MsTGANet: Automatic Drusen Segmentation From Retinal OCT Images

Meng Wang, Weifang Zhu, Fei Shi, Jinzhu Su, Haoyu Chen, Kai Yu, Yi Zhou, Yuanyuan Peng, Zhongyue Chen, and Xinjian Chen, *Senior Member, IEEE*

***Abstract*—Drusen is considered as the landmark for diagnosis of AMD and important risk factor for the development of AMD. Therefore, accurate segmentation of drusen in retinal OCT images is crucial for early diagnosis of AMD. However, drusen segmentation in retinal OCT images is still very challenging due to the large variations in size and shape of drusen, blurred boundaries, and speckle noise interference. Moreover, the lack of OCT dataset with pixel-level annotation is also a vital factor hindering the improvement of drusen segmentation accuracy. To solve these problems, a novel multi-scale transformer global attention network (MsTGANet) is proposed for drusen segmentation in retinal OCT images. In MsTGANet, which is based on U-Shape architecture, a novel multi-scale transformer non-local (MsTNL) module is designed and inserted into the top of encoder path, aiming at capturing multi-scale non-local features with long-range dependencies from different layers of encoder. Meanwhile, a novel multi-semantic global channel and spatial joint attention module (MsGCS) between encoder and decoder is proposed to guide the model to fuse different semantic features, thereby improving the model's ability to learn multi-semantic global contextual information. Furthermore, to alleviate the shortage of labeled data, we propose a novel semi-supervised version of MsTGANet (Semi-MsTGANet) based on pseudo-labeled data augmentation strategy, which can leverage a large amount of unlabeled data to further improve the segmentation performance. Finally, comprehensive experiments are conducted to evaluate the performance of the proposed MsTGANet and Semi-MsTGANet. The experimental results show that our proposed methods achieve better segmentation accuracy than other state-of-the-art CNN-based methods.**

***Index Terms*—Optical coherence tomography, drusen, transformer, segmentation.**

I. Introduction

OPTICAL coherence tomography(OCT) is a non-invasive imaging technology used to visualize the cross-sectional retinal structure [1]–[3]. Many retinal diseases can be observed in OCT images clearly, such as macular hole [4], [5], choroidal neovascularization (CNV) [6], [8], pigment epithelial detachment (PED) [9], [10], optic disc edema [11], and central serous retinopathy [12], [13], etc. Therefore, OCT plays an important role in the diagnosis and monitoring of retinal diseases [14]–[18].

Age-related macular degeneration (AMD) is an irreversible and progressive chronic retinal disease, which is one of the main causes of vision loss worldwide [19]. Drusen, a local deposit of extracellular debris between the retinal pigment epithelium (RPE) and Bruch's membrane (BM), is considered as a key clinical sign and important risk factor for the development of AMD [20]. The follow-up assessment of drusen helps to understand the progress of AMD and the effectiveness of treatment [21]. Therefore, accurate segmentation of the drusen in retinal OCT images is crucial for early diagnosis of AMD.

Many previous studies focused on the detection and segmentation of drusen in fundus images, and have achieved good results [21]–[23]. Mittal and Kumari [21]. proposed an automated method for drusen detection based on three stages including bright region enhancement, drusen regions detection by suppressing spurious regions and edge linking. Akram *et al.* [22]. adopted support vector machine (SVM) based on manually designed drusen features for drusen segmentation. To tackle the limitation of the manually selected features, Ren *et al.* [23]. adopted the VGG network as the backbone to extract rich features in fundus color images and developed a novel deep learning based method for drusen segmentation. Pham *et al.* [24]. exploited both local and global information to improve the drusen segmentation performance in fundus color photography. Yan *et al.* [25]. presented a deep random walk technique for drusen segmentation from fundus images, which is mainly composed of three parts: a deep feature extraction module to learn both semantic-level and low-level representations of image, an affinity learning module to get pixel-pixel affinities for formulating the transition matrix of random walk and a random walk module which propagates

Manuscript received July 6, 2021; revised August 30, 2021; accepted September 12, 2021. Date of publication September 14, 2021; date of current version February 2, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700; and in part by the National Nature Science Foundation of China under Grant U20A20170, Grant 61622114, and Grant 81401472. *(Corresponding author: Xinjian Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB)/Ethics Committee, and performed in line with the United States Health Insurance Portability and Accountability Act (HIPAA) and the Declaration of Helsinki.

Meng Wang, Weifang Zhu, Fei Shi, Jinzhu Su, Kai Yu, Yi Zhou, Yuanyuan Peng, and Zhongyue Chen are with the Medical Image Processing, Analysis and Visualization Lab, School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu 215006, China (e-mail: wangmeng9218@126.com).

Haoyu Chen is with The Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, Shantou 515041, China (e-mail: drchenhaoyu@gmail.com).

Xinjian Chen is with the Medical Image Processing, Analysis and Visualization Lab, School of Electronics and Information Engineering and the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, Jiangsu 215006, China (e-mail: xjchen@suda.edu.cn).

Digital Object Identifier 10.1109/TMI.2021.3112716

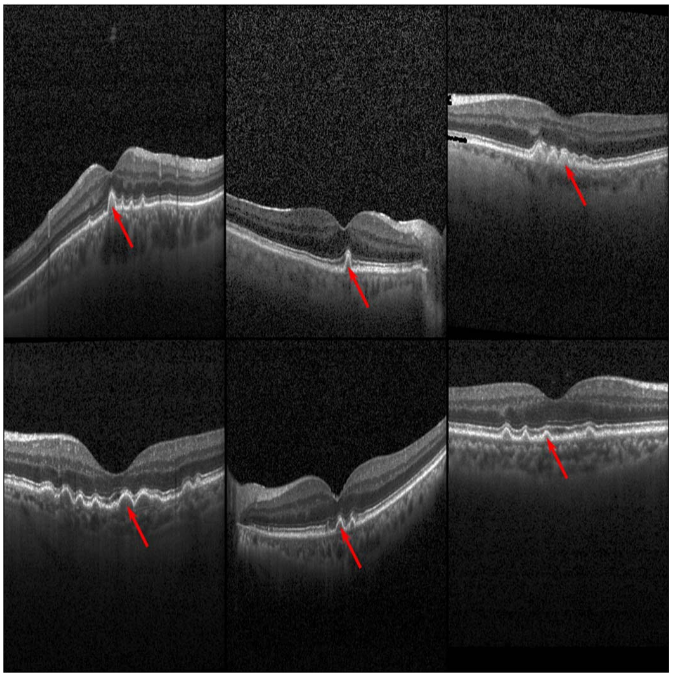


Fig. 1. Samples of drusen in OCT images. The red arrows indicate the drusens.

manual labels. By combining generalized low rank approximation of matrices with supervised manifold regularization to learn new features from image patches sampled from retinal images, Ren *et al.* [26]. proposed a supervised feature learning method for drusen segmentation in fundus images. Although these methods have achieved good performance for drusen segmentation in fundus images, the further accurate assessment of drusen based on fundus images still faces great challenges as the interference of other retinal diseases with similar pathological features, such as hard exudation [24], and the fundus image only captures the projection area information of drusen, lacking depth and spatial information [27]. Optical coherence tomography (OCT) can provide clear cross-sectional imaging of the retinal structure and has the ability to quantitatively evaluate retinal changes [28], [29]. Therefore, it has also become a promising tool for evaluating drusen [20]. Chen *et al.* [20]. exploited an automated drusen segmentation method in SD-OCT images by leveraging a priori knowledge of normal retinal morphology and anatomical features. And then, they further proposed a novel false color fusion strategy for drusen and geographic atrophy (GA) visualization in OCT images [27]. In addition, two previous CNN-based works for drusen segmentation in OCT images were proposed by Asgari *et al.* [30], [31]. Although these works have achieved promising performance in the task of evaluating drusen, the drusen segmentation in retinal OCT images is still very challenging due to the variations in size and shape of drusen, blurred boundaries, background noise interference and low contrast of OCT imaging. Fig.1 shows samples of drusen in retinal OCT images. Moreover, the lack of OCT dataset with pixel-level annotation is also a vital factor hindering the exploration of CNN-based methods for drusen segmentation.

To tackle these problems and improve the drusen segmentation accuracy, we propose a novel multi-scale transformer global attention network (MsTGANet), which integrates our newly proposed multi-scale transformer non-local module (MsTNL) and multi-semantic global channel and spatial joint attention module (MsGCS). Our main contributions are summarized as follows:

1) A novel MsTNL module is proposed and embedded into the top of encoder path to capture multi-scale non-local features with long-range dependencies from different layers in encoder.

2) To improve the model's ability to learn multi-semantic global contextual features, a novel MsGCS module is proposed and inserted between encoder and decoder.

3) By combining MsTNL and MsGCS module, a novel network named as MsTGANet is designed based on U-Shape architecture and applied to the task of drusen segmentation in OCT images.

4) We further propose a novel semi-supervised version of MsTGANet (Semi-MsTGANet) based on pseudo-labeled data augmentation strategy, which can leverage a large amount of unlabeled data to further improve the segmentation accuracy. And we also conduct extensive experiments to evaluate the effectiveness of the proposed MsTGANet and Semi-MsTGANet. The experimental results show that, compared with other state-of-the-art CNN-based methods, the segmentation accuracy of the proposed MsTGANet and Semi-MsTGANet are both improved significantly.

II. Related Works

A. Convolutional Neural Network

Recently, many segmentation networks based on fully convolutional network (FCN) [32] were proposed for semantic segmentation tasks [33]–[35]. Among them, U-Net with the encoder-decoder architecture has achieved remarkable performance in biomedical image segmentation [33]. In U-Net, the encoder to capture different level semantic features gradually by stacking convolutional layers and down-sampling operations, while a decoder with up-sampling layers is designed to recover the spatial information from the output of encoder stage by stage. Besides, to further improve the performance, skip connection is also added at each layer between the encoder and corresponding decoder to compensate the fine information loss caused by down-sampling. Although U-Net has achieved excellent performance in biomedical image segmentation, the simple skip-connection in original U-Net ignores global information and may introduce interference from local unrelated features, which causes U-Net to perform poor in some segmentation tasks with complex features. Recently, many promising studies have been proposed to try to solve these problems. Wu *et al.* [34]. proposed FastFCN by introducing a joint pyramid up-sampling module to replace dilated convolutions and capture global context information. Oktay *et al.* [35]. adopted a novel attention gate (AG) module to highlight salient features from skip connection in Attention U-Net. In addition, there are also methods aiming to improve the model's performance by exploring and integrating multi-scale contextual information in network design. PsPNet [36] was proposed to improve the model's ability of the multi-scale features capture in high-level feature maps by introducing multiple parallel pooling operations with different kernel sizes. DeepLabV3 [37] adopted multiple convolution branches with different receptive fields to improve the model's capacity to obtain multi-scale information. Based on both above advantages, Gu *et al.* [38]. proposed CE-Net by integrating two newly proposed dense atrous convolution (DAC) block and residual multi-kernel pooling (RMP) block, which has achieved promising performance in 2D medical image segmentation tasks. In recent years, many excellent methods based on attention mechanism also have been proposed for improving segmentation performance. Fu *et al.* [39]. proposed a dual attention network (DANet) to adaptively integrate local features and global dependencies. In our previous works, we proposed CPFNet [40] by designing two pyramidal modules to fuse global/multi-scale context information, which has also obtained very competitive performance in many medical image segmentation tasks. CPFNet [40] explores global/multi-scale contextual information mainly based on the soft spatial attention mechanism, which adopts the dilated convolution with shared weights to learn the feature information under different receptive fields and fuse multi-scale context information to improve the learning ability of global features. However, such multi-scale global feature extraction method still learns features in multi-size receptive fields, which cannot capture the long-distance feature correlation in the entire feature map.

B. Transformer

The transformer was originally proposed by Vaswani *et al.* [41]. to obtain the long-term dependence of timing signals in natural language processing (NLP) tasks. And many variant models based on transformer have achieved excellent results in machine translation and NLP [42]–[44]. Recently, many studies attempted to apply the transformer method in the field of computer vision to strengthen the network's ability of long-range dependencies capture in feature maps [45]–[48]. Wang *et al.* [47]. proposed a non-local network by appending the self-attention module based on transformer style on the top of backbone network to capture the non-local features with strong semantic information. Zhang *et al.* [48]. proposed a novel local relation network (LRNet) to provide greater modeling capacity than regular convolution in a more efficient manner. In addition, Dosovitskiy *et al.* [49]. proposed a novel ViT network, which can obtain results comparable to the current optimal convolutional network, with the great reduction of the computational resources required for its training. Srinivas *et al.* [50]. developed a novel method for visual recognition by combining regular CNN with transformer and achieved better performance than pure CNN based method, such as ResNet [51] and EfficientNet [52].

Although these transformer and CNN-based methods have achieved promising performance in many image processing tasks, there are still two problems that need to be resolved when these methods are applied to drusen segmentation in retinal OCT images: 1) How to improve the network's ability to capture multi-scale non-local features so as to deal with the complicated pathological manifestations of drusen in OCT images, especially in terms of size and shape. 2) How to improve the network's ability to learn multi-semantic global contextual features while suppressing noise interference, so as to solve the low contrast of OCT image imaging and the noise interference introduced by inherent technology. Therefore, to solve these problems and improve the accuracy of drusen segmentation, we propose a novel MsTGANet by combining two newly proposed modules of MsTNL and MsGCS. The MsTNL module can guide the model to capture multi-scale non-local features with long-term dependency information from different layers of the encoder. Furthermore, the MsGCS module between encoder and decoder can guide the model to extract multi-semantic global features, improving the model's capacity to learn salient features while suppressing the interference of non-related local features. By test, our proposed MsTGANet achieves higher accuracy for drusen segmentation in OCT images.

III. Methods

In this section, we first provide the structure details and core components of the proposed MsTGANet. Then, we present the loss function for optimizing the model. Finally, we introduce the semi-supervised version of MsTGANet based on pseudo-labeled data augmentation strategy, named as Semi-MsTGANet, which can further improve the segmentation accuracy of drusen in retinal OCT images by using a large amount of unlabeled data.

A. MsTGANet

1) Overview: Fig.2 shows the proposed MsTGANet, which adopts the encoder-decoder architecture as the basic framework and mainly consists four parts including encoder path, MsTNL module, MsGCS module and decoder path. It can be seen from Fig.2 that the MsTNL is inserted into the top of encoder path, in which the feature maps with different scale information from different layers of encoder are employed as MsTNL's input to capture multi-scale non-local features with long-range dependency. The MsGCS module is adopted to replace the skip-connection between encoder and decoder, which aims to guide the model to fuse multi-semantic global contextual features, so as to improve the model's ability to learn global salient features while suppressing the interference of non-related local features.

2) Encoder: It can be seen from Fig. 2 that, same as the U-Net[28], the encoder of proposed MsTGANet mainly contains five blocks, where each block consists of a MaxPool operation followed by two convolutional layers except for the first block with only two convolutional layers. The MaxPool operation is adopted to down-sample the feature maps and extend the receptive fields, while the convolutional layers are employed to extract the features in different stages.

3) MsTNL Module: As shown in Fig. 1, the complicated pathological manifestations of drusen in OCT images, especially in terms of size and shape, poses a great challenge to accurately segment drusen regions. Besides, there are also other interferences surrounding drusen, such as other lesions

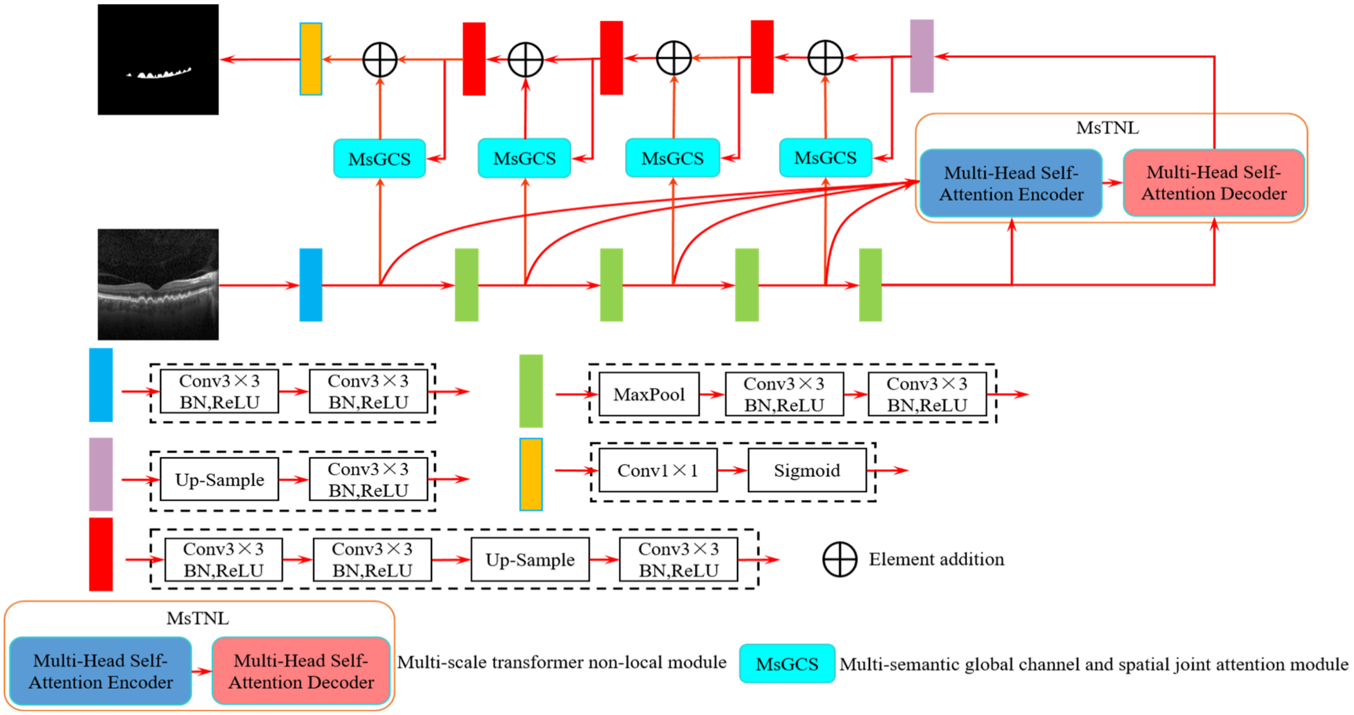

MsTNL
Multi-Head Self-Attention Encoder
Multi-Head Self-Attention Decoder
MsGCS
Conv3×3 BN,ReLU
MaxPool
Up-Sample
Conv1×1
Sigmoid
Element addition
Multi-scale transformer non-local module
Multi-semantic global channel and spatial joint attention module

Fig. 2. The architecture of MsTGANet.

and speckle noise. Therefore, improving the network's ability to learn multi-scale non-local features is essential for improving the accuracy of drusen segmentation. Many previous studies have explored the non-local features to improve the performance of image analysis tasks and have achieved excellent results [53], [54]. However, these non-local spatial interaction approaches are not cross-scale and have no location information. To this end, inspired by attention-based methods of [40] and [54], a novel multi-scale transformer non-local module (MsTNL) is proposed and appended on the top layer of encoder path to capture multi-scale non-local information with long-range dependency from different layers of encoder. Fig. 3 illustrates the details of MsTNL. As seen from Fig.3, our proposed MsTNL mainly consists of two blocks: multi-head self-attention encoder and multi-head self-attention decoder.

a) Multi-head self-attention encoder: As shown in Fig. 2 and Fig. 3, feature maps from stage1 ($\boldsymbol{F_1}$), stage2 ($\boldsymbol{F_2}$), stage3 ($\boldsymbol{F_3}$), stage4 ($\boldsymbol{F_4}$) and top layer ($\boldsymbol{F_T}$) are adopted as the input of MsTNL. First, the feature maps of $\boldsymbol{F_1}$, $\boldsymbol{F_2}$, $\boldsymbol{F_3}$, and $\boldsymbol{F_4}$ are fed into a block of down-sampling followed by a conv3 × 3 operation, respectively, to obtain the feature maps that match $\boldsymbol{F_T}$'s size and channel numbers. And then, these down-sampled features are fused by the element-level addition to get the feature maps with multi-scale information $\boldsymbol{F_A}$. To obtain the multi-scale non-local features with long dependency information, the multi-scale feature map $\boldsymbol{F_A}$ and the top layer's feature map $\boldsymbol{F_T}$ are fed into a multi-head self-attention module, which mainly contains three branches: Query, Key and Value [44]. Different from self-attention non-local module in previous studies[44], the multi-head self-attention module adopts the multi-scale features $\boldsymbol{F_A}$ as the input of Query branch, while the branches of Key and Value employs the feature maps $\boldsymbol{F_T}$ with strong semantic global information as the input. The multi-head self-attention encoder module is mainly adopted to extract the multi-scale non-local features with long dependency in $\boldsymbol{F_A}$ based on the guidance of the feature maps $\boldsymbol{F_T}$ with strong semantic global information. It can be seen from Fig.3 (a) that the proposed multi-scale self-attention module mainly consists of five steps:

(1) The convolution operations with 1 × 1 kernel size are adopted as the weights of branch Query, Key and Value to encode the feature maps $\boldsymbol{F_A}$ to $\boldsymbol{Q}$ and $\boldsymbol{F_T}$ to $\boldsymbol{K}$ and $\boldsymbol{V}$, respectively. And, the channel of $\boldsymbol{Q}$ and $\boldsymbol{K}$ also be squeezed to $1/8C$ to squeeze the channel features and improve the efficiency of the network.

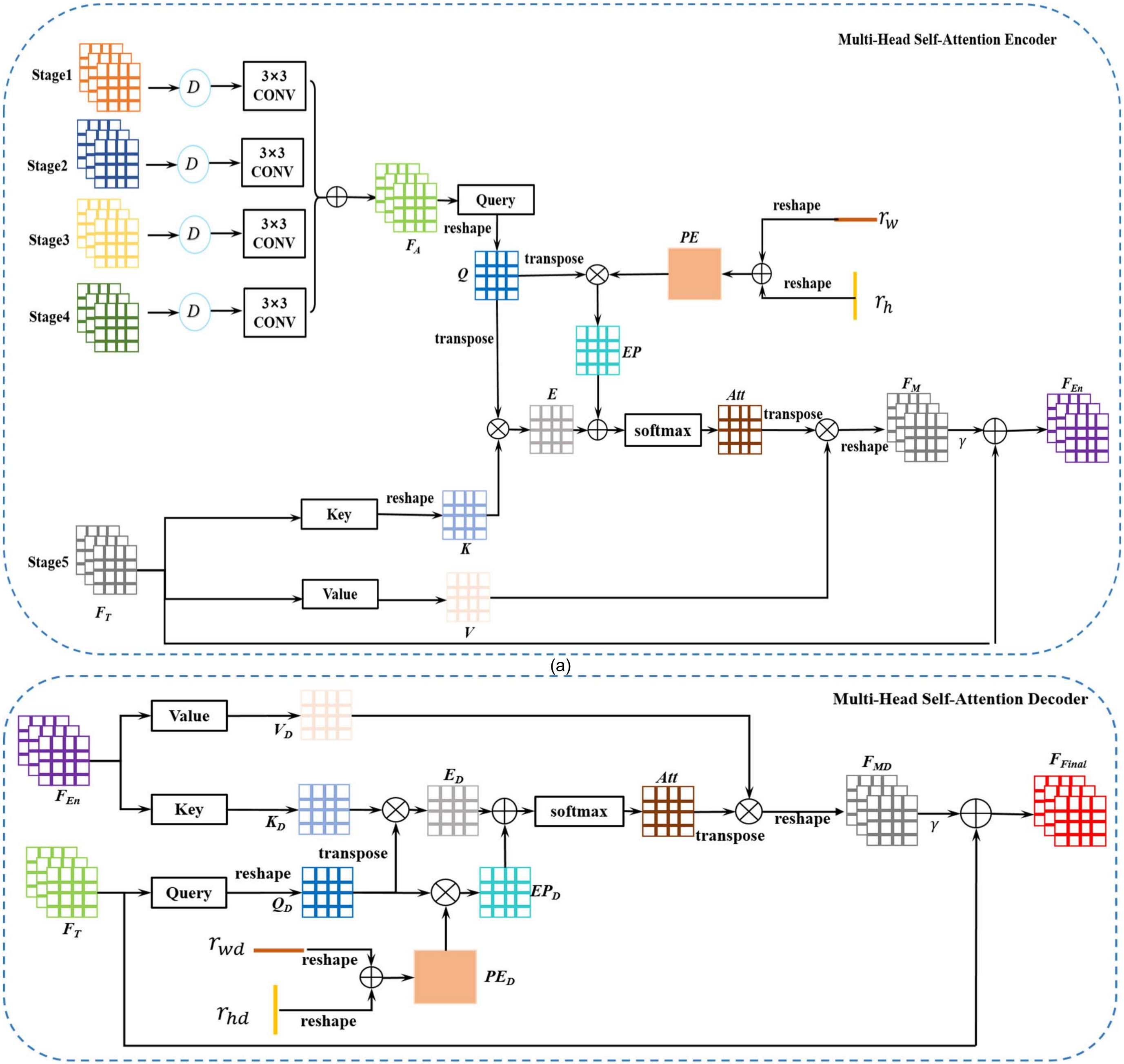
$$Q = \text{Conv}1 \times 1\,(F_A) \in R^{B,C/8,H,W} \quad (1)$$
$$K = \text{Conv}1 \times 1\,(F_T) \in R^{B,C/8,H,W} \quad (2)$$
$$V = \text{Conv}1 \times 1\,(F_T) \in R^{B,C,H,W} \quad (3)$$

where B, C, H and W represent the batch size, channel, height and width, respectively.

(2) To make the self-attention operation sensitive to position for features, position coding has been used in the model design based on the transformer architecture [41], which allows the network to focus on the long-term dependence on the feature position. Therefore, in this paper, we adopt learnable parameters to encode the feature position from the vertical and horizontal directions, respectively.

$$PE = \text{Reshape}\,(r_h) + \text{Reshape}\,(r_w) \in R^{B,C/8,H,W} \quad (4)$$

where $\mathrm{r}_h \in R^{B,C/8,H,1}$ and $\mathrm{r}_w \in R^{B,C/8,1,W}$ represent the encoding learnable vectors from the vertical and horizontal directions, respectively.



(a)

(b)

Fig. 3. Details of MsTNL. (a) Multi-head self-attention encoder (b) Multi-head self-attention decoder.

(3) Obtaining the attention map: The similarity matrix $\boldsymbol{E}$ between $\boldsymbol{Q}$ and $\boldsymbol{K}$ is calculated first to obtain the multi-scale non-local spatial correlation weight with the guidance of strong global semantic information. Then, the positional correlation from vertical and horizontal directions of the features in $\boldsymbol{Q}$, named as $\boldsymbol{EP}$, is encoded by matrix multiplication between $\boldsymbol{PE}$ and $\boldsymbol{Q}$. Finally, the attention map $\boldsymbol{Att}$ is obtained by adding $\boldsymbol{E}$ and $\boldsymbol{EP}$ followed by Softmax.

$$Q = \text{Reshape}\,(Q) \in R^{B,C/8,H\times W} \quad (5)$$

$$K = \text{Reshape}\,(K) \in R^{B,C/8,H\times W} \quad (6)$$

$$E = Q^T \circ K \in R^{B,H\times W,H\times W} \quad (7)$$

$$EP = Q^T \circ PE \in R^{B,H\times W,H\times W} \quad (8)$$

$$Att = \text{Soft}\max\,(E + EP) \in R^{B,H\times W,H\times W} \quad (9)$$

where Q^T is the transpose of Q, and $^\circ$ is matrix multiplication operation.

(4) The attention map $\boldsymbol{Att}$ and the corresponding $\boldsymbol{V}$ are weighted and summed to obtain the multi-scale non-local spatial response $\boldsymbol{F_M}$ with strong global semantics.

$$F_M = \text{Reshape}\left(V \circ Att^T\right) \in R^{B,C,W,H} \quad (10)$$

(5) Finally, the multi-scale non-local feature map $\boldsymbol{F_{En}}$ with long dependency is obtained by element-level addition between the $\boldsymbol{F_T}$ and the weighted $\boldsymbol{F_M}$ as follows:

$$F_{En} = F_T + \gamma F_M \in R^{B,C,W,H} \quad (11)$$

where γ is a learnable parameter that is initialized as 0, and gradually adjusted to assign the weight for $\boldsymbol{F_M}$ in a learnable

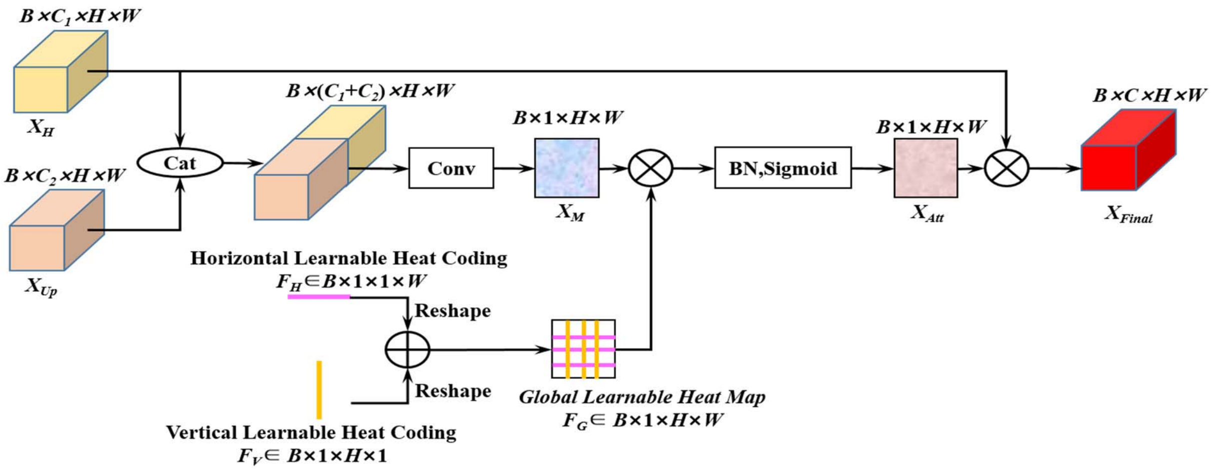


Fig. 4. The structure of multi-semantic global channel and spatial joint attention module (MsGCS).

way in the training process. As shown in Eq. 11, $\boldsymbol{F_{En}}$ is the weighted sum of multi-scale non-local features and strong global semantic features.

b) Multi-head self-attention decoder: Contrary to multi-head self-attention encoder, the multi-head self-attention decoder is mainly used to further extract the strong semantic position self-correlation information contained in the top-level feature map $\boldsymbol{F_T}$ based on the guidance of multi-scale non-local features $\boldsymbol{F_{En}}$. Similar to multi-head self-attention encoder, the proposed multi-head self-attention decoder also mainly consists of five steps, as shown in Fig.3 (b):

(1) The convolution operations with 1 × 1 kernel size are adopted as the weights of branch Query, Key and Value to encode the feature maps $\boldsymbol{F_T}$ to $\boldsymbol{Q_D}$ and $\boldsymbol{F_{En}}$ to $\boldsymbol{K_D}$ and $\boldsymbol{V_D}$, respectively.

$$Q_D = \text{Conv}1 \times 1\,(F_T) \in R^{B,C/8,H,W} \tag{12}$$

$$K_D = \text{Conv}1 \times 1\,(F_{En}) \in R^{B,C/8,H,W} \tag{13}$$

$$V_D = \text{Conv}1 \times 1\,(F_{En}) \in R^{B,C,H,W} \tag{14}$$

(2) Position encoding for $\boldsymbol{Q_D}$, which aims to encode the strong semantic feature position in $\boldsymbol{Q_D}$ from the vertical and horizontal directions, respectively.

$$PE_D = \text{Reshape}\,(r_{hd}) + \text{Reshape}\,(r_{wd}) \in R^{B,C/8,H,W} \tag{15}$$

where $r_{hd} \in R^{B,C/8,H,1}$ and $r_{wd} \in R^{B,C/8,1,W}$ represent the encoding learnable vectors from the vertical and horizontal directions, respectively.

(3) Obtaining the attention map: Like multi-head self-attention encoder, it also takes 5 steps to obtain the attention map, as follows:

$$Q_D = \text{Reshape}\,(Q_D) \in R^{B,C/8,W\times H} \tag{16}$$

$$K_D = \text{Reshape}\,(K_D) \in R^{B,C/8,W\times H} \tag{17}$$

$$E_D = Q_D^T \circ K_D \in R^{B,W\times H,W\times H} \tag{18}$$

$$EP_D = Q_D^T \circ PE_D \in R^{B,H\times W,H\times W} \tag{19}$$

$$Att_D = \text{Soft}\max\,(E_D + EP_D) \in R^{B,H\times W,H\times W} \tag{20}$$

(4) The attention map $\boldsymbol{Att}$ and the corresponding $\boldsymbol{V_D}$ are weighted and summed to obtain the spatial response $\boldsymbol{F_{MD}}$ with strong semantic positional long dependency information.

$$F_{MD} = \text{Reshape}\left(V_D \circ Att_D^T\right) \in R^{B,C,W,H} \tag{21}$$

(5) Finally, the final multi-scale strong semantic non-local feature map $\boldsymbol{F_{Final}}$ with long dependency is obtained by element-level addition between $\boldsymbol{F_T}$ and the weighted $\boldsymbol{F_{MD}}$, as follows:

$$F_{Final} = F_T + \gamma_D F_{MD} \in R^{B,C,W,H} \tag{22}$$

where γ_D is a learnable parameter that is initialized as 0, and gradually adjusted to assign the weight for $\boldsymbol{F_{MD}}$ in a learnable way during training.

4) MsGCS Module:

Although U-Net and its variants have achieved excellent performance in medical image segmentation tasks, the simple skip-connection between encoder and decoder ignores global information and may introduce interference from local unrelated features [40]. To solve this problem and further improve the drusen segmentation accuracy, we propose a novel multi-semantic global channel and spatial joint attention module(MsGCS) to replace the simple skip-connection in original U-Net, which aims to guide the model to learn multi-semantic global contextual features in channel and spatial dimensions. Fig. 4 shows the detailed structure of MsGCS module.

As shown in Fig. 4, the skip-connection feature map $\boldsymbol{X}$ with high resolution weak semantic features and the up-sampled $\boldsymbol{X_{up}}$ with low resolution strong semantic features are first fused by concatenation. And, the fused map is fed into the convolution layer for channel normalization to obtain the multi-semantic global response map $\boldsymbol{X_M}$ in channel dimension.

$$X_M = \text{Conv}\left(\text{Cat}\left(X_H, X_{up}\right)\right) \in R^{B,1,W,H} \tag{23}$$

Then, to further adaptively capture the multi-semantic global response in spatial dimension, a novel global learnable weight matrix is developed to multiply with $\boldsymbol{X_M}$ followed by batch normalization (BN) and sigmoid activation. As shown in Fig.4, the global learnable heat map $\boldsymbol{F_G}$ and multi-semantic global feature attention map $\boldsymbol{X_{att}}$ are obtained as follows:

$$F_G = \text{Reshape}\left(F_H\right) + \text{Reshape}\left(F_V\right) \in R^{B,1,W,H} \quad (24)$$

$$X_{Att} = \text{Sigmoid}\left(\text{BN}\left(F_G * X_M\right)\right) \in R^{B,1,W,H} \quad (25)$$

where $\boldsymbol{F_H} \in R^{B,1,H,1}$ and $\boldsymbol{F_V} \in R^{B,1,1,W}$ are the spatial feature positional correlation vectors in the horizontal and vertical directions, with random initial value following the standard normal distribution.

Finally, the final feature map $\boldsymbol{X_{Final}}$ is obtained by multiplying $\boldsymbol{X_H}$ with $\boldsymbol{X_{att}}$:

$$X_{Final} = X_H * X_{Att} \in R^{B,C_1,W,H} \quad (26)$$

It can be seen from Eq.26 that $\boldsymbol{X_{Final}}$ is the result of $\boldsymbol{X_H}$ weighted by $\boldsymbol{X_{att}}$, which can guide the model to learn multi-semantic global salient features in the skip-connection feature map $\boldsymbol{X_H}$ while suppressing the interference of non-related local information, thereby improving the model's segmentation performance.

5) Decoder: As shown in Fig.2, the decoder path mainly contains three components, including up-sampling layer, feature fusion operation and convolutional blocks that consist of two convolutional layers. The decoder is mainly adopted to restore the spatial information with strong multi-scale semantic features generated by MsTNL, and gradually fuse the multi-semantic global contextual information from MsGCS module via two convolutional layers.

B. Loss Function

To optimize the proposed model, a joint loss function L_{joint} including Dice loss L_{Dice} and binary cross entropy loss L_{BCE} is adopted to guide the training of the model. The joint loss function is calculated as follows:

$$L_{Joint} = L_{Dice} + L_{BCE} \quad (27)$$

$$L_{Dice} = 1 - \frac{2\left|X * Y\right|}{|X| + |Y|} \quad (28)$$

$$L_{BCE} = -\sum\nolimits_{h,w} (1 - Y) \log (1 - X) + Y \log (X) \quad (29)$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the segmentation results and the corresponding ground truth, h and w represent the coordinates of the pixel in $\boldsymbol{X}$ and $\boldsymbol{Y}$. As shown in Eq. 27, the Dice loss function is mainly used to optimize the model in image level, while the binary cross entropy loss function is employed to optimize the model in pixel level. In addition, to ensure fairness, all the comparison methods involved in this paper adopts the same loss function to optimize their model during training.

C. Semi-Supervised MsTGANet Framework

The lack of OCT dataset with pixel-level annotation is also one of the vital factors hindering the improvement of drusen segmentation accuracy. It is also very difficult and time-consuming to obtain these pixel-level annotations. To resolve this issue, we proposed a semi-supervised MsTGANet framework based on pseudo-labeled data augmentation strategy to further improve the segmentation accuracy of MsTGANet. The overall architecture of the proposed Semi-MsTGANet is shown in Fig.5. As shown in Fig. 5, the semi-supervised MsTGANet framework mainly consists of three steps:

TABLE I
SUPERVISED AND SEMI-SUPERVISED DATA STRATEGIES

	Supervised	Semi-Supervised
Training	729 retinal OCT images with ground truth from three folds.	729 retinal OCT images with ground truth from three folds+7664 retinal OCT images with pseudo labels.
Testing	243 retinal OCT images with ground truth from one fold	243 retinal OCT images with ground truth from one fold

1) The MsTGANet is first trained based on the labeled data under the guidance of the fully-supervised objective function (Eq. 27).

2) The above pre-trained MsTGANet is adopted to segment drusens in large amount of unlabeled data, and the segmentation results are employed as the pseudo labels corresponding to the unlabeled data.

3) The large amount of unlabeled data with pseudo label is mixed with labeled data to retrain MsTGANet based on the mixed-supervised loss function as follows:

$$L_{mixed} = L_{Joint}\left(X_L, Y_L\right) + L_{Joint}\left(X_U, Y_{Pseudo}\right) \quad (30)$$

where X_L and Y_L are the original data and the corresponding label, respectively. X_U and Y_{Pseudo} denote the original unlabled data and the corresponding pseudo label generated by pre-trained MsTGANet, respectively.

The results and effectiveness of semi-supervised training strategy will be discussed in Section IV in detail.

IV. EXPERIMENTS AND RESULTS

A. Dataset and Implementation Details

1) Dataset: The 8616 retinal OCT B-scans used in this paper are collected from the UCSD public dataset [62], which were established by the Shiley Eye Institute of the University of California San Diego (UCSD) and all of the images (Spectralis OCT, Heidelberg Engineering, Germany) were selected from retrospective cohorts of adult patients without exclusion criteria based on age, gender, or race. Haoyu Chen, ophthalmologist of the Joint Shantou International Eye Center (JSIEC) of Shantou University and the Chinese University of Hong Kong, guided the pixel-level annotation of 972 OCT images in the UCSD dataset. To comprehensively evaluate the performance of the proposed MsTGANet and Semi-MsTGANet, 4-fold cross-validation strategy was applied to 972 labeled images in all experiments. Meanwhile, the remaining 7644 OCT images will be employed as unlabeled data to participate the training of Semi-MsTGANet. The details for data strategies are listed in TABLE I.

2) Implementation Details: Both proposed MsTGANet and Semi-MsTGANet were performed on the public platform pytorch and RTX3090 GPU (24GB). The Adam was used

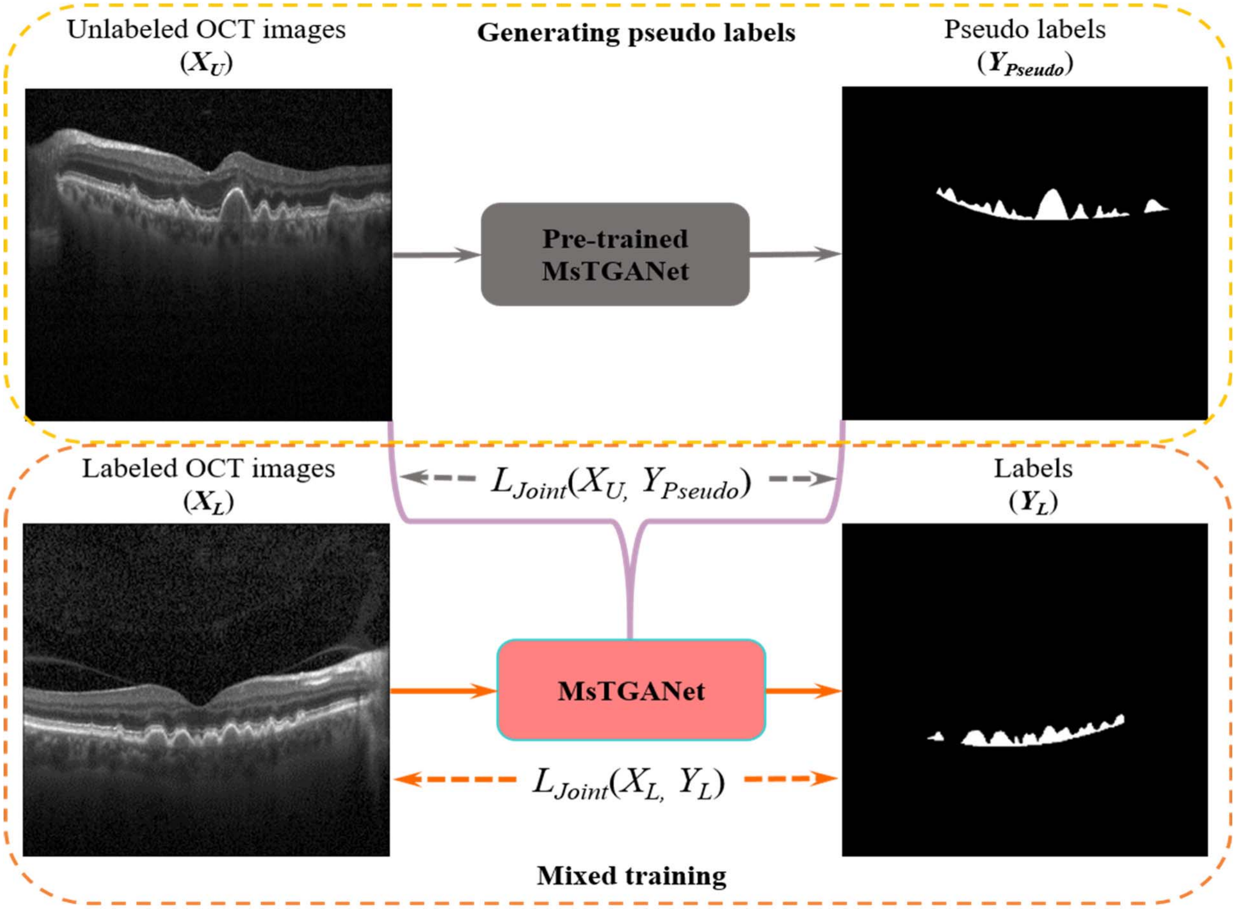


Fig. 5. The overall architecture of the proposed semi-supervised version of MsTGANet (Semi-MsTGANet).

as the optimizer. Initial learning rate and weight decay were set to 0.0005 and 0.0001, respectively. The batch size was set to 4, and the number of epochs was 100. In addition, to facilitate training and prevent loss of detailed information, the size of images was rescaled to 512 × 512. To validate the performance of the proposed MsTGANet and Semi-MsTGANet, we compared their segmentation results with other excellent networks such as FCN [32], U-Net [33], FastFCN [34], Attention U-Net(Att-UNet) [35], PsPNet [36], DeepLabV3 [37], CE-Net [38], DANet [39], CPFNet [40], GCN [56], R2UNet [57], UNet++ [58], and HRSegNet [59]. To ensure fairness, all the networks were trained with the same configuration and loss function. The code of the proposed MsTGANet and Semi-MsTGANet will be released in: https://github.com/wangmeng9218/Semi-MsTGANet.

B. Evaluation Metrics

To comprehensively and fairly evaluate the segmentation performance of different methods, we adopt four indicators to quantitatively analyze the experimental results, including Jaccard index (Jac), Dice similarity coefficient (DSC), precision (Pre) and Pearson product-moment correlation coefficient (Ppmcc), among which Jac and DSC are the most commonly used indices in validating the performance of segmentation algorithms [32], [36], [37], [39]. The formulas of the four evaluation metrics are as follows:

$$Jac = \frac{TP}{TP + TN + FP} \tag{31}$$

$$DSC = \frac{2 * TP}{2 * TP + TN + FP} \tag{32}$$

$$Pre = \frac{TP}{TP + FP} \tag{33}$$

$$Ppmcc = \frac{Cov\,(X, Y)}{\sigma_X \sigma_Y} \tag{34}$$

where *TP*, *TN*, *FP* and *FN* are true positive, true negative, false positive and false negative for pixel classification, respectively. X and Y represent the segmentation result and corresponding ground truth. Cov(.) is the covariance between X and Y. σ_x and σ_y denote the standard deviation of X and Y, respectively.

C. Qualitative Analysis

Fig. 6 shows seven segmentation results with speckle noise interference and some variation in size or shape of the proposed Semi-MsTGANet, MsTGANet, and compared with two classical networks that are widely used in medical image segmentation tasks. As shown in Fig. 6, our proposed Semi-MsTGANet achieves better segmentation performance, especially for segmenting the small size drusen (Fig. 6(c)). Instead, CE-Net performs not well in segmenting small size drusen because it adopt ResNet-34 as the feature extractor.

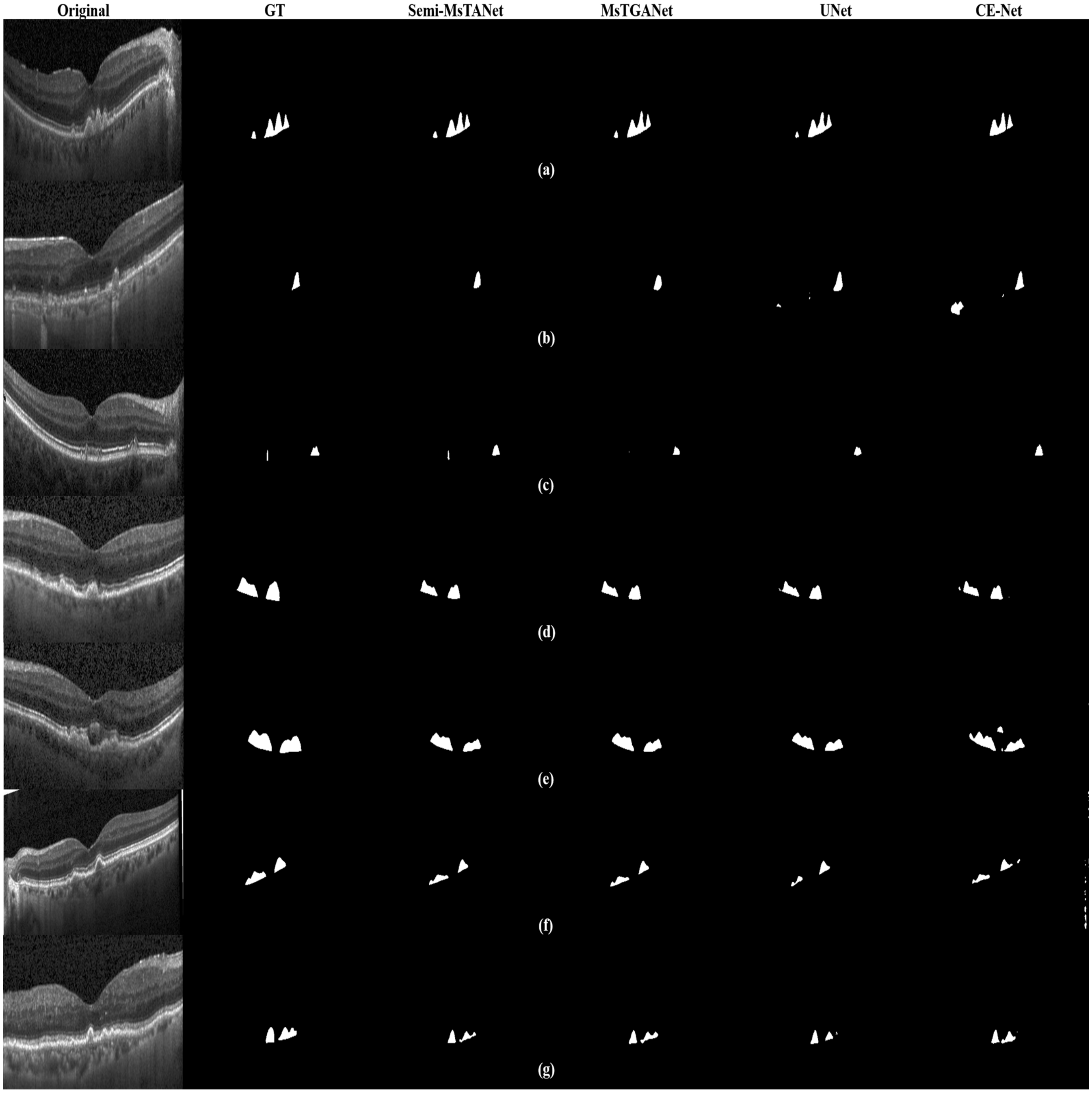


Fig. 6. Seven segmentation results with speckle noise interference and some variation in size or shape of Semi-MsTGANet, MsTGANet, and two classical networks.

The first layer of feature extractor in CE-Net is a convolutional layer with kernel size 7 × 7 and stride 2, which may cause the loss of small features for some small-sized lesion regions (Fig.6 (a) and Fig.6(c)). In U-Net and CE-Net, the feature maps from skip-connection are directly merged with the up-sampled feature map by simple concatenation. It is difficult to avoiding the interference of non-correlated local information and speckle noise, causing the false positives (Fig.6(b),(d),(e),(f)). Compared with UNet and CE-Net, the proposed MsTGANet and Semi-MsTGANet still achieves better segmentation performance under the influence of speckle noise interference and some variation in size or shape, which prove the effectiveness and robustness of our proposed methods.

D. Quantitative Evaluation

To quantitatively evaluate the retinal drusen segmentation performance, the mean and standard deviation values of four metrics including Jac, DSC, Pre, and Ppmcc for different methods are listed in TABLE II.

It can be seen from TABLE II that our proposed Semi-MsTGANet achieves better segmentation performance in

TABLE II
EVALUATION INDICES FOR DIFFERENT METHODS

Strategy	Network	Jac	DSC	Pre	Ppmcc	Time(s)
Supervised	FCN[32]	0.6779±0.0080	0.8067±0.0059	0.8067±0.0077	0.8069±0.0058	0.0177
	Att-UNet[35]	0.7362±0.0061	0.8469±0.0042	0.8366±0.0040	0.8473±0.0042	0.0150
	Fast-FCN[34]	0.6860±0.0068	0.8125±0.0048	0.8167±0.0080	0.8126±0.0048	0.0188
	GCN[56]	0.7062±0.0068	0.8267±0.0049	0.8323±0.0080	0.8268±0.0049	**0.0129**
	R2UNet[57]	0.7276±0.0068	0.8411±0.0048	0.8524±0.0150	0.8415±0.0048	0.0340
	PsPNet[36]	0.7232±0.0061	0.8383±0.0041	0.8449±0.0112	0.8386±0.0042	0.0236
	DeepLabV3[37]	0.7238±0.0116	0.8385±0.0081	0.8403±0.0144	0.8388±0.0080	0.0135
	DANet[39]	0.6956±0.0049	0.8194±0.0035	0.8175±0.0074	0.8195±0.0037	0.0294
	CE-Net[38]	0.7213±0.0063	0.8368±0.0045	0.8410±0.0144	0.8372±0.0045	0.0481
	UNet++[58]	0.7327±0.0074	0.8446±0.0052	0.8448±0.0107	0.8448±0.0052	0.0310
	HRSegNet[59]	0.7198±0.0043	0.8360±0.0031	0.8349±0.0110	0.8363±0.0031	0.0133
	CPFNet[40]	0.7251±0.0056	0.8395±0.0039	0.8408±0.0068	0.8398±0.0037	0.0159
	UNet[33]	0.7325±0.0081	0.8445±0.0056	0.8412±0.0125	0.8449±0.0055	0.0173
	UNet+MsTNL	0.7386±0.0060	0.8486±0.0041	0.8421±0.0106	0.8490±0.0040	0.0186
	UNet+MsGCS	0.7386±0.0076	0.8486±0.0054	0.8439±0.0052	0.8489±0.0053	0.0174
	MsTGANet	0.7412±0.0054	0.8502±0.0039	0.8548±0.0060	0.8504±0.0039	0.0187
Semi-Supervised	SemiGAN[60]	0.7249±0.0037	0.8398±0.0026	0.8476±0.0121	0.8397±0.0026	0.0194
	CycleGAN[61]	0.7322±0.0072	0.8448±0.0049	0.8409±0.0198	0.8448±0.0048	0.0151
	Semi-MsTGANet	**0.7597±0.0082**	**0.8629±0.0052**	**0.8633±0.0051**	**0.8626±0.0052**	0.0187

all indices. As shown in TABLE II, FCN achieves the worst results since FCN segments the target based on the features from the top layer of VGG [55], which may cause the loss of feature information of some small size drusens. Fast-FCN also achieves the bad results, especially for Jac and DSC (0.6860 and 0.8125, respectively). Compared with FCN, Fast-FCN segments the target region based on the features of top-three layers in the VGG [55] combined with joint pyramid up-sampling module. Although Fast-FCN has achieved excellent results in many segmentation tasks, it achieves low segmentation accuracy for segmenting drusen with complex pathological features because of the loss of small target features. Compared with Fast-FCN, DANet uses the dilated ResNet as its feature extractor, and introduces a dual attention module to learn the non-local features in the channel and spatial dimensions, which improves the segmentation accuracy. However, the loss of small-size features still exists, which also leads to its low performance for segmenting drusen. Compared with Fast-FCN and DANet, both PsPNet and DeepLabV3 improve the segmentation performance by introducing feature pyramid module to capture the multi-scale features, which also proves that multi-scale features are beneficial for improving segmentation performance. Contrary to FCN series, the U-shape networks series, such as U-Net, Att-UNet, R2UNet, GCN, CE-Net, CPFNet, and U-Net++ restores the resolution of the high-level features layer by layer through up-sampling operations in the decoder path, and add skip-connection at each layer between the encoder and the corresponding decoder to further alleviate the problem of details loss for small target. As shown in TABLE II, most U-shape networks achieves good results, especially U-Net and Att-UNet. Compared with U-Net, Att-UNet improves segmentation accuracy by introducing an attention gate module and embedding it at the end of each skip-connection to guide the network to focus on salient features. It can be seen from TABLE II that the proposed MsTGANet achieves better performance than other CNN-based methods, the average values of Jac, DSC, Pre, and Ppmcc reaches 0.7412, 0.8502, 0.8548, and 0.8504, respectively. In particular, compared with Att-UNet, which has the best performance among all comparison methods, all indices of MsTGANet have been improved, especially Jac and Pre increased by 0.68% and 2.18%, respectively. In addition, the average Jac, DSC, Pre and Ppmcc of the proposed MsTGANet are 2.40%, 1.40%, 1.73% and 1.38% higher than DeepLabV3, which has been widely used in many target segmentation tasks [37]. The experimental results prove the effectiveness of the proposed MsTGANet for drusen segmentation in OCT images.

As shown in TABLE II, the proposed Semi-MsTGANet achieves best performance in all evaluation metrics, with the average values of Jac, DSC, Pre and Ppmcc reaching 0.7597, 0.8629, 0.8633 and 0.8626, respectively. Compared with MsTGANet, the Jac, DSC, Pre and Ppmcc of Semi-MsTGANet have been improved significantly by 2.50%, 1.49%, 0.99% and 1.43%, respectively. And compared with the baseline network U-Net, the indices of Jac, DSC, Pre and Ppmcc of Semi-MsTGANet have been improved significantly by 3.71%, 2.18%, 2.63% and 2.09%, respectively. The experimental results show that the proposed Semi-MsTGANet can further improve the drusen segmentation performance significantly by leveraging a large amount of unlabeled data. In addition, in order to evaluate the performance of different methods more

TABLE III
STATISTICAL ANALYSIS (p-VALUE) OF THE PROPOSED MSTGANET COMPARED WITH OTHER CNN-BASED METHODS

Network	Jac	DSC
MsTGANet-FCN[32]	<5E-4	<5E-4
MsTGANet-UNet[33]	0.026	0.026
MsTGANet-Att-UNet[35]	0.053	0.046
MsTGANet-Fast-FCN[34]	<5E-4	<5E-4
MsTGANet-GCN[56]	<5E-4	<5E-4
MsTGANet-R2UNet[57]	0.001	0.001
MsTGANet-PsPNet[36]	0.031	<5E-4
MsTGANet-DeepLabV3[37]	0.026	0.027
MsTGANet-DANet[39]	<5E-4	<5E-4
MsTGANet-CE-Net[38]	<5E-4	0.001
MsTGANet-UNet++[58]	0.012	0.013
MsTGANet-HRSegNet[59]	<5E-4	<5E-4
MsTGANet-CPFNet[40]	0.001	0.001
MsTGANet-SemiGAN[60]	0.007	0.005
MsTGANet-CycleGAN[61]	0.041	0.029

comprehensively, the efficiency of different methods also be listed in TABLE II. It can be seen from TABLE II that our proposed method takes slightly longer time than UNet due to the introduction of MsTNL and MsGCS in MsTGANet. However, it can still meet the requirement of real-time processing. These experimental results show that compared with other CNN-based methods, the proposed MsTGANet and Semi-MsTGANet can achieve better segmentation performance with similar efficiency.

In addition, the comparison between the proposed Semi-MsTGANet and two commonly used semi-supervised segmentation architectures including SemiGAN [60] and CycleGAN [61] is also listed in TABLE II. It can be seen from TABLE II that the proposed Semi-MsTGANet achieves better performance in all evaluation metrics which demonstrates the effectiveness of the proposed semi-supervised framework.

E. Statistical Significance Assessment

We further investigate the statistical significance of the performance improvement for the proposed MsTGANet and Semi-MsTGANet by the paired T test, and the p-values are listed in TABLE III and TABLE IV, respectively.

As can be seen from TABLE III that compared with other excellent CNN-based methods, with the exception of Jac compared with Att-UNet ($p = 0.053$ is slightly higher than 0.05), all the improvements for Jac and DSC of MsTGANet are statistically significant with p-values less than 0.05. The results further prove the effectiveness of the proposed MsTGANet.

TABLE IV lists the p-values of the proposed Semi-MsTGANet compared with MsTGANet and other CNN-based methods. All the improvements for Jac and DSC of Semi-MsTGANet are statistically significant with p-values less than 0.05. The results further demonstrate that the proposed semi-supervised framework can leverage unlabeled data to further improve the drusen performance significantly.

TABLE IV
STATISTICAL ANALYSIS (p-VALUE) OF THE PROPOSED SEMI-MSTGANET COMPARED WITH MSTGANET AND OTHER CNN-BASED METHODS

Network	Jac	DSC
Semi-MsTGANet-FCN[32]	0.001	0.001
Semi-MsTGANet-UNet[33]	0.009	0.009
Semi-MsTGANet-Att-UNet[35]	0.014	0.013
Semi-MsTGANet-Fast-FCN[34]	0.002	0.002
Semi-MsTGANet-GCN[56]	0.003	0.003
Semi-MsTGANet-R2UNet[57]	0.006	0.006
Semi-MsTGANet-PsPNet[36]	0.021	0.006
Semi-MsTGANet-DeepLabV3[37]	0.009	0.009
Semi-MsTGANet-DANet[39]	0.002	0.002
Semi-MsTGANet-CE-Net[38]	0.008	0.008
Semi-MsTGANet-UNet++[58]	0.014	0.013
Semi-MsTGANet-HRSegNet[59]	0.004	0.004
Semi-MsTGANet-CPFNet[40]	0.004	0.004
Semi-MsTGANet-SemiGAN[60]	0.010	0.009
Semi-MsTGANet-CycleGAN[61]	0.006	0.006
Semi-MsTGANet-MsTGANet	0.033	0.030

F. Ablation Experiments

As shown in TABLE II, ablation experiments are also conducted to validate the performance of the proposed MsTNL module and MsGCS module. In this paper, U-Net is adopted as our baseline model to evaluate the effectiveness of MsTNL and MsGCS.

1) Ablation Experiment for MsTNL: As can be seen from TABLE II, compared with U-Net, the Jac, DSC, Pre, Ppmcc of UNet+MsTNL have been improved from 0.7325, 0.8445, 0.8412, and 0.8449 to 0.7386, 0.8486, 0.8421, and 0.8490 respectively, which benefits from the fact that the proposed MsTNL module can adaptively guide model to learn multi-scale non-local features with long dependency information. In addition, we also conducted experiments to compare the performance of $\boldsymbol{Q}$ and $\boldsymbol{K}$ with C and $1/16C$ channels, respectively. The results are listed in TABLE V. It can be seen from TABLE V that, compared with UNet+MsTNL(C) and UNet+MsTNL($1/16C$), our proposed UNet+MsTNL($1/8C$) achieves better performance in indices of Jac, Dsc, and Ppmcc with similar efficiency, which prove the reasonability of the proposed MsTNL module with $1/8C$.

2) Ablation Experiment for MsGCS: As shown in TABLE II, the embedding of MsGCS module into U-Net (UNet+MsGCS) also obtained better segmentation performance. Compared with U-Net, UNet+MsGCS achieves higher indices in the four evaluation indicators. The Jac, DSC, Pre, and Ppmcc of UNet+MsGCS have been improved from 0.7325, 0.8445, 0.8412, and 0.8449 to 0.7386, 0.8486, 0.8439, and 0.8489, respectively. The results show that the proposed MsGCS module is beneficial to improve the performance of model.

Finally, as shown in TABLE II that compared with U-Net, the segmentation performance of MsTGANet (UNet+MsTNL+ MsGCS) has been improved significantly. The average Jac, DSC, Pre and Ppmcc of MsTGANet have been improved from 0.7325, 0.8445, 0.8412 and

TABLE V
EVALUATION INDICES FOR MsTNL MODULE WITH DIFFERENT CHANNELS

Network	Jac	DSC	Pre	Ppmcc	Time(s)
UNet+MsTNL(1/16C)	0.7366±0.0058	0.8472±0.0042	**0.8440±0.0052**	0.8475±0.0042	0.0186
UNet+MsTNL(1/8C)	**0.7386±0.0060**	**0.8486±0.0041**	0.8421±0.0106	**0.8490±0.0040**	0.0186
UNet+MsTNL(C)	0.7378±0.0078	0.8480±0.0054	0.8432±0.0134	0.8483±0.0055	0.0188

0.8449 to 0.7412, 0.8502, 0.8548 and 0.8504, respectively. The ablation experiment results show that the proposed MsTGANet (UNet+MsTNL+ MsGCS) can improve the drusen segmentation accuracy in OCT images significantly.

V. CONCLUSION

In this paper, we propose a novel MsTGANet to improve the accuracy of drusen segmentation in retinal OCT images. In MsTGANet, two newly proposed MsTNL module and MsGCS module are designed, and both modules are combined with the U-shape architecture to improve the model's ability to learn the multi-scale features with long dependency information and the capacity of multi-semantic global contextual feature capture. It is the first time that a multi-scale transformer method has been developed and applied to the drusen segmentation task to explore multi-scale long-term dependency information, also the first time to propose a novel adaptively global attention method that integrate features from channel and spatial dimensions to improve the model's capacity to capture multi-semantic global contextual features. Furthermore, a novel Semi-MsTGANet based on pseudo-labeled data augmentation strategy also be proposed to alleviate the impact of insufficient labeled data, which can leverage unlabeled data to further improve the segmentation performance. We conducted comprehensive experiments to validate the segmentation performance of the proposed MsTGANet and Semi-MsTGANet. The experimental results show that compared with other state-of-the-art CNN-based networks, the segmentation performance of the proposed MsTGANet and Semi-MsTGANet have been improved significantly.

In our future works, we will collect more OCT data with drusen from different OCT scanners and acquisition modes to build a larger and more comprehensive database to further evaluate the performance and robustness of the proposed MsTGANet. In addition, we will also explore different semi-supervised learning strategies based on MsTGANet to further improve the performance of drusen segmentation in OCT images.

REFERENCES

[1] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[2] D. Xiang *et al.*, "Automatic segmentation of retinal layer in OCT images with choroidal neovascularization," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5880–5891, Dec. 2018.

[3] X. Chen, M. Niemeijer, L. Zhang, K. Lee, M. D. Abràmoff, and M. Sonka, "Three-dimensional segmentation of fluid-associated abnormalities in retinal OCT: Probability constrained graph-search-graph-cut," *IEEE Trans. Med. Imag.*, vol. 31, no. 8, pp. 1521–1531, Aug. 2012.

[4] L. Zhang, W. Zhu, F. Shi, H. Chen, and X. Chen, "Automated segmentation of intraretinal cystoid macular edema for retinal 3D OCT images with macular hole," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 1494–1497.

[5] L. Ye, W. Zhu, D. Bao, S. Feng, and X. Chen, "Macular hole and cystoid macular edema joint segmentation by two-stage network and entropy minimization," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 735–744.

[6] E. Talisa *et al.*, "Spectral-domain optical coherence tomography angiography of choroidal neovascularization," *Ophthalmology*, vol. 122, no. 6, pp. 1228–1238, 2015.

[7] X. Xi *et al.*, "Automated segmentation of choroidal neovascularization in optical coherence tomography images using multi-scale convolutional neural networks with structure prior," *Multimedia Syst.*, vol. 25, no. 2, pp. 95–102, Apr. 2019.

[8] S. Zhu, F. Shi, D. Xiang, W. Zhu, H. Chen, and X. Chen, "Choroid neovascularization growth prediction with treatment based on reaction-diffusion model in 3-D OCT images," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1667–1674, Nov. 2017.

[9] C. K. Chan *et al.*, "Optical coherence tomography-measured pigment epithelial detachment height as a predictor for retinal pigment epithelial tears associated with intravitreal bevacizumab injections," *Retina*, vol. 30, no. 2, pp. 203–211, 2010.

[10] Z. Sun *et al.*, "An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, Feb. 2016.

[11] F. Costello, L. Malmqvist, and S. Hamann, "The role of optical coherence tomography in differentiating optic disc drusen from optic disc edema," *Asia–Pacific J. Ophthalmol.*, vol. 7, no. 4, pp. 271–279, 2018.

[12] D. Xiang *et al.*, "Automatic retinal layer segmentation of OCT images with central serous retinopathy," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 283–295, Jan. 2019.

[13] H. Chen *et al.*, "Quantitative analysis of retinal layers' optical intensities on 3D optical coherence tomography for central retinal artery occlusion," *Sci. Rep.*, vol. 5, no. 1, p. 9269, Aug. 2015.

[14] B. Hassan, G. Raja, T. Hassan, and M. U. Akram, "Structure tensor based automated detection of macular edema and central serous retinopathy using optical coherence tomography images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 33, no. 4, pp. 455–463, 2016.

[15] M. Wang *et al.*, "Semi-supervised capsule cGAN for speckle noise reduction in retinal OCT images," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1168–1183, Apr. 2021.

[16] J. Fang, Y. Zhang, K. Xie, S. Yuan, and Q. Chen, "An improved mpb-cnn segmentation method for edema area and neurosensory retinal detachment in SD-OCT images," in *Proc. Int. Workshop Ophthalmic Med. Image Anal.* Cham, Switzerland: Springer, 2019, pp. 130–138.

[17] J. Yang, Z. Ji, S. Niu, Q. Chen, S. Yuan, and W. Fan, "RMPPNet: Residual multiple pyramid pooling network for subretinal fluid segmentation in SD-OCT images," *OSA Continuum*, vol. 3, no. 7, pp. 1751–1769, 2020.

[18] F. Shi *et al.*, "Automated 3-D retinal layer segmentation of macular optical coherence tomography images with serous pigment epithelial detachments," *IEEE Trans. Med. Imag.*, vol. 34, no. 2, pp. 441–452, Feb. 2015.

[19] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," *Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[20] Q. Chen *et al.*, "Automated drusen segmentation and quantification in SD-OCT images," *Med. Image Anal.*, vol. 17, no. 8, pp. 1058–1072, Dec. 2013.

[21] D. Mittal and K. Kumari, "Automated detection and segmentation of drusen in retinal fundus images," *Comput. Electr. Eng.*, vol. 47, pp. 82–95, Oct. 2015.

[22] M. U. Akram, S. Mujtaba, and A. Tariq, "Automated drusen segmentation in fundus images for diagnosing age related macular degeneration," in *Proc. Int. Conf. Electron., Comput. Comput. (ICECCO)*, Nov. 2013, pp. 17–20.

[23] X. Ren, X. Zheng, X. X. Dong, and X. Cui, "Deep feature extraction via adaptive collaborative learning for drusen segmentation from fundus images," *Signal, Image Video Process.*, vol. 1, no. 1, pp. 1–8, 2020.

[24] Q. Pham, S. Ahn, S. J. Song, and J. Shin, "Automatic drusen segmentation for age-related macular degeneration in fundus images using deep learning," *Electronics*, vol. 9, no. 10, pp. 1617–1628, 2020.

[25] F. Yan *et al.*, "Deep random walk for drusen segmentation from fundus images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 48–55.

[26] X. Ren *et al.*, "Drusen segmentation from retinal images via supervised feature learning," *IEEE Access*, vol. 6, pp. 2952–2961, 2017.

[27] Q. Chen, T. Leng, S. Niu, J. Shi, L. D. Sisternes, and D. L. Rubin, "A false color fusion strategy for drusen and GA visualization in OCT images," *Retina Philadelphia, PA*, vol. 34, no. 12, pp. 2346–2371, 2014.

[28] N. Jain *et al.*, "Quantitative comparison of drusen segmented on SD-OCT versus drusen delineated on color fundus photographs," *Invest. Ophthalmol. Vis. Sci.*, vol. 51, no. 10, pp. 4875–4883, 2010.

[29] X. Xi, X. Meng, Z. Qin, X. Nie, Y. Yin, and X. Chen, "IA-Net: Informative attention convolutional neural network for choroidal neovascularization segmentation in OCT images," *Biomed. Opt. Exp.*, vol. 11, no. 11, pp. 6122–6136, 2020.

[30] R. Asgari, S. Waldstein, F. Schlanitz, M. Baratsits, U. S. Erfurth, and H. Bogunović, "U-Net with spatial pyramid pooling for drusen segmentation in optical coherence tomography," in *Proc. Int. Workshop Ophthalmic Med. Image Anal.* Cham, Switzerland: Springer, 2019, pp. 77–85.

[31] R. Asgari *et al.*, "Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 192–200.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[34] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation," 2019, *arXiv:1903.11816*. [Online]. Available: http://arxiv.org/abs/1903.11816

[35] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[37] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[38] Z. Gu *et al.*, "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[39] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[40] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[41] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[43] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*. [Online]. Available: http://arxiv.org/abs/1901.02860

[44] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*. [Online]. Available: http://arxiv.org/abs/1906.08237

[45] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.

[46] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019, *arXiv:1906.05909*. [Online]. Available: http://arxiv.org/abs/1906.05909

[47] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[48] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 323–339.

[49] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: http://arxiv.org/abs/2010.11929

[50] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," 2021, *arXiv:2101.11605*. [Online]. Available: http://arxiv.org/abs/2101.11605

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[53] H. Zhang, I. Goodfellow, D. Metaxas, and A, "Odena. Self-attention generative adversarial network," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[54] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[56] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.

[57] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: http://arxiv.org/abs/1802.06955

[58] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[59] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[60] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5688–5696.

[61] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[62] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.