net [12] proposed a fully convolutional neural network that counters the loss of information caused by max-pooling by re-introducing the original image at multiple points within the network. The above methods demonstrate that multi-scale input and deep supervision can efficiently improve the segmentation.

Motivated by the above discovery, in this paper, we propose pyramid U-Net for accurate retinal vessel segmentation. In pyramid U-Net, the proposed pyramid-scale aggregation blocks (PSABs) are employed in both the encoder and decoder to aggregate features at higher, current and lower levels. In this way, coarse-to-fine context information is shared and aggregated in each scale thus to improve the location of capillaries. To further improve performance, two optimizations including pyramid inputs enhancement and deep pyramid supervision are applied to PSABs in the encoder and decoder,

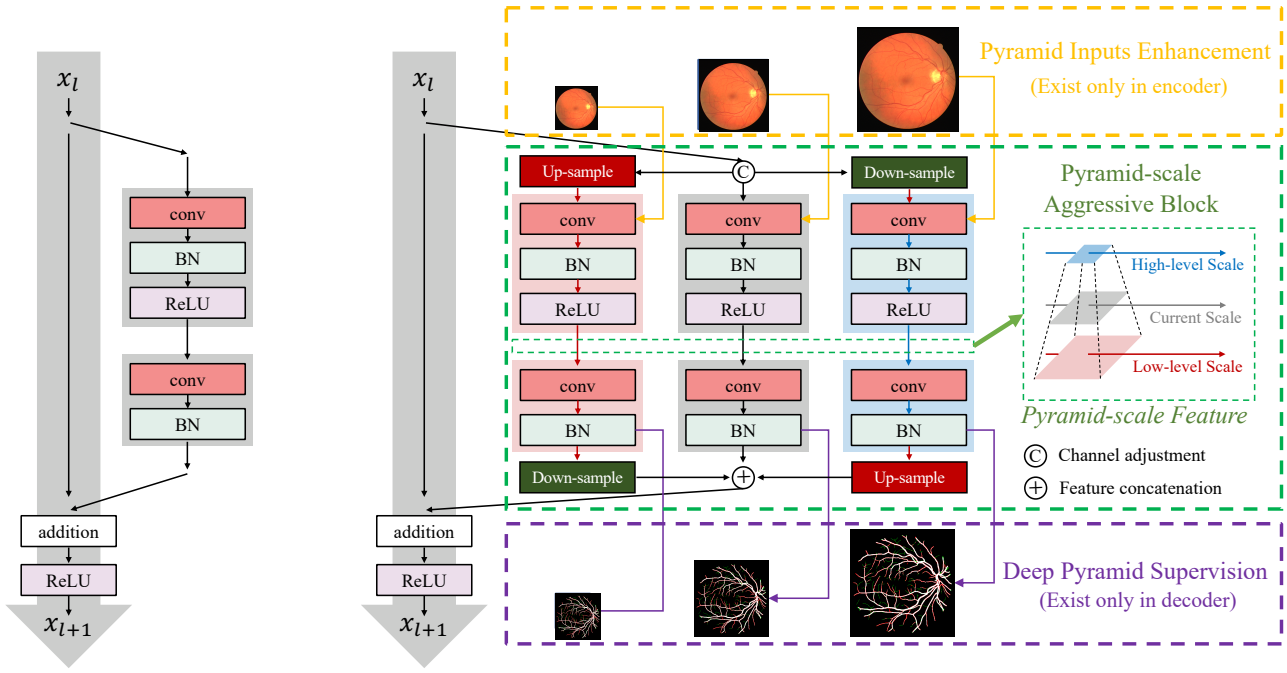

Fig. 1. Illustration of (a) ResNet blocks and (b) our pyramid-scale aggregation blocks (PSABs). PSABs (green rectangle) not only aggregate features with current scale but also from both the higher-level scale and lower-level scales containing coarse-to-fine context information. Meanwhile, pyramid input enhancement (yellow rectangle) and deep pyramid supervision (purple rectangle) are employed to fuse the inputs with corresponding scales and optimize the various-scale features learned in PSABs.

respectively. For PSABs in the encoder part, scaled input images are added into corresponding blocks as extra inputs to counter the information loss from the pooling layer. While for PSABs in the decoder part, scaled intermediate outputs are supervised by the scaled segmentation labels to optimize hierarchical representations and accelerate the training process. We have conducted comprehensive experiments on two retinal vessel image segmentation datasets including DRIVE [13] and CHASE-DB1 [14] with various segmentation networks including U-Net [15], DeepVessel [16], CE-Net [17]. The experimental results show that our method can significantly improve the segmentation and achieves state-of-the-art performance on the above two public datasets.

Our contributions in this paper can be summarized as follows. **1)** We introduce the pyramid-scale aggregation blocks (PASBs), which aggregate features at higher, current and lower levels to improve the segmentation performance. **2)** We propose pyramid input enhancement and deep pyramid supervision to further boost the performance. The former can reduce the loss of information caused by re-scaling in the encoder, while the latter can deeply supervise the learning process in the decoder. **3)** We conduct comprehensive experiments on two public datasets, and experimental results show that our methods can achieve state-of-the-art performance on both datasets.

2. METHODS

In this section, the details of pyramid U-Net are presented as shown in Fig. 2. We first introduce pyramid-scale aggregation block, and then describe two optimizations including pyramid input enhancement and deep pyramid supervision.

2.1. Pyramid-scale Aggregation Block

Pyramid-scale aggregation blocks (PSABs) are based on the widely adopted ResNet block [18]. The structure of ResNet block [18] is illustrated in the dashed box in Fig. 1, which is defined as

$$X_{l+1} = f(X_l) + X_l, \tag{1}$$

where X_l and X_{l+1} are the input and output of the current layer, while $f(\cdot)$ represents the learning process of the current layer. Fig. 1 illustrates the detailed structure of traditional ResNet blocks and our PSABs. Different from ResNet blocks, PSABs perform processing at three parallel pyramid scales including the higher, current and lower scales. In each scale, the processing steps are almost the same as that in traditional ResNet blocks. Some extra steps such as up-sampling and down-sampling are adopted at higher and lower scales to adjust scales. In order to reduce the potential increase of computational cost, the number of channels of the input X_l

has been reduced to half, while the number of channels of resized inputs X_l^p and X_l^d are reduced to one-fourth. The outputs of channel adjustment are fed to the processing steps at three scales and are processed in parallel. The three outputs at higher, current and lower scales are then concatenated. The whole process is formulated as follows,

$$\widetilde{X}_{l+1} = H(f(\widehat{X}_l^p), f(\widehat{X}_l), f(\widehat{X}_l^d)) + X_l, \quad (2)$$

where X_l^p and X_l^d are the up-sampled and down-sampled results of the current input X_l with channel adjustment, respectively. Meanwhile, $\widehat{X}_l^p$, $\widehat{X}_l$ and $\widehat{X}_l^d$ are the enhancement results by the pyramid input enhancement, which is detailed in section 2.2. $H(\cdot)$ represents the aggregation process, which performs re-scaling and feature concatenation. $\widetilde{X}_{l+1}$ is the strengthened results of X_{l+1} by PASB.

To improve the efficiency of feature extraction, we also employ an attention mechanism [19, 20] in PSAB as follows,

$$\mathbf{\Phi}(\widetilde{\mathbf{X}}_{l+1}) = \mathbf{W}(\Phi_{Avg}(\widetilde{\mathbf{X}}_{l+1})) + \mathbf{W}(\Phi_{Max}(\widetilde{\mathbf{X}}_{l+1})). \quad (3)$$

$$\mathbf{\Psi}(\widetilde{\mathbf{X}}_{l+1}) = \boldsymbol{\sigma}(\mathbf{\Phi}(\widetilde{\mathbf{X}}_{l+1}) \otimes \widetilde{\mathbf{X}}_{l+1}. \quad (4)$$

where $\mathbf{\Psi}(\cdot)$ is the operation of attention process, $\mathbf{W}$ is the conventional operation using 1×1 kernels for channel adjustment, and $\boldsymbol{\sigma}$ is the activation function. Average-pooling $\Phi_{Avg}(\cdot)$ and max-pooling $\Phi_{Max}(\cdot)$ are adopted for aggregating channel information.

2.2. Pyramid Input Enhancement

Pyramid input enhancement introduces the input image to PSABs to enlarge feature fusion. Particularly, the input image is scaled at higher, current and lower scales which are fed to three parallel processing steps at different scales in the PSAB to reduce the loss of information caused by scaling. The above three pyramid-scale images are concatenated with corresponding outputs of up-sampling, down-sampling and channel adjustment, respectively. For clearly, X_l is denoted as the input of the current layer, and X_l^p and X_l^d are results at higher and lower scales, respectively. Meanwhile, I_{l-1}, I_l and I_{l+1} are the corresponding scaled inputs of X_l^d, X_l and X_l^p, which have the same size. The fusion process of the current scale is formulated as follows,

$$\widehat{X}_{l-1} = H(X_l^d, \mathbf{W}^d(I_{l-1})), \quad (4)$$

$$\widehat{X}_l = H(X_l, \mathbf{W}(I_l)), \quad (5)$$

$$\widehat{X}_{l+1} = H(X_l^p, \mathbf{W}^p(I_{l+1})), \quad (6)$$

where $\mathbf{W}^p(\cdot)$, $\mathbf{W}^d(\cdot)$ and $\mathbf{W}(\cdot)$ represents 3×3 convolutional operations and is applied before concatenating to the features with pyramid-scale, and $H(\cdot)$ denotes the channel adjustment.

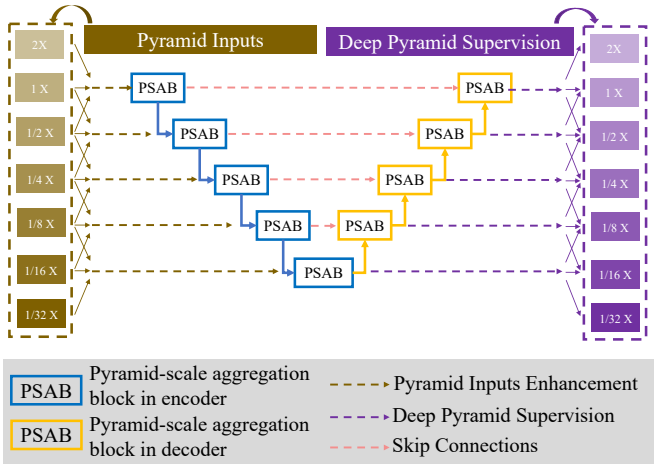

Fig. 2. Structure of the pyramid U-Net.

2.3. Deep Pyramid Supervision

In order to optimize hierarchical representations from multiple scales, deep pyramid supervision is further adopted. To realize deep pyramid supervision, the learned feature maps at multiple scales from each PSAB in the decoder is fed into a plain 3×3 convolutional layer followed by a sigmoid function. The deep pyramid supervision at the lth scale of the decoder can be defined as,

$$L_l = L(Y_l^p, M_{l-1}) + L(Y_l, M_l) + L(Y_l^d, M_{l+1}). \quad (7)$$

The ground truths M are scaled to the same size of the pyramid-scale feature maps for deep supervision, e.g., Y_l^p, Y_l and Y_l^d are supervised by the corresponding ground truth M_{l-1}, M_l and M_{l+1}, respectively. Each of the auxiliary feature masks is followed by a dropout layer (set to 50%) and a convolutional layer followed by a Softmax layer to get the auxiliary outputs. Our loss combines cross-entropy loss and Intersection over Union (IoU) loss.

3. EXPERIMENTS

3.1. Datasets

We used two public available retinal vessel datasets, DIRVE [13] and CHASE-DB1 [14] for evaluation. The DRIVE dataset contains 40 images with a resolution of 565 $\times$ 584 pixels, which were acquired using a Canon CR5 non-mydriatic 3CCD camera with a 45-degree field of view (FOV). Meanwhile, the FOV of each image is circular with a diameter of approximately 540 pixels, the images have been cropped around the FOV. The images are resized to 448 $\times$ 448 pixels. The set of 40 images has been divided into a training and a test set, both containing 20 images. In particular, we pre-train our network on PASCAL VOC 2012 to alleviate potential over-fitting due to the limited dataset for

retinal vessel segmentation. The CHASE-DB1 dataset consists of 28 vascular patch images with a resolution of 999 $\times$ 960. Following the configuration in [21], we use the first 20 images as the training set and the remaining 8 images as the test set.

3.2. Implementations and Evaluation

We implemented all experiments in PyTorch platform and trained models on an Nvidia GeForce Titan X machine with 12 GB memory. Meanwhile, we employed CE-Net as our backbones to implement PSABs, pyramid input enhancement and deep pyramid supervision. During training, we adopted Adaptive Moment Estimation (Adam) with a batch size of 4 and a weight decay of 0.0001. Training techniques like data argumentation (horizontal flip, vertical flip and diagonal flip) and learning rate decay are equipped. To evaluate our model, we used Sensitivity (Sen), Specificity (Spec), Accuracy (Acc) and Area Under the ROC Curve (AUC) as evaluation metrics. The aforementioned metrics are calculated as follows: $\mathrm{Sen} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $\mathrm{Spec} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$, $\mathrm{Acc} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$. If pixels are correctly classified to objects or backgrounds, they will be annotated as True Positive (TP) or True Negatives (TN), respectively. Meanwhile, pixels are misclassified to objects or backgrounds are labeled as False Positive (FP) or False Negatives (FN), respectively.

Table 1. Performance comparison of pyramid U-Net with state-of-the-art methods on the DRIVE dataset.

Method	Sens	Spec	Acc	AUC
Azzopardi [22] (2015)	0.7655	0.9704	0.9442	0.9614
Roychowdhury [23] (2015)	0.7395	0.9782	0.9494	0.9672
U-Net [15] (2015)	0.7531	0.9645	0.9445	0.9601
DeepVessel[16] (2016)	0.7612	0.9768	0.9523	0.9752
CE-Net [17] (2019)	**0.8309**	0.9747	0.9545	0.9779
Pyramid U-Net	0.8213	**0.9807**	**0.9615**	**0.9815**

3.3. Comparison with State-of-the-art Works

We compared our pyramid U-Net with existing state-of-the-art works including U-Net [15], DeepVessel [16] and CE-Net [17] on DRIVE and DECHASE1 datasets. The results are summarized in Table 1 and Table 2. On the DRIVE dataset, compared with the state-of-the-art method CE-Net, pyramid U-Net obtains a higher performance on Spec (improved by 0.6%), Acc (improved by 0.7%), AUC (improved by 0.36%) and a competitive result on Sen, which is only 0.32% worse. Meanwhile, our proposed pyramid U-Net outperforms state-of-the-art methods on all metrics on the CHASE-DB1 dataset. The visual comparison is shown in Fig. 3. White and black pixels are correct predictions of object and background, respectively, while red and green pixels are incorrect predictions. We can notice that our proposed pyramid U-Net evidently improves the segmentation performance. Especially for those narrow, low-contrast and ambiguous retinal vessels, which are highlighted by yellow rectangles.

Table 2. Performance comparison of pyramid U-Net with state-of-the-art methods on the CHASE-DB1 dataset.

Method	Sens	Spec	Acc	AUC
Azzopardi [22] (2015)	0.7585	0.9587	0.9387	0.9487
U-Net [15] (2015)	0.7675	0.9631	0.9409	0.9705
DeepVessel [16] (2016)	0.7412	0.9701	0.9609	0.9790
CE-Net [17] (2019)	0.7841	0.9725	0.9583	0.9787
Pyramid U-Net	**0.8035**	**0.9787**	**0.9639**	**0.9832**

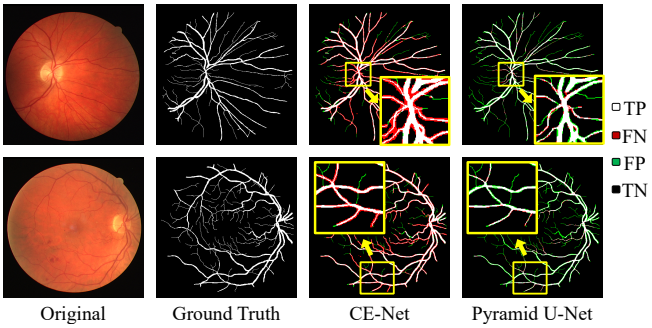

Fig. 3. Visual comparison of CE-Net [17] and our pyramid U-Net with corresponding original images and ground truths on the DRIVE dataset.

4. CONCLUSION

In this paper, we propose the pyramid U-Net for retinal vessel segmentation. In pyramid U-Net, the proposed pyramid-scale aggregation blocks (PSABs) are employed in both the encoder and decoder to aggregate features at higher, current and lower levels. To further improve performance, two optimizations including pyramid inputs enhancement and deep pyramid supervision are applied to PSABs in the encoder and decoder, respectively. We have conducted comprehensive experiments on two retinal vessel image segmentation datasets including DRIVE [13] and CHASE-DB1 [14]. Experimental results show that our pyramid-scale aggregation block can efficiently improve the segmentation performance.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.62006050).

6. REFERENCES

[1] Carol Yim-lui Cheung, Yingfeng Zheng, Wynne Hsu, Mong Li Lee, Qiangfeng Peter Lau, Paul Mitchell, Jie Jin Wang, Ronald Klein, and Tien Yin Wong, "Retinal vascular tortuosity, blood pressure, and cardiovascular risk factors," *Ophthalmology*, vol. 118, no. 5, pp. 812–818, 2011.

[2] Rubina Sarki, Khandakar Ahmed, Hua Wang, and Yanchun Zhang, "Automated detection of mild and multi-class diabetic eye diseases using deep learning," *HICS*, vol. 8, no. 1, pp. 1–9, 2020.

[3] Dinesh Pandey, Xiaoxia Yin, Hua Wang, and Yanchun Zhang, "Accurate vessel segmentation using maximum entropy incorporating line detection and phase-preserving denoising," *CVIU*, vol. 155, pp. 162–172, 2017.

[4] Xiaowei Xu, Qing Lu, Lin Yang, Sharon Hu, Danny Chen, Yu Hu, and Yiyu Shi, "Quantization of fully convolutional networks for accurate biomedical image segmentation," in *CVPR*, 2018, pp. 8300–8308.

[5] Jianxu Chen, Sreya Banerjee, Abhinav Grama, Walter J Scheirer, and Danny Z Chen, "Neuron segmentation using deep complete bipartite networks," in *MICCAI*. Springer, 2017, pp. 21–29.

[6] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *CVPR*, 2016, pp. 2487–2496.

[7] Jiawei Zhang, Yuzhen Jin, Jilan Xu, Xiaowei Xu, and Yanchun Zhang, "Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation," *arXiv preprint arXiv:1812.00352*, 2018.

[8] Jiawei Zhang, Yanchun Zhang, Shanfeng Zhu, and Xiaowei Xu, "Constrained multi-scale dense connections for accurate biomedical image segmentation," in *BIBM*. IEEE, 2020, pp. 877–884.

[9] Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai, "Multiscale network followed network model for retinal vessel segmentation," in *MICCAI*. Springer, 2018, pp. 119–126.

[10] Clément Playout, Renaud Duval, and Farida Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *MICCAI*. Springer, 2018, pp. 101–108.

[11] S. E. A. Raza, L. Cheung, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Mimo-net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images," in *ISBI*, April 2017, pp. 337–340.

[12] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *MedIA*, vol. 52, pp. 199–211, 2019.

[13] Joes Staal, Michael D Abramoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE TMI*, vol. 23, no. 4, pp. 501–509, 2004.

[14] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE TBE*, vol. 59, no. 9, pp. 2538–2548, 2012.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing, 2015.

[16] 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 *MICCAI*. Springer, 2016, pp. 132–139.

[17] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE TMI*, 2019.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[20] Sizhe Li, Jiawei Zhang, Chunyang Ruan, and Yanchun Zhang, "Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation," in *BIBM*. IEEE, 2019, pp. 818–825.

[21] Qiaoliang Li, Bowei Feng, LinPei Xie, Ping Liang, Huisheng Zhang, and Tianfu Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE TMI*, vol. 35, no. 1, pp. 109–118, 2015.

[22] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical image analysis*, vol. 19, no. 1, pp. 46–57, 2015.

[23] Sohini Roychowdhury, Dara D Koozekanani, and Keshab K Parhi, "Iterative vessel segmentation of fundus images," *IEEE TBE*, vol. 62, no. 7, pp. 1738–1749, 2015.