

# Accurate retinal vessel segmentation in OCT-A images based on MsTCG-Net

Sumin Qi <sup>1, a</sup>, Baoyu Cui <sup>1, b \*</sup>, Mengqi Zhang <sup>1, c \*</sup>, Jing Meng <sup>2, d</sup> and Bangqiang Qi <sup>1, e</sup>

<sup>1</sup> Qufu Normal University, School of Cyber Science and Engineering, Qufu, China, 273165

<sup>2</sup> Qufu Normal University, School of computing, Rizhao, China, 276827

<sup>a</sup> qixm@qfnu.edu.cn, <sup>b</sup> bycuiqfu@163.com, <sup>c</sup> mengqiZ2000@163.com, <sup>d</sup> jingmeng@qfnu.edu.cn, <sup>e</sup> 493940122@qq.com

**Abstract.** Optical coherence tomography angiography (OCT-A) is a non-invasive visualization imaging technology with high-resolution that can more clearly image tiny blood vessels. Using OCT-A imaging technology, certain ophthalmic diseases can be better diagnosed by the morphological changes of retinal blood vessels. However, the task of segmenting retinal vessels is still very challenging due to the large variation in vessel size and shape and the presence of noise. In this paper, by introducing a transformer with a self-attention mechanism, we propose a novel multi-scale transformer-based channel and global attention network (MsTCG-Net) for segmentation of blood vessels in retinal OCT-A images. In MsTCG-Net, transformer-based channel joint attention (TC) block and transformer-based global joint attention (TG) block are proposed to capture multi-semantic features from spatial and channel dimension and fuse global contextual semantic features from different layers of encoder. Experimental results show that our proposed method achieves better segmentation performance than other state-of-the-art U-Net-based methods.

**Keywords:** Optical coherence tomography angiography; Vascular segmentation; Transformer.

## 1 Introduction

Fundus images contain a wealth of information, with a high density of retinal base vessels, whose changes are crucial for diagnosing a variety of systemic diseases. Consequently, routine ocular examinations are instrumental in the early identification and treatment of these conditions.

In the past, medical professionals commonly manually segmented retinal blood vessel images. However, this manual process was inefficient and error-prone. To address this issue, experts and scholars have developed several efficient automatic segmentation methods. The contemporary discourse on retinal vessel segmentation is predominantly centered around two imaging modalities: color fundus photography and Optical Coherence Tomography Angiography (OCT-A). The former exhibits certain limitations, particularly in the precise visualization of the microvasculature in central regions and its sensitivity to variable lighting conditions. In stark contrast, OCT-A imaging affords high-fidelity representations of the retinal vascular network, effectively circumventing the intrinsic limitations associated with traditional fundoscopic imaging [1], making OCT-A-guided vascular segmentation a leading methodology in the research field.

In recent years, convolutional neural network-based methods for retinal vessel segmentation have demonstrated significant effectiveness [2]. In particular, the introduction of U-Net [3] has established itself as an efficient segmentation framework. Furthermore, the various iterations of U-Net have further enhanced the segmentation performance [4-5]. Recently, transformer-based architectures have been incorporated into the domain of medical image segmentation to address the limitations of deep learning models in capturing long-range dependencies, with a current trend favoring the hybrid of transformers and U-Net as a foundational framework [6-7].

Therefore, in this paper, a new multi-scale transformer-based channel and global attention network (MsTCG-Net) is proposed as shown in Fig. 1. A novel transformer-based channel joint attention (TC) block is proposed and embedded between the encoder and decoder to improve the

ability of model to learn spatial and channel feature information. A novel transformer-based global joint attention (TG) block is designed to capture multi-scale global features with long-range dependency from different layers in the encoder. Furthermore, a new self-attention weight coefficient is designed and applied to the proposed MsTCG-Net to perform the retinal vessel segmentation task.

The rest of this paper is organized as follows. The section 2 describes MsTCG-Net, TC block and TG block. The experimental setup and results are shown in section 3. Conclusions and future work are carried out in section 4.

## 2 Methods

### 2.1 Network Architecture

Inspired by [8-9], we propose the MsTCG-Net shown as Fig. 1, which takes U-Net [3] as the basic framework and mainly consists of four parts: encoder, TC block, TG block, and decoder. The encoder has five blocks with sequentially applied convolution, batch normalization, and activation. The decoder mirrors the encoder, recovering spatial and channel details via multi-scale features from TC and TG blocks.

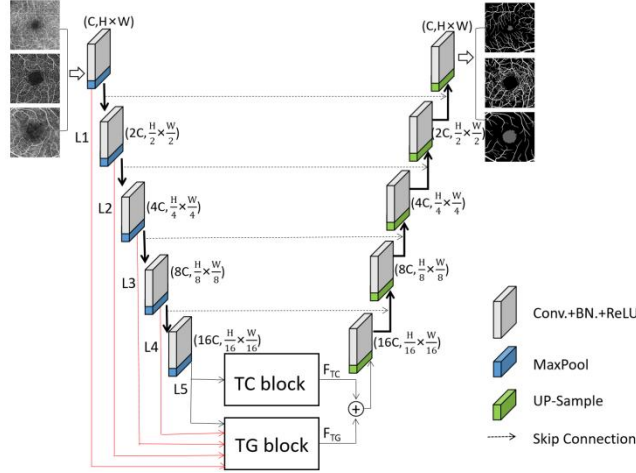


Fig. 1 MsTCG-Net

### 2.2 TC Block

To address the issue that transformers only focus on spatial information and neglect channel information, we set a new TC block and self-attention weight coefficient to enhance the extraction of multi-semantic contextual feature information in channel and spatial dimensions. The detailed structure of TC block is shown in Fig. 2, which is mainly composed of T(F5) in space dimension, C(F5) in channel dimension and F5 as described in Eq. (1).

$$F_{TC} = T(F5) + C(F5) + F5 \# (1)$$

LayerNorm applies normalization to  $A_t$  and  $V_t$  to obtain a non-local spatial response with strong semantics as Eq. (2). Attention map  $A_t$  is obtained by a Softmax operation, where similarity matrix  $Q_t^T$  and  $K_t$  are first multiplied to generate a new self-attention weight coefficient, and  $d_t$  is the dimension of  $Q_t$  and  $K_t$ .

$$T(F5) = \text{LayerNorm}(V_t \cdot A_t) = \text{LayerNorm} \left( V_t \cdot \text{Softmax} \left( \frac{Q_t^T \cdot K_t}{\log_2 d_t} \right) \right) \# (2)$$

$I_c$ ,  $M_c$  and  $N_c$  are results of reshaping F5.  $\text{Softmax}(I_c \cdot M_c)$  is used as a scale factor to highlight significant feature information in channel dimension as Eq. (3).

$$C(F5) = \text{Softmax}(I_c \cdot M_c) \cdot N_c \# (3)$$

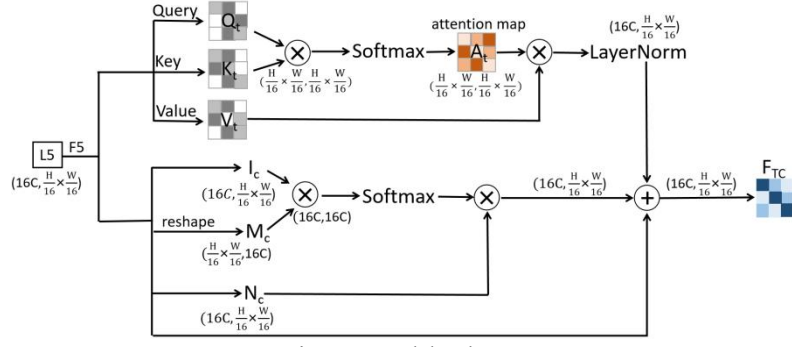


Fig. 2 TC block

## 2.3 TG Block

As shown in Fig. 3, the TG block accepts  $F_1 \sim F_5$  as input. It performs dimension normalization by down-sampling  $F_1 \sim F_4$  and achieves multi-scale integration through a convolution and summation process. The result is  $F_{add}$ , which matches the dimensions and channel count of  $F_5$ . Both  $F_{add}$  and  $F_5$  are then processed by the TG block to capture multi-scale global features with long-range dependencies. Different from the self-attention module [10] in previous studies, the TG block uses the  $F_{add}$  as the input of the Query branch, while the Key and Value branches use  $F_5$  as the input. The detailed structure of TG block is shown in Fig. 3, which is mainly composed of  $TG(F_5)$  in space dimension and  $F_5$ , that can be described as Eq. (4).

$$F_{TG} = TG(F_5) + F_5 \# (4)$$

Attention map  $A$  is obtained by a Softmax operation as Eq. (5), where similarity matrix  $Q^T$  and  $K$  are first multiplied to yield a new self-attention weight coefficient as  $E$ , position features are through matrix multiplication between  $P$  and  $Q^T$ .  $P_w$  and  $P_h$  represent learnable vectors encoded from the horizontal and vertical directions, which are optimized with training.

$$A = \text{Softmax}(E + Q^T \cdot P) = \text{Softmax}\left(\frac{Q^T \cdot K}{\log_2 d} + Q^T \cdot \text{reshape}(P_w + P_h)\right) \# (5)$$

$TG(F_5)$  is performed to obtain a global feature map with strong semantics as Eq. (6), where  $F_s$  represents the weighting of attention map  $A$  and the corresponding  $V$ .  $\lambda$  is a learnable parameter initialized to 0 and gradually adjusted during training to assign weights to  $F_s$  in a learnable manner.

$$TG(F_5) = \lambda \cdot F_s = \lambda \cdot (V \cdot A) \# (6)$$

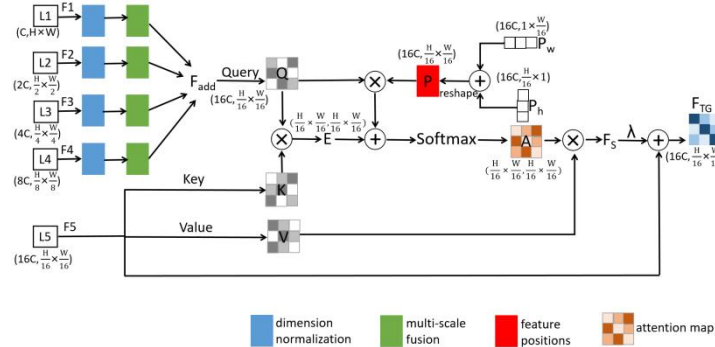


Fig. 3 TG block

## 3 Experimental Setup and Results

### 3.1 Datasets

In our experiments, we used two public datasets: ROSE-1(SVC+DVC) [11] from the Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, encompassing data from 39 participants, and OCTA-500 [12-13], which comprises OCTA\_6M and OCTA\_3M

subsets and six projection types. We selected the OCTA\_FULL (average) projection for our experiments.

We use an Intel(R) Xeon(R) Gold 6130 CPU @2.10GHZ, and we employ ADAM for optimization with an initial learning rate of 0.0001 and weight decay set to 0.0005. All images are resized to 512×512 pixels. The code of the proposed MsTCG-Net will be released in: <https://github.com/CBY1121/MsTCG-Net>.

### 3.2 Loss Function

Our experimental use mean square error (MSE) loss function as Eq. (7), where  $\hat{y}$  and  $\hat{y}_i$  is the predicted value,  $y$  and  $y_i$  is the true value, and  $n$  is the number of elements.

$$Loss(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad \#(7)$$

### 3.3 Experimental Results

(c)~(f) in Fig. 4~Fig. 6 shows the segmentation results of each of the models. In general, all models show similar performance on large-diameter vessels. However, in comparison to other models, our proposed MsTCG-Net demonstrates superior segmentation for small-diameter vessels due to its adaptive integration of local features and global dependencies. Meanwhile, the model suppresses irrelevant noise and ensures the continuity and integrity of tiny blood vessels to the greatest extent.

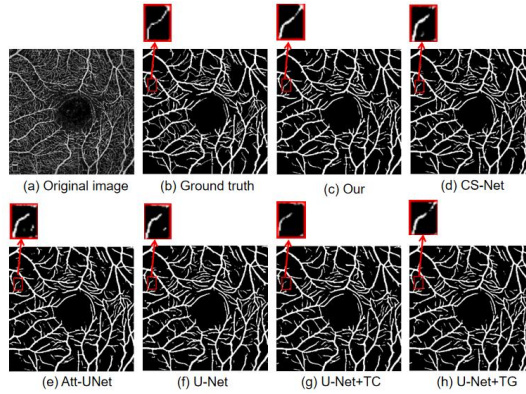


Fig. 4 Comparison of blood vessel segmentation of different models on ROSE-1(SVC+DVC).

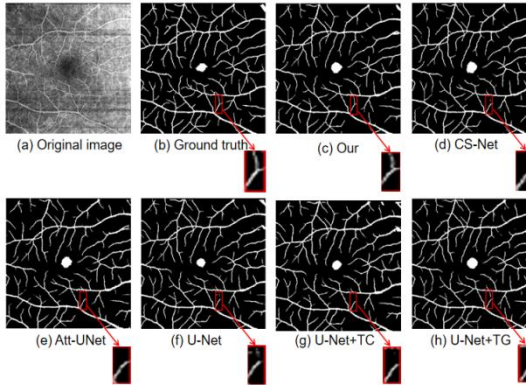


Fig. 5 Comparison of blood vessel segmentation of different models on OCTA\_6M.

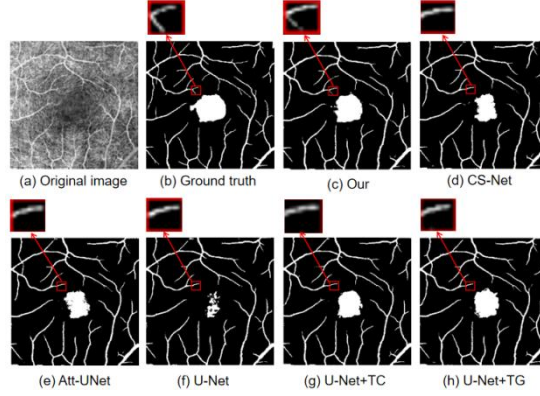


Fig. 6 Comparison of blood vessel segmentation of different models on OCTA\_3M.

Table 1 compares our model's performance to other advanced methods, showing our model excels in most metrics. It also includes ablation studies on the TC and TG blocks. The results of ablation experiments show that the proposed MsTCG-Net can improve the segmentation accuracy of retinal vessels in OCT-A images.

Table 1 Evaluation metrics for different networks on different datasets

Datasets	Networks	ACC	SEN	SP	Dice	Kappa	FDR	AUC
ROSE-1 (SVC+DVC)	U-Net [3]	0.9163	<b>0.7572</b>	0.9548	0.7727	0.7214	0.2070	0.9432
	Att-Unet [4]	0.9178	0.7441	0.9596	0.7728	0.7227	0.1913	0.9424
	CS-Net [5]	0.9183	0.7434	0.9605	0.7734	0.7237	0.1891	0.9427
	U-Net+TC	0.9197	0.7400	0.9631	0.7755	0.7269	0.1799	0.9443
	U-Net+TG	0.9195	0.7454	0.9614	0.7764	0.7274	0.1852	0.9444
	<b>MsTCG-Net</b>	<b>0.9210</b>	0.7339	<b>0.9660</b>	<b>0.7771</b>	<b>0.7294</b>	<b>0.1687</b>	<b>0.9447</b>
OCTA_6M	U-Net [3]	0.9514	0.7640	0.9764	0.7855	0.7581	0.1892	0.8702
	Att-Unet [4]	0.9531	0.7704	0.9776	0.7933	0.7669	0.1799	0.8740
	CS-Net [5]	0.9519	0.7741	0.9757	0.7898	0.7626	0.1916	0.8749
	U-Net+TC	0.9512	0.7932	0.9724	0.7917	0.7641	0.2079	0.8828
	U-Net+TG	0.9530	0.7919	0.9747	0.7976	0.7710	0.1948	0.8833
	<b>MsTCG-Net</b>	<b>0.9559</b>	<b>0.7936</b>	0.9777	<b>0.8084</b>	<b>0.7835</b>	<b>0.1746</b>	<b>0.8856</b>
OCTA_3M	U-Net [3]	0.9671	0.8135	<b>0.9853</b>	0.8375	0.8193	0.1320	0.8994
	Att-Unet [4]	0.9679	0.8341	0.9839	0.8461	0.8282	0.1391	0.9090
	CS-Net [5]	0.9707	0.8501	0.9852	0.8604	0.8440	0.1273	0.9176
	U-Net+TC	0.9725	0.8713	0.9846	0.8708	0.8553	0.1286	0.9279
	U-Net+TG	0.9725	0.8729	0.9843	0.8704	0.8550	0.1306	0.9286
	<b>MsTCG-Net</b>	<b>0.9741</b>	<b>0.8866</b>	0.9847	<b>0.8801</b>	<b>0.8656</b>	<b>0.1253</b>	<b>0.9357</b>

We further research statistical significance of proposed MsTCG-Net performance improvement in ACC and SEN by paired *T*-test, and the *p*-values are listed in Table 2, respectively.

Table 2 Statistical analysis (p-value) of proposed MsTCG-Net compared to other models

Datasets	Networks	ACC	SEN
ROSE-1(SVC+DVC)	MsTCG-Net&U-Net [3]	0.0086	<5e-4
	MsTCG-Net&Att-Unet [4]	<5e-4	0.0022
	MsTCG-Net&CS-Net [5]	0.0050	0.1435
OCTA-500	MsTCG-Net&U-Net [3]	0.0108	0.0014
	MsTCG-Net&CS-Net [5]	0.0585	<5e-4

## 4 Conclusions and Future Work

In this paper, we introduce the MsTCG-Net, which enhances the model's capability to learn multi-scale features and long-range dependencies by TC and TG modules with a U-shaped architecture. Additionally, we introduce a novel self-attention weight coefficient to accentuate vascular details. Experimental evidence demonstrates that MsTCG-Net outperforms existing U-Net-based networks in the task of retinal vessel segmentation. In our future work, we will collect more retinal OCT-A data to further evaluate the performance and robustness of the proposed MsTCG-Net.

## References

- [1] Spaide R F, Klancnik J M, Cooney M J. Retinal vascular layers imaged by fluorescein angiography and optical coherence tomography angiography. *JAMA ophthalmology*, 2015, 133(1): 45-50.
- [2] Liskowski P, Krawiec K. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 2016, 35(11): 2369-2380.
- [3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//*Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer International Publishing, 2015: 234-241.
- [4] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [5] Mou L, Zhao Y, Chen L, et al. CS-Net: Channel and spatial attention network for curvilinear structure segmentation//*Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*. Springer International Publishing, 2019: 721-730.
- [6] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [7] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation//*European conference on computer vision*. Cham: Springer Nature Switzerland, 2022: 205-218.
- [8] Feng S, Zhao H, Shi F, et al. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 2020, 39(10): 3008-3018.
- [9] Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020: 10076-10085.
- [10] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- [11] Ma Y, Hao H, Xie J, et al. ROSE: a retinal OCT-angiography vessel segmentation dataset and new model. *IEEE transactions on medical imaging*, 2020, 40(3): 928-939.
- [12] Li M, Chen Y, Ji Z, et al. Image projection network: 3D to 2D image segmentation in OCTA images. *IEEE Transactions on Medical Imaging*, 2020, 39(11): 3343-3354.
- [13] Li M, Zhang Y, Ji Z, et al. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020, 5: 16.