Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240513

Laddernet retinal vessel segmentation algorithm fusing attentional mechanisms

Hui Tan[1], Caijun Gan[2,3,*]

[1]Hui Tan, School of Computer and Information Security, Guilin University Of Electronic Technology, Guilin City, Guangxi, China
[2]Caijun Gan, School of Electronic Information and Automation,Guilin University Of Aerospace Technology, No. 2 Jinji Road, Guilin City, Guangxi, China

[3]imagan@qq.com
*corresponding author

Abstract. A retinal blood vessel segmentation algorithm fusing the attention mechanism is proposed to address the problems of insufficient accuracy and incorrect segmentation of minutiae in existing retinal blood vessel segmentation algorithms. The network structure used in this algorithm is Laddernet (LCT-UNet for short), which fuses the Convolutional Block Attention Module (CBAM) and Transformer, i.e., the residual module of Laddernet, which shares the weights, is structurally optimised to decrease the parameter count; After the residual module, CBAM is incorporated to adaptively adjust the weights of channels and spatial features to enhance the feature representation; Transformer with global self-attention mechanism is introduced at the bottom of the network to overcome the shortcomings of UNet's inability to model long-range relationships and spatial dependencies, and gradually refine the features so that the segmentation results have finer boundary and detail information. The model validity was verified on the public datasets DRIVE and CHASEDB1, and the experimental results showed that the ACC of LCT-UNet was 95.81% and 96.88%, SE was 79.74% and 81.12%, SP was 98.15% and 98.83%, and AUC was 98.22% and 98.99%, and the F1 values were 82.88% and 85.14%, respectively, and the comprehensive segmentation performance of the LCT-UNet network showed significant enhancement over the algorithms such as U-Net, R2UNet, and Laddernet.

Keywords: Image segmentation, retinal vessels, Laddernet, Transformer, CBAM.

1. Introduction

Fundus blood vessels can reflect the dynamics of the body's systemic blood circulation as well as its health status. Segmentation of blood vessels in retinal images is a necessary part of disease diagnosis based on fundus vascular features. Segmentation of blood vessels in retinal images is a necessary part of disease diagnosis based on fundus vascular features. Traditionally, this process is mainly done manually, which has limitations Traditionally, this process is mainly done manually, which has limitations such as low efficiency, high workload, and high influence of doctors' personal experience. Traditionally, this process is mainly done manually, which has limitations such as low efficiency, high workload, and high influence of doctors' personal experience.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/75/20240513

A novel structure combining U-Net [1] and Laddernet, has been proposed by Zhuang J T [2]: in each decoding layer, Laddernet introduces a Ladder Block for layer-by-layer feature propagation and correction from the bottom to the top layer. By passing and correcting features, Laddernet can By passing and correcting features, Laddernet can better handle fuzzy or noisy data in segmentation tasks, providing more accurate and robust segmentation results.

Transformer was originally a neural network structure based on Self-Attention mechanism (SAM) in the field of NLP [3]. In the application of image segmentation, Transformer is used to build the encoding and decoding phases of the segmentation task. According to researchers such as Jienneng Chen [4], Transformer is able to encode to capture global contextual information by transforming image blocks from the feature map of the CNN into labeled sequences. And in the decoding stage, the encoded features are up-sampled and combined with the high-resolution feature maps from the CNN to ensure precise localization. This combined U-Net approach recovers local spatial information, enabling Transformers to be used as an effective encoding tool in medical image segmentation tasks.

Attention mechanisms are mainly used to allow models to selectively focus on the important parts and ignore the unimportant parts when processing sequential data (e.g., natural language, audio, and images, etc.). CBAM involves the element-by-element multiplication of the output features from both the Channel Attention and Spatial Attention Modules to derive the enhanced attention-focused features.C. Zhu et al [5] proposed to introduce a series of CBAMs (CBAM) and Dense Hollow Convolution Modules (DVCM) (U-Net) at the bottom of the network to achieve a more accurate localization of the image segmentation task. segmentation task. and Dense Cavity Convolution Module (DAC) to expand the field of view and grasp more detailed vascular characteristics.

The above Laddernet, Transformer and CBAM all have their own advantages in image semantic segmentation, and if the three can be organically combined, it is possible to achieve complementary advantages and further improve the performance of retinal blood vessel segmentation. Based on this idea, this paper proposes Laddernet that integrates CBAM and Transformer. paper proposes Laddernet that integrates CBAM and Transformer, and the network structure takes Laddernet as the basic framework and makes the following improvements. Improvements.

(1) Modelling long-range relationships and spatial dependencies by introducing Transformer in order to capture contextual key information.
(2) By deploying the CBAM module after the residual block, diverse weights are realized to be applied to the foreground detail features, which in turn highlights the key features and reduces the attention to the secondary features.
(3) By improving the shared weights residual module, the total number of parameters is reduced and the generalization ability of the model is enhanced, thus improving the efficiency of network training

2. Method

2.1. Structure of the LCT-UNet design

The LCT-UNet encoder and decoder jump connection section uses a feature summation operation to connect the respective branch features of the encoder and decoder together. The decoder part, which consists of a total of 4 upsampling steps, where each upsampling step consists of 1 shared-weight residual block, 1 2-fold transposed upsampling, introduces a long jump connection to sum the feature maps extracted by the 4-layer convolutional network in the encoder part with the feature maps of the corresponding decoder upsampling layer, in order to recover the shallow detail feature information in the decoder path. Rather than utilizing conventional convolutional layers combined with a residual unit that shares weights, CBAM is incorporated into the encoder section to dynamically focus on both channel and spatial aspects of the features through channel and spatial attention mechanisms. Meanwhile, the underlying network employs a Transformer to grasp long-range dependencies and merge global feature details. The architecture of the LCT-UNet algorithm is depicted in Figure 1.

DOI: 10.54254/2755-2721/75/20240513

The Laddernet residual module before optimization differs from the traditional residual convolution block in that the two convolution layers in the same block have the same weights, thus greatly reducing the number of model parameters. On this basis, the fitting resistance is increased by replacing Dropout with Dropblock [6] and, at the same time, removing Batch Normalization. The comparison of the residual modules before and after optimization is shown in Figure 2.

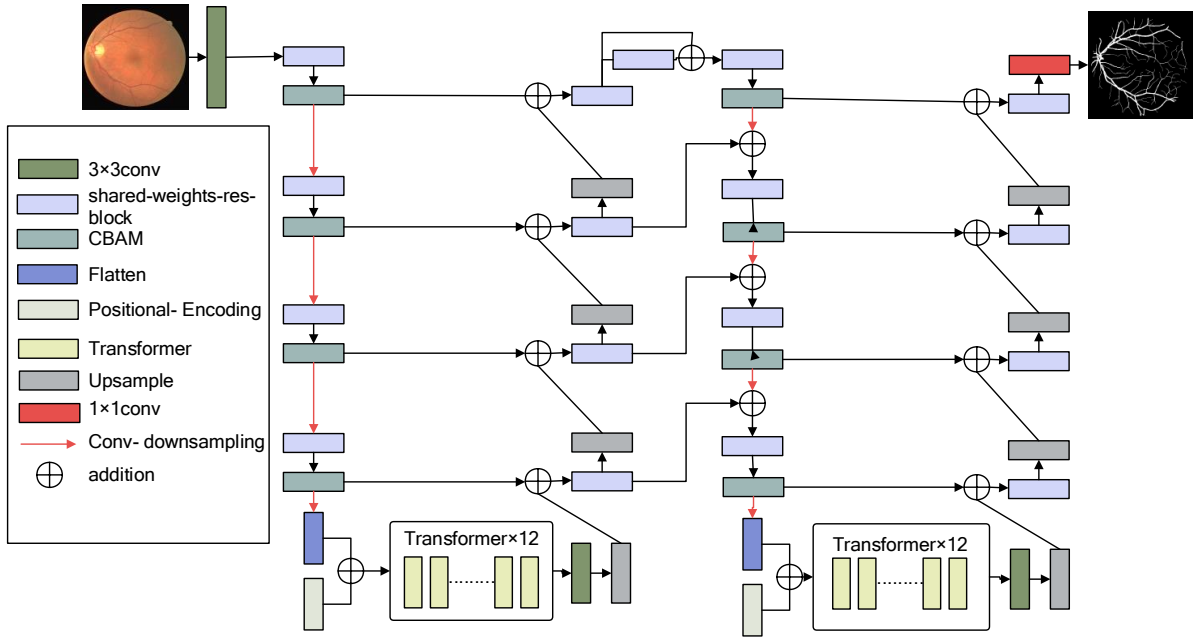


Figure 1. Structure of the LCT-UNet algorithm.

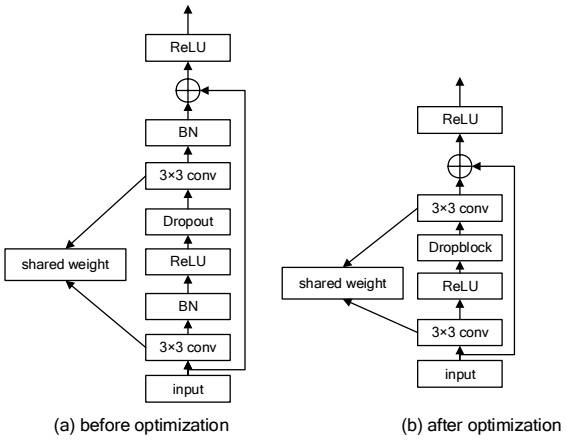


Figure 2. Optimization of shared weights residual module.

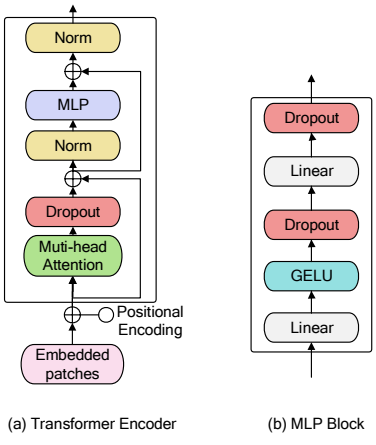


Figure 3. Structure of Transformer.

2.2. Transformer Module Design in LCT-UNet

First, to satisfy the Transformer's input format, the Transformer reshapes the input image into a sequence of 2-dimensional image block vectors to obtain image blocks $x = [x_1, x_2, \ldots, x_N] \in R^{N\times(P\times P)\times C}$, where the number of image block $N = HW/P^2$, the size of each image block is $P \times P$, and C is the number of

DOI: 10.54254/2755-2721/75/20240513

channels. Next, the serialized image blocks are mapped to a potential D-dimensional space using a trainable linear mapping. The position information is embedded into the sequence of image blocks using a learnable position encoding method, which further describes the position information of the individual image blocks again the whole space. The specific operation is shown in equation (1):

$$z_0=[x_1E_{patch};x_2E_{patch}\ldots;x_NE_{patch}]+E_{pos} \tag{1}$$

where z_0 denotes the labeled feature map at the pixel level of layer 0, and $E_{patch}\in R^{(P^2\times C)\times D}$ denotes the linear mapping mechanism, and $E_{pos}\in R^{N\times D}$denotes the positional encoding method.

Subsequently, the embedded image block is fed into the Transformer module consisting of Multi-headed Self-Attention (MSA) and Multi Layer Perception (MLP) block.The MSA module integrates multiple independent attention mechanisms and the attention function consists of 3 point-to-point (*Q, K, V*) linear layers, and the MSA representation is shown in equation (2). LN is applied after each module and residual connections are added. In this paper the number of Transformer layers in the encoder is set to 12 and the number of self-attention heads is 8. The Transformer Encoder and MLP Block are shown in Fig. 3. The output of the mth Transformer layer is computed as shown in equation (3) and (4):

$$MSA(Q,K,V)=Softmax(\frac{W_QW_K^T}{d^{1/2}})W_V \tag{2}$$

Where $W_Q\in R^{N\times d}$, $W_K^T\in R^{N\times d}$, $W_V\in R^{N\times d}$, d are the output dimensions.

$$z_m^{'}=LN(Dropout(MSA(z_{m-1}))+z_{m-1}) \tag{3}$$

$$z_m=LN(MLP(z_m^{'})+z_m^{'}) \tag{4}$$

Where $z_m^{'}$ is the output feature representation of the MSA module of the mth layer, z_m is the output feature representation of the mth Transformer layer, LN denotes the normalization operation, Dropout denotes the regularization operation, and MLP denotes the multilayer perceptron block.

2.3. CBAM in LCT-UNet

CBAM [7] combines a channel and a spatial attention module in series to better highlight key features while minimizing focus on background elements, aiding in the extraction of retinal vessel texture features.The channel attention module, firstly, performs parallel operations of global maximum pooling and global average pooling for each channel of the input feature map $F^{C\times H\times W}$ respectively to obtain two feature maps of $F^{C\times 1\times 1}$. Immediately after that, it is fed into the shared network layer (FC) consisting of two fully connected layers for processing, with the number of neurons in the first fully connected layer being *C/r* (r is the reduction rate), and the second fully connected layer being restored to the original number of channels. The outputs of the two paths are summed and the weight coefficients (ranging from 0 to 1) for each channel, M_c, are obtained by a sigmoid activation function. The channel attention module is represented as shown in equation (5):

$$M_c=\sigma\left(FC(AvgPool(F))+FC(MaxPool(F))\right) \tag{5}$$

where σ denotes the sigmoid function.The resulting feature map $F^{'}$, obtained by multiplying M_c with the input feature map *F*, serves as the input for the spatial attention module, Calculated as in equation (6).Firstly, $F^{'}$ undergoes two separate global max pooling and global average pooling operations along its channel dimension, creating two $F^{1\times H\times W}$ feature maps. These maps are then concatenated along their channel dimensions. Following this, the combined feature maps are processed by a 7×7 convolution layer to produce a single-channel spatial attention map. Subsequently, a sigmoid activation function is applied to determine the significance weight of each spatial position (with weights ranging from 0 to 1), denoted as M_s. The final step involves an element-wise multiplication of this feature map with the

DOI: 10.54254/2755-2721/75/20240513

module's input feature map F' , resulting in the CBAM's output features, M_{CBAM}.The output features of the spatial attention module and the CBAM final output representation is shown in equation (7), (8):

$$F' = M_c F \tag{6}$$

$$M_s = \sigma\left(Conv^{7\times7}\left(\left[AvgPool(F');MaxPool(F')\right]\right)\right) \tag{7}$$

$$M_{CBAM} = M_s F' \tag{8}$$

3. Experimentation and Discussion

3.1. Dataset

This paper conducted thorough experiments on two widely recognized fundus image datasets, DRIVE and CHASEDB1. The DRIVE dataset, created by Niemeijer et al [8], consists of 40 color fundus photographs, divided equally into training and testing groups, each with dimensions of 565 × 584 pixels. On the other hand, the CHASEDB1 dataset [9] comprises 28 color fundus photographs from both eyes of 14 school-aged children, offering a resolution of 999 × 960 pixels. Traditionally, the initial 20 images are designated for training, while the remaining 8 serve for testing. Both datasets feature binary maps of the fundus vasculature annotated by two specialists, with the annotations from the first expert being utilized for the experiments..

3.2. Ablation experiments

Table 1 displays the comparative performance of segmentation of major modules in the network on the CHASEDB1 dataset through ablation experiments. It is observed from the table that LadderNet outperforms UNet overall. The inclusion of refined residual blocks, denoted as ‘res’ in rows 2 to 3, showcases a significant improvement in fitting resistance by replacing Dropout with Dropblock. Rows 2, 3, and 4 demonstrate that introducing attention mechanisms in the encoder helps in more precise extraction of crucial features, reducing redundant information, and enhancing the model’s ability to capture both main vessels and fine branches in retinal vessels. Rows 2, 3, and 5 further prove that integrating Transformer structures can amalgamate global feature information, thereby enhancing the model’s perception and modeling capabilities of global information, which is more conducive to retinal vessel segmentation.

Table 1. Comparison of the performance of the modules for ablation on the CHASEDB1 dataset

Model	ACC	SE	SP	AUC	F1
UNet[1]	0.9604	0.7870	0.9777	0.9779	0.7828
Laddernet[2]	0.9656	0.7978	0.9818	0.9839	0.8031
Laddernet+res	0.9671	0.7797	0.9904	0.9891	0.8397
Laddernet+res +CBAM	0.9677	0.7734	**0.9918**	0.9901	0.8406
Laddernet+res +Transformer	0.9683	0.7871	0.9907	**0.9904**	0.8455
LCT-UNet	**0.9688**	**0.8112**	0.9883	0.9899	**0.8514**

3.3. Comparison experiment of different segmentation algorithms

In order to verify the effectiveness of LCT-UNet network on retinal vessel segmentation, experiments were carried out on both the DRIVE and CHASEDB1 datasets using identical parameter configurations, and the final segmentation prediction maps were compared for qualitative analysis. Meanwhile, in order to show the segmentation details of the networks more clearly, the experiments intercepted the local blood vessel details for comparison, and the segmentation effects of the three networks, UNet, Laddernet and LCT-UNet, are shown in Figure 4. The first 2 rows are the comparison of segmentation details in the DRIVE dataset, and the comparison algorithms are obviously insufficient for the segmentation of fine retinal vessels, and the UNet algorithm even appears to ignore the segmentation of fine vessels,

DOI: 10.54254/2755-2721/75/20240513

resulting in more background redundancy. The last 2 rows are the comparison of segmentation details in the CHASEDB1 dataset, and the comparison algorithm has the problem of broken and insufficient vessel segmentation, especially under the influence of the strong noise background in row 4, the comparison algorithm has the problem of broken vessels and false vessels, while the ability of the LCT-UNet network to capture the more subtle vascular features and the global contextual features ensures the connectivity of the subtle vessels.

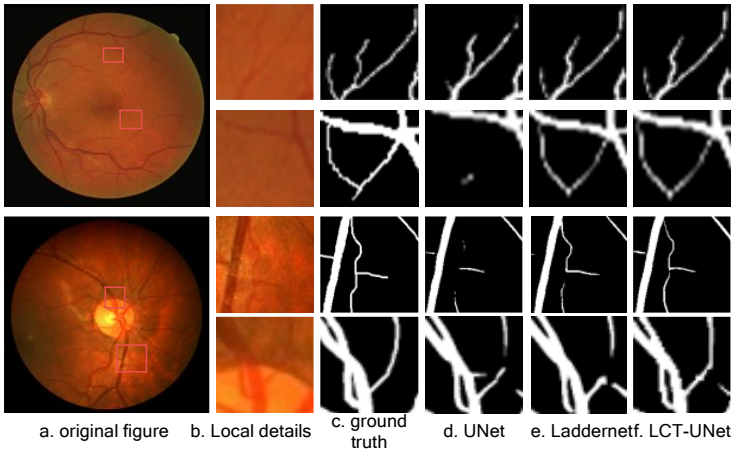


Figure 4. Comparison of algorithmic segmentation results.

To demonstrate the LCT-UNet network's enhanced performance in retinal vessel segmentation, it was quantitatively evaluated against several other segmentation algorithms, and the comparison results are shown in Table. For the five metrics, ACC, SE, SP, AUC, and F1, the LCT-UNet network outperformed the Laddernet algorithm by 0.20%, 1.18%, 0.05%, 0.29%, and 0.86%, respectively, on the DRIVE dataset, and by 0.32%, 1.32%, 1.32%, and 1.86%, respectively, on the CHASEDB1 dataset. On the CHASEDB1 dataset, LCT-UNet network is 0.32%, 1.34%, 0.65%, 0.60% and 4.83% higher than Laddernet algorithm respectively, so LCT-UNet network can improve the segmentation performance of the algorithm for Laddernet algorithm; on the DRIVE and CHASEDB1 datasets, LCT-UNet network is higher than the 2nd state-of-the-art algorithm in terms of the indicators of ACC, SE, AUC and F1. In the DRIVE and CHASEDB1 datasets, LCT-UNet network's ACC, SE, AUC and F1 metrics are higher than the 2nd state-of-the-art algorithm in ACC, SE, AUC and F1 metrics. In summary, LCT-UNet network has stronger anti-interference ability and balanced performance than other algorithms, and also has better global feature extraction ability of blood vessels and image feature recovery level.

Table 2. Performance comparison of the algorithms on the DRIVE dataset

Model	ACC	SE	SP	AUC	F1
UNet[1]	0.9531	0.7537	0.9820	0.9755	0.8142
R2Unet[10]	0.9556	0.7792	0.9813	0.9784	0.8171
Laddernet[2]	0.9561	0.7856	0.9810	0.9793	0.8202
MAUNet[11]	0.9557	0.7890	0.9799	0.9774	0.8192
IterNet[12]	0.9573	0.7735	**0.9838**	0.9816	0.8205
LCT-UNet	**0.9581**	**0.7974**	0.9815	**0.9822**	**0.8288**

DOI: 10.54254/2755-2721/75/20240513

Table 3. Performance comparison of the algorithms on the CHASEDB1 dataset.

Model	ACC	SE	SP	AUC	F1
UNet[1]	0.9604	0.7870	0.9777	0.9779	0.7828
R2Unet[10]	0.9634	0.7756	0.9820	0.9815	0 7928
Laddernet[2]	0.9656	0.7978	0.9818	0.9839	0.8031
MAU-Net[11]	0.9620	0.7798	0.9822	0.9791	0.8037
IterNet[12]	0.9655	0.7970	0.9823	0.9851	0.8073
LCT-UNet	**0.9688**	**0.8112**	**0.9883**	**0.9899**	**0.8514**

4. Conclusion

In order to effectively solve the problems of insufficient accuracy and segmentation errors in the minutiae of existing retinal blood vessel segmentation algorithms, this paper proposes a Laddernet retinal blood vessel segmentation algorithm incorporating an attention mechanism. The fitting resistance of the model is further increased by optimizing the shared weights residual module. Embedding CBAM in the encoder focuses on extracting the useful regions required for retinal vessel features. The Transformer with its own global self-attention mechanism is introduced at the bottom of the encoder to capture the long-distance dependency between features and integrate the global feature information, which alleviates the problems of vessel breakage and subtle vessels being overlooked. The experiment results indicate that a majority of inaccuracies in vessel segmentation stem from intense disturbances, such as irregular lighting, unclear lesions, and noise disruptions. When compared to other leading algorithms for retinal segmentation, LCT-UNet notably reduces instances of incorrect and missed segmentation, demonstrating superior resistance to interference and enhancing the precision of algorithms designed for retinal vessels segmentation.

References

[1] Ronneberger O, Fischer P, Brox T.U-Net:Convolutional networks for biomedical image segmentation[C] // Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention,2015:234-241

[2] Zhuang J T. LadderNet: multi-path networks based on U-Net for medical image segmentation[OL]. [2022-03-29]. http://arxiv.org/abs/1810.07810 .

[3] VASWANI, Ashish, et al. Attention is all you need. advances in neural information processing systems, 2017, 30.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation[J]. Computer Vision and Pattern Recognition. arXiv:2102.04306, 2021. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] // Proc of the 31st Conference on in Neural Information Proces- sing Systems, 2017:5998-6008.

[5] C. Zhu, X. Niu, L. Zuo and Z. Liu, "Fused Residual Attention Dense Double-U Network Retinal Vessel Segmentation Algorithm," 2023 International Conference on Intelligent Supercomputing and BioPharma (ISBP), Zhuhai, China, 2023, pp. 109-115, doi: 10.1109/ISBP57705.2023.10061315.

[6] G Ghiasi, T Y Lin and Q V Le, "Dropblock: a regularization method for convolutional networks[J]", Advances in neural information processing systems,. 2018.

[7] S Woo, J Park, J Y Lee et al., "Cbam: Convolutional block attention module[C]", Proceedings of the European conference on computer vision (ECCV), pp. 3-19 , 2018.

[8] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," Ieee Transactions on Medical Imaging, vol. 23, 2004:501-509.

DOI: 10.54254/2755-2721/75/20240513

[9] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," Ieee Transactions on Medical Imaging , 2006:1214-1222.

[10] Alom M Z, Hasan M, Yakopcic C, et al. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation[J]. arXvi. 1802.06955, 2018.

[11] Li H, Wang Y K, Wan C, et al. MAU-Net: a retinal vessels segmentation method[C] //Proceedings of the Annual International Conference of the IEEE Engineering in Medicine & Biology Society. Los Alamitos: IEEE Computer Society Press, 2020: 1958-1961.

[12] Li L Z, Verma M, Nakashima Y, et al. IterNet: retinal image segmentation utilizing structural redundancy in vessel networks[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2020: 3645-3654