Lightweight Retinal Layer Segmentation With Global Reasoning

Xiang He, Weiye Song, Yiming Wang, *Member, IEEE*, Fabio Poiesi, Ji Yi, Manishi Desai, Quanqing Xu, Kongzheng Yang, and Yi Wan

***Abstract*— Automatic retinal layer segmentation with medical images, such as optical coherence tomography (OCT) images, serves as an important tool for diagnosing ophthalmic diseases. However, it is challenging to achieve accurate segmentation due to low contrast and blood flow noises presented in the images. In addition, the algorithm should be lightweight to be deployed for practical clinical applications. Therefore, it is desired to design a lightweight network with high performance for retinal layer segmentation. In this article, we propose LightReSeg for retinal layer segmentation which can be applied to OCT images. Specifically, our approach follows an encoder–decoder structure, where the encoder part uses multiscale feature extraction and a transformer block for fully exploiting the semantic information of feature maps at all scales and making the features have better global reasoning capabilities, while the decoder part, we design a multiscale asymmetric attention (MAA) module for preserving the semantic information at each encoder scale. The experiments show that our approach achieves a better segmentation performance compared with the current state-of-the-art method TransUnet with 105.7 M parameters on both our collected dataset and two other public datasets, with only 3.3 M parameters.**

Manuscript received 22 December 2023; revised 12 April 2024; accepted 14 April 2024. Date of publication 21 May 2024; date of current version 5 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 51975336 and Grant 62205181; in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF017; in part by the Natural Science Outstanding Youth Fund of Shandong Province under Grant 2023HWYQ-023; in part by the Taishan Scholar Foundation of Shandong Province under Grant tsqn202211038; in part by NIH under Grant R01NS108464; in part by the Key Technology Research and Development Program of Shandong Province under Grant 2020JMRH0202 and Grant 2022CXGC020701; in part by the Shandong Province New Old Energy Conversion Major Industrial Tackling Projects under Grant 2021-13; in part by the Key Research and Development Project of Jining City under Grant 2021DZP005; and in part by the Shandong University Education Teaching Reform Research Project under Grant 2022Y133, Grant 2022Y124, and Grant 2022Y312. The Associate Editor coordinating the review process was Dr. Zhenghua Chen. *(Corresponding author: Yi Wan.)*

Xiang He is with the School of Mechanical Engineering and the Joint SDU-NTU Center for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, Shandong 250061, China (e-mail: he_xiang@mail.sdu.edu.cn).

Weiye Song, Quanqing Xu, Kongzheng Yang, and Yi Wan are with the School of Mechanical Engineering, Shandong University, Jinan, Shandong 250061, China (e-mail: songweiye@sdu.edu.cn; 202214333@mail.sdu.edu.cn; 202234432@mail.sdu.edu.cn; wanyi@sdu.edu.cn).

Yiming Wang and Fabio Poiesi are with the Fondazione Bruno Kessler, 38123 Trento, Italy (e-mail: ywang@fbk.eu; poiesi@fbk.eu).

Ji Yi is with the Department of Biomedical Engineering and the Department of Ophthalmology, Johns Hopkins University, Baltimore, MD 21231 USA (e-mail: jiyi@jhu.edu).

Manishi Desai is with the Department of Ophthalmology, Boston Medical Center, Boston University School of Medicine, Boston, MA 02118 USA (e-mail: madesai@bu.edu).

Digital Object Identifier 10.1109/TIM.2024.3400305

***Index Terms*— Lightweight, multiscale asymmetric attention (MAA), retinal layer segmentation, visible-light optical coherence tomography (OCT).**

I. Introduction

WITH the accelerated pace of lifestyle and the popularity of electronic products, many people lack awareness of eye protection, leading to an increasing incidence of retinal diseases. Many retinal diseases are accompanied by changes in the thickness of the retinal layers. For instance, in patients with glaucoma, the retinal nerve fiber layer (NFL) is thinner and the optic disk is atrophied compared with the healthy eyes [1], [2]. In diabetic retinopathy, the macular region of the retina thickens and becomes edematous [3]. Current research indicates that the changes in the retinal layer begin to occur in the early stages of the disease [4], [5], and therefore, automatic analysis of the retinal layers and monitoring their morphological changes can help understand the disease progression and provide early treatment.

OCT can be used for noninvasive 3-D structural imaging of the retina [6], with which experienced ophthalmologists can manually inspect the retinal layers to identify the disease progression. However, manual inspection is inefficient and tedious. It is also challenging for various levels of ophthalmologists to ensure consistency and objectivity of the inspection. Although imaging techniques with a higher resolution are available now, such as visible-light OCT images [7], [8], [9], this does not address completely the above status quo. Recent convolutional neural networks (CNNs), as typified by U-net [10], have achieved promising results regarding semantic segmentation in the field of medical image analysis. The seminar work ReLayNet [11] has triggered widespread interest in applying CNNs to segment retinal layers. ReLayNet follows a multiscale encoder–decoder structure, which has inspired many mainstream semantic segmentation frameworks to adopt, such as U-net [10] and Attention_Unet [12]. However, such structure can be limited in two main aspects. First, most U-shaped encoder–decoder structures perform the multiscale feature extraction followed by feature fusion via only residual connections. Thus, it is limited to maintain and fully exploit the semantic information of feature maps at all scales, resulting in false positives in the background region when performing retinal layer segmentation. Second, many of these methods do not pay attention to the actual application needs of OCT

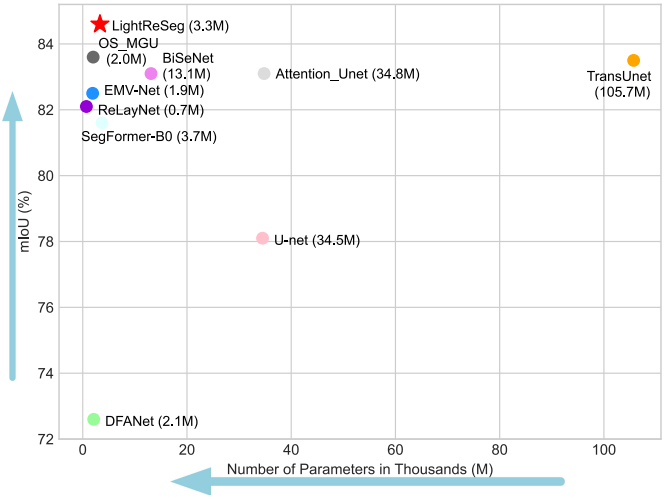


Fig. 1. Model size versus the segmentation accuracy in terms of mIoU of the state-of-the-art retinal layer segmentation methods. Our LightReSeg achieves the highest mIoU compared with SOTA methods while maintaining a smaller model size.

devices, often pursuing accuracy unilaterally, for example, such multiscale reasoning often comes at the cost of a large number of parameters, which can be computationally demanding for the deployment of real-time clinical applications.

To overcome the above-mentioned limitations, we propose LightReSeg, a novel multiscale encoder–decoder network for end-to-end retinal layer segmentation, which is a lightweight segmentation approach tailored to practical OCT device requirements. To reduce segmentation errors in the background region, we propose to incorporate a transformer block [13] to features with global reasoning and to introduce a multiscale asymmetric attention (MAA) module to better preserve the semantic information at each encoder scale. To achieve computation efficiency, both the backbone and the MAA module adopt lightweight feature extractors, such as depthwise separable convolution and asymmetric convolution [14], [15]. LightReSeg achieves the state-of-the-art segmentation accuracy with only 3.3 M parameters, a significantly light model, as shown in Fig. 1.

In summary, our contributions are as follows.

1) We propose LightReSeg, a novel encoder–decoder structure for retinal layer segmentation, which exploits global information in a lightweight manner, to improve the segmentation accuracy with reduced computational complexity, and it provides valuable experience for algorithms designed for practical applications in medical devices.
2) We propose a novel attention module, the MAA module, to jointly work with transformer, to address erroneous segmentation in the background area by allowing the model to reason in a global manner.
3) We perform our method on the visible-light OCT images for the first time and score the best segmentation performance, and it provides experience for the algorithm to perform robustly on datasets from different domains.

This article is organized as follows: Section II describes the research progress in the field of biological tissue segmentation and retinal layer segmentation. Section III introduces the overall framework, the feature extractor, MAA module, and the lightweight designs of the approach. Section IV describes the datasets, performance metrics, implementation details, analyzes quantitatively and qualitatively the experimental results, and performs ablation experiments. Finally, the conclusions of the article are given in Section V.

II. Related Work

In this section, we first provide an overview of the research progress of semantic segmentation, especially lightweight semantic segmentation methods, followed by a more specific coverage of deep-learning-based retinal layer segmentation.

A. Medical Image Segmentation

Medical image segmentation has made rapid progress under the promotion of deep learning, especially in the fields of biological tissue recognition and lesion detection, which has sparked a large amount of research on medical image segmentation. The morphology of the optic disk and optic cup, as well as the cup-to-disk ratio, can be used to assess glaucoma, Wang et al. [16] proposed an automatic segmentation approach based on CNNs to accurately segment the optic disk and optic cup from fundus images for glaucoma detection. Precise segmentation of retinal vessels from fundus images is essential for intervention in numerous diseases, and high-precision segmentation of retinal vessels still remains a challenging task; Li et al. [17] proposed a dual-path progressive fusion network, named DPF-Net, which achieved better segmentation ability by effectively fusing global and local features. Abdominal multiorgan image segmentation plays a crucial role in the diagnosis and treatment of many diseases. Traditional methods of manual depiction are inefficient and subjective, and therefore, Chen et al. [18] proposed TransUnet based on CNN and transformer block for abdominal multiorgan segmentation. TransUnet was the first model to introduce the transformer block into the U-shaped encoder–decoder structure and achieved very high accuracy in segmentation. There are also segmentation studies dedicated to lesion regions, such as Astaraki et al. [19] who segmented different types of lung pathological regions to enable screening for lung diseases. In addition, some lightweight segmentation methods for practical applications have also been proposed, such as SegFormer [20], A-net [21], and EMV-Net [22].

B. Retinal Layer Segmentation

The morphology of the retinal layer is associated with the development of many ophthalmic diseases, and therefore, there are many studies on retinal layer segmentation based on CNN and OCT images. The success of ReLayNet opened the door to the application of CNNs for retinal layer segmentation [11], a typical U-shaped network based on an encoder–decoder structure, which achieved the best performance for retinal layer segmentation at that time. Since layer segmentation of OCT retinal images is prone to speckle noise, intensity inhomogeneity, Wang et al. proposed a boundary-aware U-Net

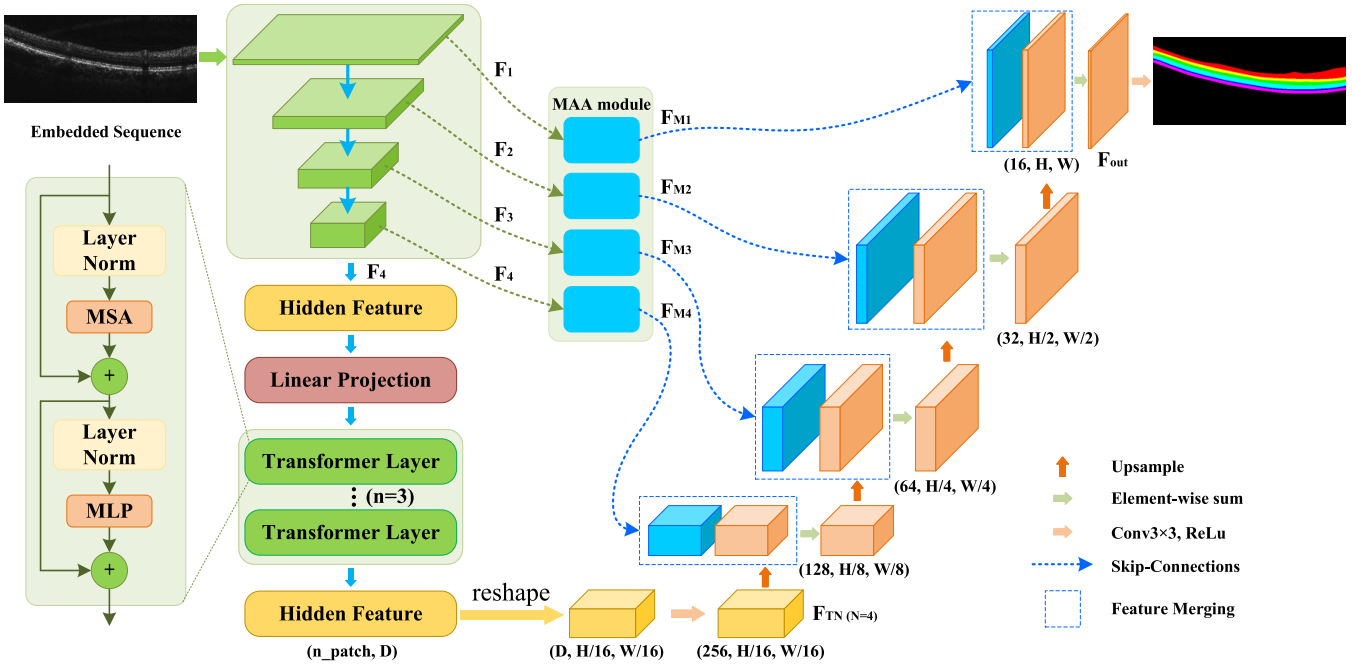


Fig. 2. Network of LightReSeg follows a U-shaped encoder–decoder structure. The encoder takes as the input a retinal image of dimension $(3, H, W)$ and performs multiscale feature extraction that outputs feature maps of N scales, where N is set to 4 in our design. The last feature map is further fed to the transformer layers through a linear transformation to extract features that reasons in long range to help the reduction in segmentation errors in the background region. The resulted features after the transformer layers are then fused via reshaping and upsampling with the multiscale encoder features that are optimized by the proposed MAA module. The final fused feature map F_{out} is exploited for the retinal layer segmentation via convolutions.

for retinal layer segmentation by detecting accurate boundaries [23]. To solve the topological errors caused by not considering the order of retinal layers, Liu et al. [24] proposed a novel deep-learning-based framework that used the distance maps of layer surfaces to convert the layer segmentation task into a multitasking problem for classification and regression. To ensure the continuity of the retinal layer boundary and obtain more accurate segmentation results, Hu et al. [25] proposed a coarse-to-fine retinal layer boundary segmentation method based on the embedded residual recurrent network and the graph search, and it has achieved good performance in both qualitative and quantitative indicators. In addition, considering the impact of different neurological diseases on the retinal layer, Gende et al. [26] presented a fully automatic approach for the retinal layer segmentation in multiple neurodegenerative disorder scenarios, and the results indicate that the model is more robust. There are also studies aimed at practical application deployment, such as He et al. proposed a lightweight retinal layer segmentation approach that achieved good performance with a lower number of parameters [22].

III. Methodology

A. Framework Overview

Our proposed LightReSeg is a U-shaped network based on an encoder–decoder structure as shown in Fig. 2. The network takes a retinal layer OCT image as input. The first N scale feature maps $(F_1, \ldots, F_{N-1}, F_N)$ are extracted by the encoder which consists of a multiscale feature extractor, respectively. To further deepen the global reasoning capability in the depth direction, the feature map F_N has to pass through a depth feature encoder: transformer block [13]. The feature map F_{TN} further encoded by the transformer block has no change in shape and size, but it facilitates long-range global reasoning. To better preserve details at different scales, multiscale features $(F_1, \ldots, F_{N-1}, F_N)$ are then merged with its corresponding layers in the decoder part via our proposed MAA module, which outputs the feature maps $(F_{M1}, \ldots, F_{MN-1}, F_{MN})$ with the same shape size as its input. F_{TN} output from the transformer block is merged with F_{MN} for feature fusion, and then after twofold size upsampling, it continues to fuse with F_{MN-1} for feature fusion and then follows a similar operation until the fusion of the upsampled restored feature map with F_{out} is completed, and after channel adjustment, the final retinal layer segmentation image is output. LightReSeg is designed for real-time clinical operations, where lightweight designs are used where possible.

We present in full the multiscale encoder in Section III-B followed by the proposed MAA module that serves for multiscale feature fusion in Section III-C. Finally, we describe the design details for computation efficiency in Section III-D.

B. Multiscale Encoder

The multiscale encoder extracts N scale features to fully exploit the semantic information of feature maps at all scales; further, to make the features have better global reasoning capabilities, we incorporate a transformer block [13], which is a module that further optimizes the Nth scale feature map extracted from the deepest scale of the encoder. To ensure the dimensionality is consistent with the transformer block port, we preprocess the feature map F_N before being processed by

the transformer block. The feature map F_N is first converted into a 2-D patch sequence as

$$F_N \Rightarrow F'_N = \left\{ x_p^k \in R^{P^2 \times c} \mid k = 1, \ldots, Z \right\}, \quad Z = \frac{HW}{P^2} \tag{1}$$

where c is the number of channels in the F_N feature map, P is the size of the patch, and H and W are the height and width of the F_N feature map, respectively.

We use a trainable linear projection to map the vector patch x_p to a latent D-dimensional embedding space. To encode the spatial information of the patches, we learn specific position embeddings that are added to preserve the positional information in the patch embedding as

$$z_0 = \left[x_0; x_p^1 E; x_p^2 E; \ldots; x_p^N E \right] + E_{\text{pos}} \tag{2}$$

where $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ represents the patch embedding projection, x_0 is an additional learnable vector concatenated with the remaining vectors to integrate the information of all the remaining vectors, $[\cdot; \ldots; \cdot]$ is the concatenation operator, and $E_{\text{pos}} \in \mathbb{R}^{(Z+1) \times D}$ represents the position embedding [13].

Then, inputting z_0 into the transformer layer, as shown in Fig. 2, first goes through the first layer norm (LN) layer and then enters the multihead self-attention (MSA) layer; here, $\text{MSA}(\text{LN}(z_0)) + z_0$ forms the residual structure and get the z'_0. Then, after passing through the second LN layer, it enters the multilayer perceptron (MLP) layer; here, $\text{MSA}(\text{LN}(z'_0)) + z'_0$ once again forms the residual structure and get z_1.

At this point, the first transformer layer is finished. We have set three transformer layers in our approach, so we have to cycle three times

$$z_L = \left[x'_0; x_p^{1'}; x_p^{2'}; \ldots; x_p^{N'} \right] \Rightarrow \left[x_p^{1'}; x_p^{2'}; \ldots; x_p^{N'} \right]. \tag{3}$$

When z_L outputs from the transformer block, we have to remove the x'_0 vector from the z_L, because only then we can reshape it to the same size of F_N.

C. MAA Module

To better preserve the semantic information at each encoder scale, the MAA module is designed in the skip connection part of the encoder and decoder by a large convolution kernel and channel attention mechanism. It is also set with multiscale feature fusion, and it is based on the ideas of asymmetric convolution [15], multiscale feature extraction [27], [28], and channel attention mechanism [29]

$$F'_0 = \text{Conv}_{5\times5}(F_0) \tag{4}$$

$$F'_{0i} = \text{Conv}_{ki\times1}\left[\text{Conv}_{1\times ki}\left(F'_0\right)\right], \quad k = 3, 7, 11 \tag{5}$$

$$F''_{0i} = CA\left(F'_{0i}\right). \tag{6}$$

The workflow of the MAA module is illustrated in Fig. 3, and it begins with the input of a feature map F_0 extracted from the feature encoder, assuming a shape size of (C, H_0, W_0). F_0 enters MAA and goes through a 5×5 convolution kernel to expand the perceptual field, and then has to enter a multiscale convolution to extract information at different scales [see (4), (5)]. Each scale goes through a channel attention (CA) module to increase the weight of the feature map on the channel dimension [see (6)]. As shown in the channel attention module in Fig. 3, the feature $F'_{0i} \in \mathbb{R}^{C\times H\times W}$ reshape to $F'_{0i\_cn} \in \mathbb{R}^{C\times N}$, where $N = H \times W$, and $\in$ represents the shape of the feature map. After that, we perform a matrix multiplication between F'_{0i_cn} and the transpose of it and get the $F'_{0i\_cc} \in \mathbb{R}^{C\times C}$. Eventually, we apply the softmax layer to get the channel attention map $A \in \mathbb{R}^{C\times C}$

$$F'_{0i\_cc} = F'_{0i\_cn} \times \left(F'_{0i\_cn}\right)^T \tag{7}$$

$$A = \text{Softmax}\left(F'_{0i\_cc}\right). \tag{8}$$

[Equations (7) and (8)]. Here, the matrix $A \in \mathbb{R}^{C\times C}$ contains information about the weights between channels, e.g., $(A \in \mathbb{R}^{C\times C})_{ij}$ represents the impact of ith channel on the jth channel. In addition, we multiply the matrix between the transpose of $A \in \mathbb{R}^{C\times C}$ and

$$A_1 \in \mathbb{R}^{C\times N} = A^T \times F'_{0i\_cn} \tag{9}$$

$$F''_{0i} = \alpha A_1 \in \mathbb{R}^{C\times N \overset{\text{reshape}}{\rightarrow} C\times H\times W} + F'_{0i}. \tag{10}$$

F'_{0i_cn}, and reshape the result to $\mathbb{R}^{C\times H\times W}$. Then we multiply the result by a scale parameter α and perform an elementwise sum operation with F'_{0i} to obtain the final output F''_{0i} [see (9) and (10)]. After the CA module, all the feature maps are summed and convolved by 1×1 kernel to generate a weighted feature map, and we use the weighted feature map to reweight F_0 by elementwise multiplying ($\bigotimes$) it with the input F_0

$$\text{att} = \text{Conv}_{1\times1}\left[\sum_{i=1}^{3} F''_{0i} + CA\left(F'_0\right)\right] \tag{11}$$

$$F_{M0} = \text{att} \bigotimes F_0. \tag{12}$$

Finally, we obtain F_{M0}. According to the above procedure, we can input the extracted four scales feature maps $F_1, \ldots, F_{N-1}, F_N$ into the MAA module to get $F_{M1}, \ldots, F_{MN-1}, F_{MN}$, respectively, as shown in Fig. 2.

Instead, the ASPP uses a series of dilated convolution layers with different rates to obtain larger receptive fields; however, dilated convolutions may lose local detail information. In contrast, the MAA module uses asymmetric convolutions with large kernels, which allows capturing multiscale features without sacrificing too much detail. Furthermore, before the 1×1 convolution compresses the channels in the MAA module, a channel attention mechanism is adopted, enhancing the module's focus on the channel dimension and better preserving important features.

D. Lightweight Designs

The basic principle of our design of LightReSeg is to improve the overall performance of the model with as few parameters as possible, so we can meet the computational efficiency of deployments for real-time clinical applications. Thus, we make some lightweight designs for part of the structure of the model. For encoder extraction of multiscale feature maps, we use depthwise separable convolutions (DS-Conv) for our downsampling [14]. DS-Conv considers the channels and spatial regions of the feature map separately,

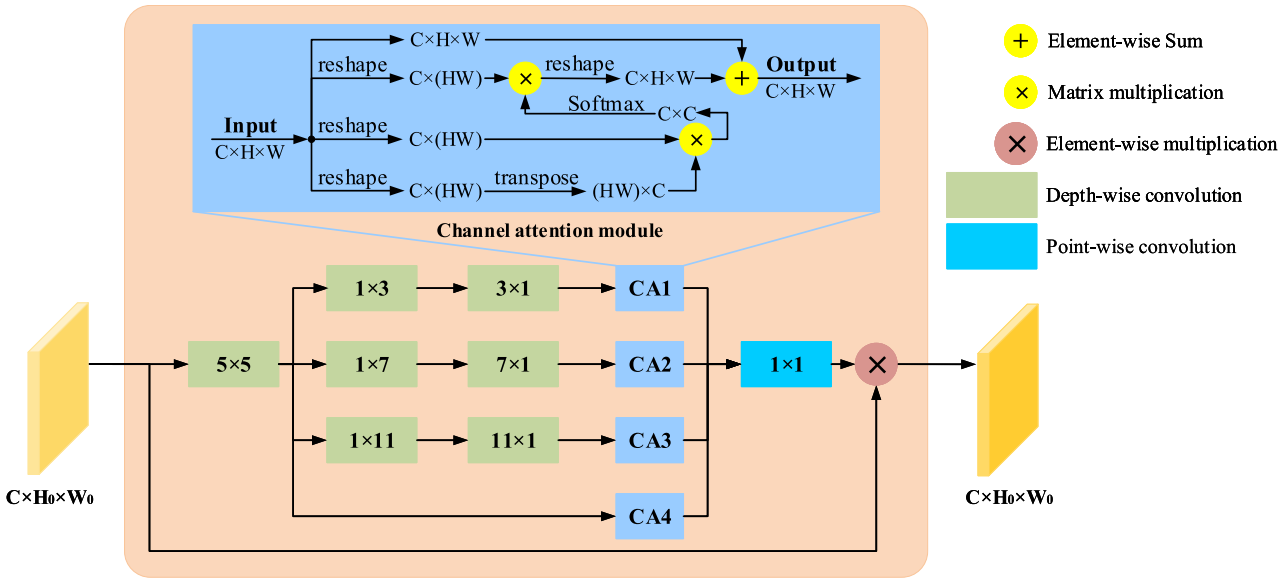


Fig. 3. Structure of the MAA module.

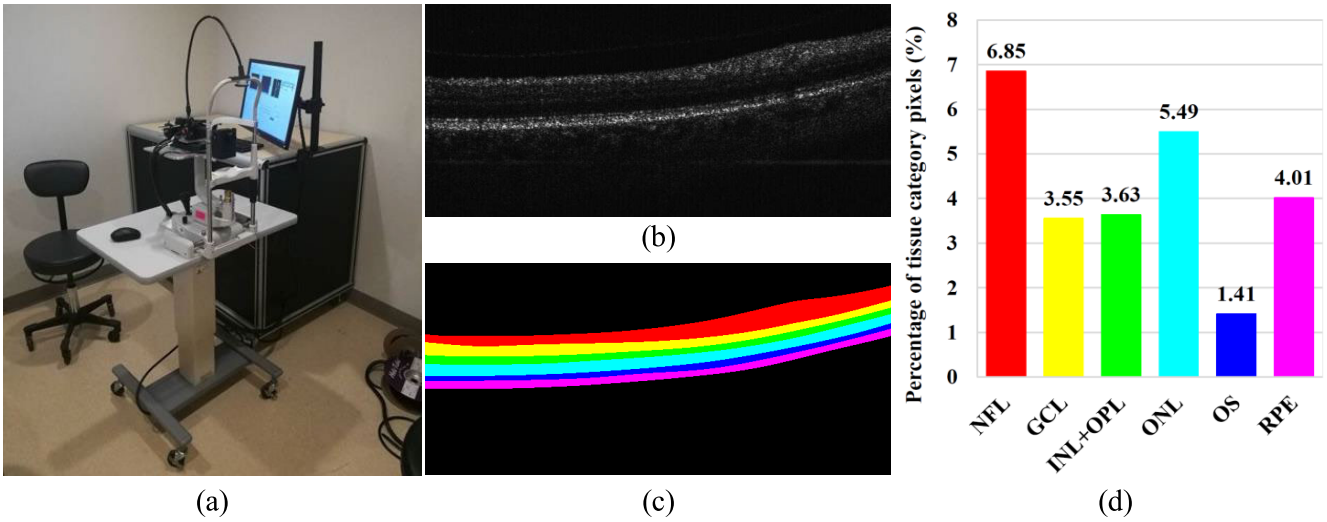


Fig. 4. (a) Visible-light OCT device for capturing images. (b) Visible-light OCT B-scan image. (c) Annotation image (ground truth). (d) Average percentage of pixels on OCT images for each tissue layer besides the background (the percentage of background is 75.06%).

takes different convolution kernels for different channels, and then uses point convolution to aggregate the channel information, achieving a richer feature representation with fewer parameters to learn. We find that using standard convolution for downsampling is better than pooling, but convolution increases the number of parameters, so we use DS-Conv to ensure the segmentation performance while decreasing the burden of the number of parameters. In addition, as mentioned in Section III-C, multiscale asymmetric convolution is used in our MAA module. Initially, our idea is to perform multiscale extraction of feature information like ASPP [28], but we find that using standard convolution would cause a greater burden on the number of parameters, since the MAA module is supposed to perform extraction of the N scale feature maps, which would certainly further increase the number of parameters. To reduce the number of parameters in the MAA module, we use asymmetric convolution instead of standard convolution, shown in Fig. 3, and the overall performance of LightReSeg is nearly the same. In particular, the transformer block has a significant proportion of the number of parameters in LightReSeg. We have optimized the internal parameters of the transformer block structure, such as the setting of heads $=$ 8 for multiple heads in MSA and dim_head $=$ 64 for hidden linear layers in MSA [13]. All these designs to minimize the number of parameters are made under the precondition of ensuring the performance of LightReSeg.

IV. Experiment

In this section, we compare our LightReSeg against the state-of-the-art methods for retinal layer segmentation on: 1) two publicly available OCT datasets with unhealthy eyes, where the Glaucoma dataset is for the glaucoma disease and the DME dataset is for diabetic macular edema (DME) and 2) a new dataset, named as Vis-105H, which we create on top of raw images collected from [33] which correspond to

TABLE I

MULTIPLE APPROACHES TO MULTIPLE METRICS [S_{MPA} (%), S_{MIoU} (%), S_{DSC} (%), S_{PA} (%)] EVALUATION ON THE VIS-105H DATASET, NO. 1 AND NO. 2 PLACES ARE GIVEN IN BOLD IN BLACK AND BOLD IN GRAY FOR EACH METRIC, RESPECTIVELY

Method	Indicators	Tissue Layers						S_{mPA}	S_{mIoU}
		NFL	GCL	INL/OPL	ONL	OS	RPE		
ReLayNet [11]	S_{DSC}	93.6	85.6	85.5	93.5	87.9	94.1	97.3	82.1
	S_{PA}	93.3	82.1	86.0	97.2	87.5	91.1		
OS_MGU [30]	S_{DSC}	**94.8**	86.8	86.7	**94.1**	**88.6**	**94.5**	**97.7**	**83.6**
	S_{PA}	93.5	82.8	87.4	**97.8**	**88.1**	91.2		
DFANet [31]	S_{DSC}	91.1	80.0	80.8	88.3	73.5	88.8	95.1	72.6
	S_{PA}	91.3	77.0	82.9	93.3	74.4	86.3		
BiSeNet [32]	S_{DSC}	**94.9**	87.3	**88.3**	93.9	84.8	**94.5**	**97.8**	83.1
	S_{PA}	**95.4**	**83.9**	**89.7**	96.6	84.9	91.2		
Attention_Unet [12]	S_{DSC}	94.6	**87.4**	87.6	93.6	87.1	93.7	97.6	83.1
	S_{PA}	**94.6**	83.4	87.6	**97.6**	87.5	91.2		
EMV-Net [22]	S_{DSC}	94.6	85.3	85.5	93.7	87.9	94.5	97.6	82.5
	S_{PA}	93.5	81.9	86.2	97.1	**90.1**	**91.7**		
U-net [10]	S_{DSC}	92.6	82.5	83.4	91.5	82.6	92.3	96.8	78.1
	S_{PA}	90.3	80.8	84.8	96.6	80.8	90.0		
SegFormer-B0 [20]	S_{DSC}	94.5	83.4	84.1	93.4	88.2	94.2	97.4	81.6
	S_{PA}	92.8	79.0	86.9	**97.6**	86.3	91.3		
TransUnet [18]	S_{DSC}	94.4	86.9	87.5	**94.1**	87.5	**94.7**	**97.7**	83.5
	S_{PA}	93.4	83.2	88.0	97.0	87.2	**91.6**		
LightReSeg (Our)	S_{DSC}	94.3	**87.5**	**89.1**	**94.7**	**89.1**	**94.5**	**97.8**	**84.6**
	S_{PA}	94.3	**84.5**	**89.4**	96.9	87.8	**91.6**		

only healthy human eyes with visible light OCT. Moreover, we present a thorough ablation study with our Vis-105H dataset to validate the effectiveness of the proposed MAA and the introduction of the transformer block.

A. Datasets

1) Vis-105H Dataset: We create the Vis-105H dataset for the method evaluation which contains images captured using a custom *visible light* OCT device that takes optic disk cubes from each subject [34], [35], [36], as shown in Fig. 4(a). The raw images are originally collected and presented in [33], and they correspond to 21 eyes from 14 healthy subjects, all of whom with written informed consent prior to the participation. The inclusion criteria include: 1) age $\geqslant$ 40 years; 2) best-corrected visual acuity better than 20/40; 3) no previous intraocular surgery of any kind; and 4) no known retinal disease, resulting in 21 scanning cubes with 5376 B-scan images. Two professional ophthalmologists further select five images with high imaging quality from every 256 B-scans, as shown in Fig. 4(b), with the selection criteria including: 1) low noise; 2) clear retinal layer boundaries; and 3) suitability for ophthalmologists to use for diagnosis. With the supervision of medical ophthalmology professionals, we annotate these 105 preprocessed retinal visible-light OCT images of healthy human eyes for semantic segmentation and complete the Vis-105H dataset for seven-class semantic segmentation including the background, as shown in Fig. 4(c). Specifically, we exploit the graphics software Inkscape to annotate the six layers, NFL, ganglion cell layer (GCL) and inner plexiform layer (IPL), inner nuclear layer (INL) and out plexiform layer (OPL), outer nuclear layer (ONL), outer segment (OS), and retinal pigment epithelium (RPE), as red, yellow, green, light blue, dark blue, and pink, respectively, and to annotate the remaining regions as background black. Each annotated image is further examined by a professional ophthalmologist and exported in PNG format with a size of 660 $\times$ 300 pixels. Fig. 4(d) shows the average pixel percentages of all the categories on the OCT image except for the background, from which we can observe that the most dominant two classes are the NFL and ONL layers while the OS and GCL layers are the least dominant classes. The Vis-105H dataset is divided into a training set, a validation set, and a test set consisting of 75, 15, and 15 images, respectively. The data for the Vis-105 dataset are conducted at BMC, and both BU and BMC hold ownership rights to the data.

2) Glaucoma Dataset: The dataset is collected from 61 different subjects [30], where 12 radial OCT B scans of each subject are collected using DRI OCT-1 Atlantis at the Ophthalmology Department of Shanghai General Hospital. All the images are scanned at 20.48 $\times$ 7.94 mm field of view in the optic nerve head region. Under the supervision of a glaucoma specialist, two ophthalmologists manually annotate these images into the optic disk and nine retinal layers. The image size is 1024 $\times$ 992, and the dataset is divided into training, validation, and test sets by 148, 48, and 48 images, respectively.

3) DME Dataset: The dataset is collected by Chiu et al. [37] using the Duke Enterprise Data Unified Content Explorer search engine to retrospectively identify DME subjects within the Duke Eye Center. It consists of 110 OCT B-scans obtained from 10 patients with DME with a size of 496 $\times$ 768 pixels, and each B-scan is annotated with nine layers. The dataset is divided by 88, 11, and 11 into training, validation, and test sets, respectively. All the subjects are scanned with the macula at the center.

TABLE II

MULTIPLE APPROACHES TO MULTIPLE METRICS [S_{MPA} (%), S_{MIOU} (%), S_{DSC} (%), S_{PA} (%)] EVALUATION ON THE DME DATASET, NO. 1 AND NO. 2 PLACES ARE GIVEN IN BOLD IN BLACK AND BOLD IN GRAY FOR EACH METRIC, RESPECTIVELY

Method	Indicators	Tissue Layers								S_{mPA}	S_{mIoU}
		NFL	GCL/IPL	INL	OPL	ONL/ISM	ISE	OS/RPE	Fluid		
ReLayNet [11]	S_{DSC}	82.0	93.4	79.4	77.2	86.8	86.6	**86.5**	53.8	95.5	69.6
	S_{PA}	79.5	93.3	76.2	76.7	87.7	89.7	87.2	52.0		
OS_MGU [30]	S_{DSC}	**82.6**	**93.6**	**79.5**	**78.4**	87.2	87.1	85.6	57.7	**95.6**	69.8
	S_{PA}	**80.7**	**93.7**	77.0	77.2	**88.3**	89.5	84.6	52.1		
DFANet [31]	S_{DSC}	77.4	90.3	73.6	72.6	85.8	85.7	85.5	49.8	94.7	64.8
	S_{PA}	73.7	91.4	70.4	70.8	84.1	89.4	85.9	47.3		
BiSeNet [32]	S_{DSC}	80.9	92.7	77.9	75.3	86.3	86.1	**86.6**	56.7	95.4	68.3
	S_{PA}	76.7	93.0	76.8	72.2	86.9	87.5	**89.8**	51.0		
Attention_Unet [12]	S_{DSC}	80.5	91.5	77.6	75.9	86.8	86.6	86.4	55.6	95.3	68.0
	S_{PA}	79.4	92.4	74.4	74.0	86.8	88.9	86.6	49.9		
EMV-Net [22]	S_{DSC}	80.4	92.5	78.5	74.8	85.6	86.3	86.3	**59.9**	95.3	68.4
	S_{PA}	74.9	91.6	**81.8**	71.3	81.2	89.4	87.6	**60.2**		
U-net [10]	S_{DSC}	82.2	92.8	79.1	78.2	86.7	**87.4**	**86.6**	56.2	95.5	69.4
	S_{PA}	**81.4**	92.9	76.2	76.5	87.6	**90.7**	86.7	52.2		
SegFormer-B0 [20]	S_{DSC}	81.7	**93.7**	79.1	77.0	86.6	86.6	85.7	57.6	95.6	69.2
	S_{PA}	77.5	**95.0**	74.9	77.0	**89.8**	**89.9**	**87.8**	49.1		
TransUnet [18]	S_{DSC}	81.4	93.1	**79.8**	77.4	**87.4**	87.2	**86.6**	**60.1**	**95.7**	**70.0**
	S_{PA}	80.0	93.0	**77.7**	**77.6**	88.2	89.5	86.7	53.9		
LightReSeg (Our)	S_{DSC}	**82.4**	**93.7**	**79.8**	**78.8**	**87.9**	**87.3**	86.0	59.7	**95.7**	**70.5**
	S_{PA}	79.7	93.5	77.1	**78.1**	87.8	89.6	85.6	**57.8**		

TABLE III

MULTIPLE APPROACHES TO MULTIPLE METRICS [S_{MPA} (%), S_{MIOU} (%), S_{DSC} (%), S_{PA} (%)] EVALUATION ON THE GLAUCOMA DATASET, NO. 1 AND NO. 2 PLACES ARE GIVEN IN BOLD IN BLACK AND BOLD IN GRAY FOR EACH METRIC, RESPECTIVELY

Method	Indicators	Tissue Layers										S_{mPA}	S_{mIoU}
		NFL	GCL	IPL	INL	OPL	ONL	IS/OS	RPE	Choroid	OD		
ReLayNet [11]	S_{DSC}	79.4	**64.9**	**70.8**	**76.4**	**81.1**	90.5	**86.0**	**93.1**	87.9	77.8	94.2	**67.0**
	S_{PA}	77.2	**60.7**	**70.6**	77.0	**80.5**	90.0	84.9	85.2	88.6	77.4		
OS-MGU [30]	S_{DSC}	80.7	62.0	69.6	74.2	78.4	89.6	84.7	**83.0**	88.5	83.7	**94.8**	66.7
	S_{PA}	82.8	56.1	68.7	74.4	75.7	88.3	85.0	**85.7**	88.7	86.8		
DFANet [31]	S_{DSC}	79.8	60.5	68.2	74.3	77.4	89.4	85.0	81.3	88.2	83.9	94.7	65.8
	S_{PA}	80.0	55.9	66.1	71.9	73.1	91.7	**85.7**	80.8	89.6	88.6		
BiSeNet [32]	S_{DSC}	78.0	61.5	67.4	70.1	77.1	88.5	82.4	78.3	86.9	**84.3**	94.5	64.0
	S_{PA}	76.4	57.0	65.6	66.2	77.4	89.5	82.3	81.6	90.1	**89.3**		
Attention_Unet [12]	S_{DSC}	79.5	60.1	64.7	72.8	78.8	**90.6**	**86.3**	82.8	**88.8**	83.2	94.7	66.0
	S_{PA}	84.2	57.9	59.4	71.0	75.0	91.9	**85.6**	81.8	89.1	84.8		
EMV-Net [22]	S_{DSC}	81.2	**65.3**	69.5	72.1	79.1	90.0	83.9	81.5	88.5	83.4	**94.8**	66.6
	S_{PA}	83.7	59.3	67.8	65.8	74.8	88.6	81.5	83.4	89.9	86.8		
U-net [10]	S_{DSC}	**81.9**	63.7	69.4	73.5	77.5	89.7	85.4	82.3	87.8	83.3	**94.8**	66.6
	S_{PA}	83.5	58.1	68.3	73.7	74.6	91.7	84.2	83.4	**91.0**	86.5		
SegFormer-B0 [20]	S_{DSC}	**82.1**	64.3	69.4	**76.4**	76.6	89.9	84.4	81.2	**88.6**	**84.2**	**94.8**	66.9
	S_{PA}	**86.7**	58.5	63.9	**81.7**	69.7	**92.6**	82.8	79.8	87.7	**90.7**		
TransUnet [18]	S_{DSC}	80.6	60.0	69.4	76.1	80.2	**90.8**	85.8	82.6	88.5	82.4	94.7	**67.0**
	S_{PA}	**84.7**	53.3	69.0	77.6	78.5	**93.6**	83.0	80.3	86.9	84.3		
LightReSeg (Our)	S_{DSC}	80.9	64.3	**70.5**	**77.2**	**80.4**	90.1	85.4	82.7	88.4	83.3	**94.9**	**67.8**
	S_{PA}	83.5	**61.1**	**70.8**	**78.8**	**78.7**	90.2	85.0	**85.3**	**90.5**	83.5		

B. Performance Metrics

We evaluate the segmentation performance quantitatively using the Dice similarity coefficient (DSC), mean intersection over union (mIoU), pixel accuracy (PA), and mean PA (mPA). They can be computed as

$$S_{\text{DSC}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{13}$$

$$S_{\text{mIoU}} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}} \tag{14}$$

$$S_{\text{PA}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{15}$$

and

$$S_{\text{mPA}} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{16}$$

where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. k represents the number of categories. In addition, we count the number of parameters for compared approaches to represent their computation complexity.

C. Implementation Details

LightReSeg is implemented based on the PyTorch and trained with the Adam optimizer with the cross-entropy loss. The initial learning rate is set to 0.001 which is then gradually halved every 40 epochs. Data augmentation is applied

on all the three datasets, including horizontal flipping with probability P = 0.5, random center rotation within plus or minus 20 degrees with probability P = 0.5, median and motion blur processing with P = 0.5 probability, random Gaussian noise addition, and random brightness and contrast with P = 0.5 probability. All the experiments are conducted on an NVIDIA GeForce RTX 3090 Graphics card. The code of the proposed LightReSeg could be found at: https://github.com/Medical-Image-Analysis/LightReSeg.

D. Comparison

We compare LightReSeg with the state-of-the-art approaches including ReLaynet [11], EMV-Net [22], Attention_Unet [12], BiSeNet [32], DFANet [31], OS_MGU [30], U-net [10], and TransUnet [18]. We report the results on the above-mentioned three datasets.

1) Quantitative Analysis: Table I shows the comparison of our approach with the state-of-the-art methods on the Vis-105H dataset. We can see that our approach LightReSeg scores the best performance in both S_{mPA} and S_{mIoU} metrics, with $+1\%$ improvement in terms of the S_{mIoU} compared with the second-best performing method OS_MGU [30]. In terms of S_{DSC} metric, our approach achieves the highest segmentation performance in the GCL layer, INL/OPL layer, ONL layer, and OS layer. Regarding the S_{PA} metric, our approach achieves the first place in the GCL layer, and the second place in the INL/OPL layer and RPE layer, which also shows the outstanding performance of our proposed approach. The segmentation performance of ReLayNet, the lightest algorithm in the field of retinal layer segmentation, is -2.5% lower than our proposed LightReSeg in terms of S_{mIoU} metric. We further perform a statistical significance test, using the Wilcoxon rank sum test, to compare the Dice score performance of different methods on each layer. When comparing our approach with ReLayNet, we observe a P_{value} of 0.031250 ($p < 0.05$), indicating a statistically significant difference. The poor performance of ReLayNet might be due to its rather simple feature fusion strategy for combining features from the encoder and decoder, i.e., skip connections, leading to limited semantic information at the fusion. Instead, LightReSeg further introduces the MAA module for the feature fusion at multiple scales. Moreover, we add transformer layers at the bottom of the encoder for efficient global reasoning which further improves the segmentation accuracy. SegFormer is a new lightweight model in the field of semantic segmentation and has a good performance in many tasks [20], but as can be seen from Table I, SegFormer-B0 only ranks seventh in terms of S_{mIoU} and S_{mPA} metrics, while in terms of S_{DSC} metric, its segmentation performance of all the layers is much lower than ours. In terms of S_{mPA} metric, only the ONL layer is higher than our by 0.7% points, and the segmentation performance in the rest of the layers is much lower than our approach. We analyze that the poor segmentation performance of SegFormer-B0 on this task may be mainly due to two factors: first, its excessive pursuit of lightweight results in limited ability to extract detailed features; second, the upsampling part of SegFormer-B0's decoder is too rough, resulting in the lack of fine detail recovery.

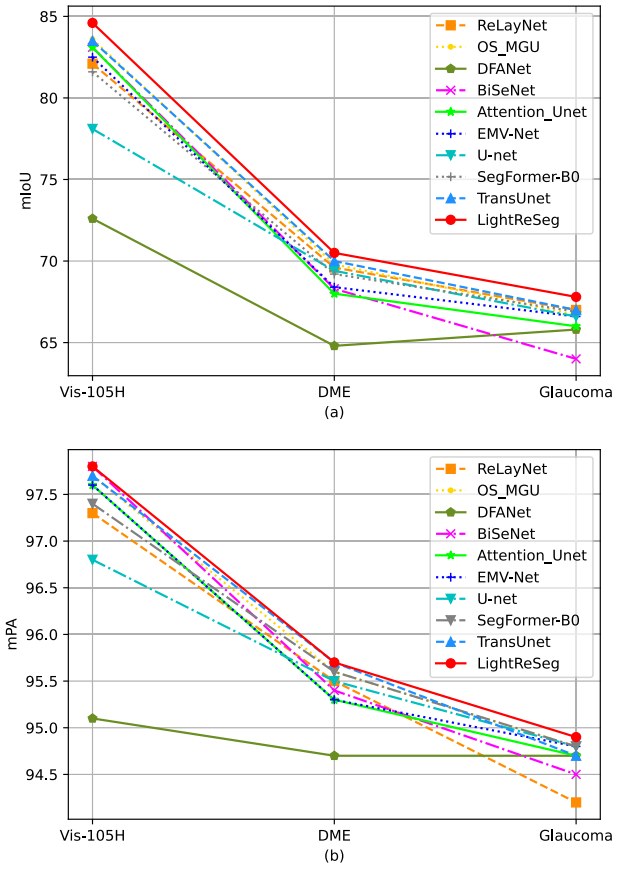


Fig. 5. (a) and (b) Performance of the nine mainstream approaches on the three retinal layer segmentation datasets measured by mIoU and mPA metrics, respectively.

We evaluate our approach on the DME dataset, as shown in Table II. We achieve the best performance in terms of both S_{mPA} and S_{mIoU}. In terms of S_{DSC} we achieve first place in four layers (i.e., GCL/IPL, INL, OPL, and ONL/ISM) and second place in another two layers (i.e., NFL and ISE), while in terms of S_{mPA}, we are on par with the state-of-the-art performance. Among other approaches, such as TransUnet, OS_MGU, and ReLayNet, although they are more advanced in S_{mPA} metric, they are 0.5%, 0.7%, and 0.9% points lower than the first place in S_{mIoU} metric, respectively, which also indicates that our model also has a strong comprehensive performance on the DME dataset. We further perform the statistical significance test using the Wilcoxon rank sum test. For example, when comparing our method with TransUnet on the same domain, we observe a P_{value} of 0.017629 ($p < 0.05$), indicating a statistically significant difference. Similar statistically significant differences are observed, with P_{values} of 0.023437 ($p < 0.05$), 0.039062 ($p < 0.05$), and 0.017755 ($p < 0.05$), respectively, when comparing our method with OS_MGU, EMV-Net, and SegFormer-B0. We further evaluate our approach on the Glaucoma dataset, as shown in Table III. Our LightReSeg achieves the best results on both the overall metrics S_{mPA} and S_{mIoU}. In addition, our approach also scores better on most of the layers in terms of both S_{DSC} and S_{PA}.

Fig. 5(a) and (b) shows the performance of the ten mainstream approaches on the three datasets measured

by mIoU and mPA metrics, respectively. Our approach LightReSeg achieves the best performance on all the three datasets in terms of S_{mIoU} and S_{mPA}. TransUnet is the second best-performing method with a small margin lower than ours on all the three datasets; however, its number of parameters is 32 times of the one of our approach. The lightest model ReLayNet is inferior to our method in terms of performance by -2.5%, -0.9%, and -0.8% on the Vis-105H, DME, and Glaucoma datasets, respectively. Therefore, our approach achieves state-of-the-art performance with a relatively lightweight model.

In addition, we find that regardless of the method, certain layers (NFL, ONL, RPE) have significantly higher accuracy compared with other layers, as shown in Table I. We believe that class imbalance is the most likely cause, as the number of samples in a certain class is much larger than in other classes, resulting in the model learning better for that class during training and exhibiting higher accuracy in evaluation. From Fig. 4(d), we can see that the average pixel proportion of NFL, ONL, and RPE layers is relatively high, and Table I also shows that these three layers have higher accuracy compared with other layers. In addition, there are differences in shape, color, texture, and other aspects among different layers, which may make certain categories of targets easier to segment while others are more difficult. In summary, we believe that the best way to solve this problem is to first address class imbalance.

2) Qualitative Analysis: Fig. 6 shows the segmentation prediction maps of several mainstream segmentation methods in a retinal B-scan image that contains blood flow information disturbance. As shown in Fig. 6(j), our approach achieves much better segmentation performance. The prediction graphs of all the approaches in Fig. 6(c), (d), (f), and (h) showed false-positive errors, specifically, the region above the NFL layer or below the RPE layer of the retina is incorrectly identified as the retinal layer and segmented, resulting in the wrong segmentation target. The TransUnet in Fig. 6(e) performs well, but two locations on the boundary between the NFL and GCL layers are inaccurately segmented, as marked by the two locations in Fig. 6(e). The same occurs in EMV-Net in one place, as shown in Fig. 6(g). In comparison, our approach is not only without false-positive errors but also segmentation boundaries are more accurate.

Fig. 7 shows the comparative prediction plots of several mainstream segmentation approaches in a single B-scan image over the retinal optic nerve head region. As shown in Fig. 7(j), our approach shows a better segmentation performance. The black dashed region in Fig. 7(d), where the NFL and GCL layer boundaries appear distinctly jagged, and the black dashed region in Fig. 7(f), where a distinct honeycomb shape appears, are signs of unsmooth and inaccurate layer boundary prediction, and the same occurs in Fig. 7(c), (e), and (h). In addition, the problem of segmentation category error occurs, such as the red dashed area in Fig. 7(d), where the segmentation results in incorrectly segmenting part of the OPL layer into ONL layers, and this segmentation category error occurs at least twice in Fig. 7(e)–(g). In contrast, our approach not only has no problem with segmentation category error but also the segmented boundary is smoother than other approaches.

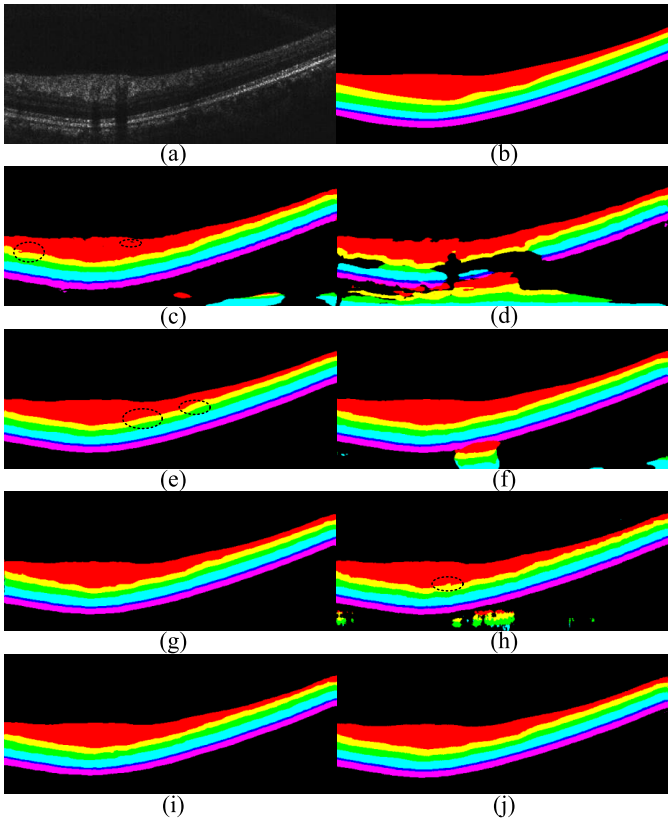


Fig. 6. Comparison of segmentation prediction maps of mainstream approaches on the Vis-105H dataset. (a) Original image. (b) Ground truth (c) Prediction map of U-net. (d) Prediction map of DFANet. (e) Prediction map of TransUnet. (f) Prediction map of Attention_Unet. (g) Prediction map of EMV-Net. (h) Prediction map of ReLayNet. (i) Prediction map of SegFormer-B0. (j) Prediction map of LightReSeg. The black dashed line framed area in the image is the region where the boundary of the prediction layer is not smooth or inaccurate.

E. Ablation Study

1) Performance of Different Modules: We conduct ablation experiments on the Vis-105H dataset with the main purpose of verifying the effect of the MAA module and the contribution of the transformer block on the overall method segmentation performance. The ablated methods are: Base is the U-shaped backbone; Base_MAA and Base_trans3 are MAA and three transformer layers added to Base, respectively; Base_MAA_trans3 and Base_MAA_trans6 are based on Base_MAA with three and six transformer layers, respectively. Note that Base_MAA_trans3 represents the final version of our proposed LightReSeg. As shown in Table IV, the Base_MAA approach improves over the Base approach, S_{mPA} and S_{mIoU} by $+0.4\%$ and $+1.9\%$, respectively. All the six layers except for the OS layer are also significantly improved in S_{DSC}. We also observe that Base_MAA_trans3 improves over Base_trans3 in terms of both S_{mPA} and S_{mIoU}. Fig. 8(c) and (d) shows that the segmentation performance of the model is greatly improved by adding the MAA module to the model, and the results of mis-segmentation are significantly reduced in the background region below the RPE layer.

Regarding the transformer block, as shown in Table IV, Base_trans3 improves $+0.5\%$–$+2.4\%$ over Base in terms of

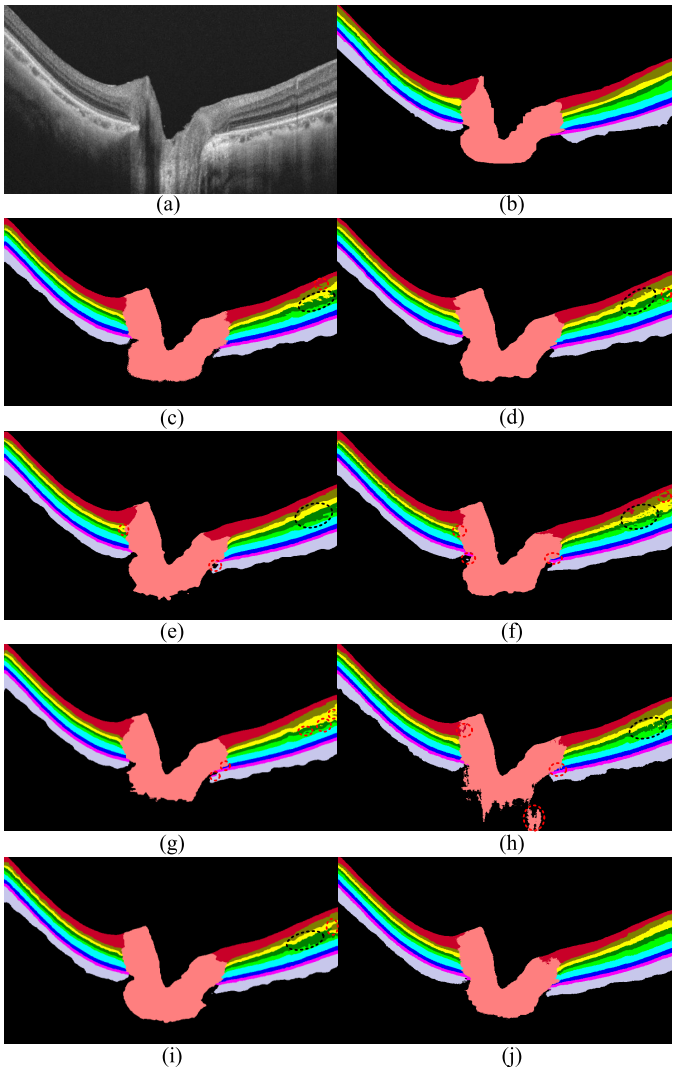


Fig. 7. Comparison of segmentation prediction maps of mainstream approaches on the Glaucoma dataset. (a) Original image. (b) Ground truth (c) Prediction map of OS_MGU. (d) Prediction map of DFANet. (e) Prediction map of TransUnet. (f) Prediction map of Attention_Unet. (g) Prediction map of EMV-Net. (h) Prediction map of ReLayNet. (i) Prediction map of SegFormer-B0. (j) Prediction map of LightReSeg. The black dashed line framed area in the image is the region where the boundary of the prediction layer is not smooth or inaccurate, the red dashed area is where the prediction map has a segmentation category error.

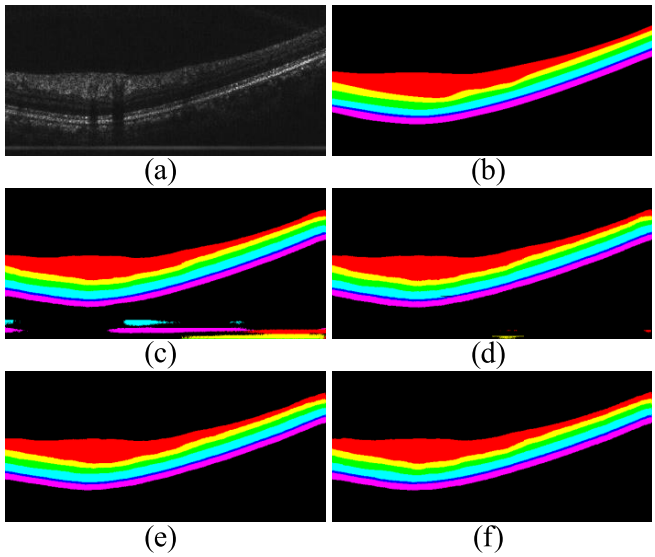


Fig. 8. Prediction images of our method under different ablation settings. (a) Original image. (b) Ground truth. (c) Base framework. (d) Base framework plus MAA module. (e) Base framework plus three transformer blocks. (f) Base framework plus MAA module and six transformer blocks.

S_{mPA} and S_{mIoU}, respectively, and the segmentation ability of each layer is improved in terms of S_{DSC}. In addition, we also find that increasing the number of transformer layers can further improve the segmentation performance within a certain range. For example, Base_MAA_trans6 adds three more transformer layers than Base_MAA_trans3, which improves S_{mPA} and S_{mIoU} by $+0.1\%$–0.6%, respectively, and this comes at a cost of $+40\%$ more of parameters. Fig. 8(c) and (e) shows qualitatively that adding three sequential transformer blocks to the model contributes to a big improvement in the segmentation performance of the model, and there is no segmentation error in the background region below the RPE layer.

To further interpret the role of different modules, we use a modified version of Grad-CAM [38] to visualize the feature activations in different layers of the model (see Fig. 9). Specifically, in accordance with the model settings in Table IV, we extract the feature activation heat maps for different retinal layers at different model structures and at different positions of the model, respectively. By comparing Fig. 9(e) and (i), with the addition of the MAA module, we can see that the heat map of the corresponding NFL layer in Fig. 9(i) is further enhanced compared with Fig. 9(e). The same trend is also observed in Fig. 9(h) and (l). It is further found that with the addition of the MAA module to the Base_trans3 model structure, the feature region corresponding to the segmentation category is also further enhanced, as shown in Fig. 9(p) and (t). Adding the transformer blocks to the model also causes changes in the heat map, as in Fig. 9(f) and (n). The heat map of the corresponding RPE layer region is activated more comprehensively with the addition of the transformer blocks, and the OS and background layers of the adjacent regions also show different degrees of activation.

When analyzing the activation maps, we observe that not only the regions corresponding to the predicted layers prominently highlighted but also the feature weights of the adjacent layers are also significantly amplified. To interpret this behavior, we conduct further analysis and find that the boundaries of the retinal layers exhibit a high degree of correlation with the adjacent layers. As in Fig. 9(e), the focus is mainly on the NFL layer, but adjacent layers (e.g., background and GCL) also show various degrees of activation. Similarly, in Fig. 9(r), the RPE layer of the retinal layer is the main focus, while information from the OS and background layers at the upper and lower boundaries of the retinal layer is also used. This phenomenon suggests that the model uses information from adjacent layers to enhance its predictions. In addition, we also export the feature activation maps for different deep network layers in the prediction of the RPE layer, and we find that the heat map of the deep network

TABLE IV

PROPOSED APPROACH PERFORMS MULTIPLE METRICS EVALUATION OF ABLATION EXPERIMENTS ON THE VIS-105H DATASET

Method	S_{DSC} (%) Results of Tissue Layers						S_{mPA} (%)	S_{mIoU} (%)	Params (M)
	NFL	GCL	INL/OPL	ONL	OS	RPE			
Base	94.0	84.6	84.8	93.2	87.9	92.6	97.2	81.3	1.61
Base_MAA	94.4	86.9	87.7	94.0	87.4	94.0	97.6	83.2	1.72
Base_trans3	94.7	86.9	87.3	94.1	89.0	94.0	97.7	83.7	3.20
Base_MAA_trans3(Our)	94.3	87.5	89.1	94.7	89.1	94.5	97.8	84.6	3.30
Base_MAA_trans6	95.0	87.6	89.0	94.8	89.7	95.1	97.9	85.2	4.68

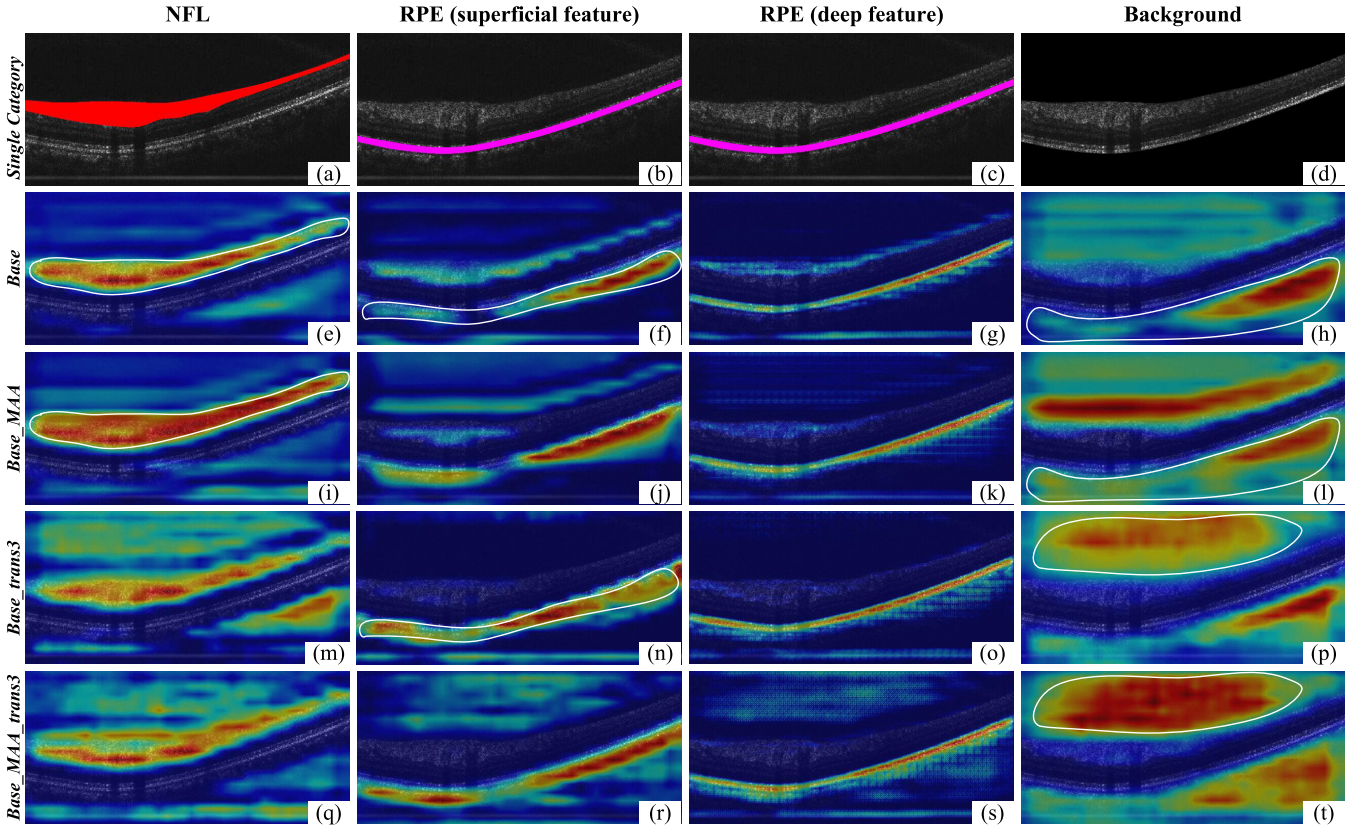


Fig. 9. Heatmap for different retinal layers at different model structures and at different positions of the model, (a)–(d) for the single category prediction, (e)–(h) for the Base structure, (i)–(l) for the Base_MAA structure, (m)–(p) for the Base_trans3 structure, and (q)–(t) for the Base_MAA_trans3 (Our) structure. The white line highlights the contrast area in the graph.

focuses on more detailed and accurate regions compared with the heat map of the superficial network. As in Fig. 9(o), the regions where the heat map is focused on activation are highly coincident with the RPE layer, while the regions activated by the heat map in Fig. 9(n) are much coarser, which is similarly shown in Fig. 9(r) and (s). This observation suggests that the deeper extract image features of the model network contain more accurate and insightful semantic information. Overall, the visualization of deep feature activation provides valuable insights into the model's decision process, and the model can use multiple layers of information to obtain more robust segmentation results.

2) Lightweight Setup: To evaluate the potential impact of DS-Conv on feature extraction capabilities, we replace all DS-Conv in the encoder section with standard convolutions and conduct comparative analyses [39], as shown in Table V. The results indicate that using standard convolutions instead of DS-Conv does not enhance feature extraction in our optimized model and leads to an 11% increase in the parameter count. This demonstrates that a simple increase in the model's parameter count does not always result in performance.

TABLE V

IMPACT OF DS-CONV ON MODEL PERFORMANCE

Method Settings	$S_{mPA}(\%)$	$S_{mIoU}(\%)$	Params(M)
Encoder with DS-Conv	97.8	84.6	3.30
Encoder with Conv	97.8	84.3	3.69

For assessing the impact of model lightweight on performance, we modify the number of channels in the encoder's multiscale features, both halving and doubling them, and the specific results are shown in Table VI. We find that based on our model, when the channel of the multiscale features extracted by the encoder is reduced to 0.5× and 0.25×,

TABLE VI
IMPACT OF MODEL LIGHTWEIGHT ON PERFORMANCE

Method Settings	mPA(%)	mIoU(%)	Params(M)
0.25x	96.9	79.2	0.48
0.5x	96.9	79.1	1.18
1x	97.8	84.6	3.30
2x	97.8	84.3	9.87
4x	97.7	84.1	33.48

TABLE VII
STATISTICAL INFERENCE SPEED OF DIFFERENT METHODS ON THREE DATASETS, AND THE BEST PERFORMANCE ON EACH DATASET IS GIVEN IN BOLD

Method	Params	Vis-105H	DME	Glaucoma
ReLayNet [11]	$0.7M$	$0.08s$	$0.20s$	$0.06s$
EMV-Net [22]	$1.9M$	$0.15s$	$0.31s$	$0.13s$
OS_MGU [30]	$2.0M$	$0.12s$	$0.30s$	$0.07s$
DFANet [31]	$2.1M$	$0.12s$	$0.31s$	$0.08s$
SegFormer-B0 [20]	$3.7M$	$0.12s$	$0.28s$	$0.07s$
BiSeNet [32]	$13.1M$	$0.08s$	$0.20s$	$0.05s$
U-net [10]	$34.5M$	$0.10s$	$0.20s$	$0.08s$
Attention_Unet [12]	$34.8M$	$0.10s$	$0.22s$	$0.09s$
TransUnet [18]	$105.7M$	$0.14s$	$0.32s$	$0.11s$
LightReSeg (Our)	$3.3M$	$0.11s$	$0.27s$	$0.07s$

the model's parameter count decreased, but its segmentation performance significantly deteriorated. When we increase the channel count to $2\times$ and $4\times$, there is no significant change in the model's accuracy, yet the parameter counts are $3\times$ and $11\times$ the original size, respectively. From this, we can also conclude that our lightweight design maintains accuracy while using the fewest possible parameters.

F. Limitation

Although LightReSeg boasts only 3.3 M trainable parameters and delivers the optimal retinal layer segmentation outcomes, in real-world applications, model parameter size is not the sole determinant of algorithm efficacy. Inference speed emerges as a crucial metric for assessing algorithmic effectiveness. As evidenced by Table VII, our proposed method exhibits inference speeds of 0.11, 0.27, and 0.07 s on the three datasets, respectively. While these figures already signify remarkable efficiency, there remains scope for further enhancement. For example, although the segmentation accuracy of ReLayNet is far inferior to our approach, it must be admitted that its inference efficiency is slightly higher than ours. Moreover, in the context of practical clinical applications, constraints related to dataset limitations preclude the inclusion of all the device data types. Imagery obtained through disparate devices may exhibit domain discrepancies, inevitably resulting in false-positive errors in the segmentation output for unfamiliar new devices. Consequently, apart from enlarging the dataset, enhancing the model's domain generalization competency is a subject worthy of further investigation. The supplementary segmentation result evaluation system will also augment ophthalmologists' confidence in the segmentation outcomes.

V. CONCLUSION

In this article, we propose a novel lightweight method LightReSeg for retinal layer segmentation. Our method introduces a transformer-based block in the encoder part to enable global reasoning for reducing errors in the background region and an attention mechanism, named MAA module, to best exploit rich semantic information for the multiscale feature fusion to improve the segmentation accuracy. The extensive evaluation shows that our approach achieves the best segmentation performance on both our collected Vis-105H dataset and two other public ones, i.e., DME and Glaucoma, indicating that our model has good reliability in the face of noise and uncertainty in the data. To improve the efficiency, our method also incorporates lightweight designs. The ablation experiments detailed in Section IV-E2 demonstrate that our lightweight design is precisely adequate, maintaining the highest level of accuracy while significantly reducing the number of parameters compared to other high-performing methods. This further highlights the practicality of our proposed approach, particularly evident in the real-time preview functionality of small-field-of-view OCTA imaging. In the future, we will collect more clinical data to further improve our approach and continuously contribute to the efficient performance of OCT devices.

REFERENCES

[1] F. K. Horn et al., "Correlation between local glaucomatous visual field defects and loss of nerve fiber layer thickness measured with polarimetry and spectral domain OCT," *Investigative Opthalmol. Vis. Sci.*, vol. 50, no. 5, p. 1971, May 2009.

[2] F. Pollet-Villard, C. Chiquet, J.-P. Romanet, C. Noel, and F. Aptel, "Structure-function relationships with spectral-domain optical coherence tomography retinal nerve fiber layer and optic nerve head measurements," *Investigative Opthalmol. Vis. Sci.*, vol. 55, no. 5, p. 2953, May 2014.

[3] A. Ajaz, H. Kumar, and D. Kumar, "A review of methods for automatic detection of macular edema," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102858.

[4] C. Brandl et al., "Retinal layer thicknesses in early age-related macular degeneration: Results from the German AugUR study," *Investigative Opthalmol. Vis. Sci.*, vol. 60, no. 5, p. 1581, Apr. 2019.

[5] J. A. van de Kreeke et al., "Retinal layer thickness in preclinical Alzheimer's disease," *Acta Ophthalmolog.*, vol. 97, no. 8, pp. 798–804, 2019.

[6] D. Huang et al., "Optical coherence tomography," *Science*, vol. 254, pp. 1178–1181, Nov. 1991.

[7] X. Shu, L. Beckmann, and H. F. Zhang, "Visible-light optical coherence tomography: A review," *J. Biomed. Opt.*, vol. 22, no. 12, Dec. 2017, Art. no. 121707.

[8] W. Song, W. Shao, and J. Yi, "Wide-field and micron-resolution visible light optical coherence tomography in human retina by a linear-K spectrometer," in *Bio-Optics: Design and Application*, 2021, Paper DM2A-4.

[9] W. Song, W. Shao, W. Yi, and J. Yi, "Linear k-domain wide-field and micron-resolution visible light human retinal optical coherence tomography," *Proc. SPIE*, vol. 11941, Mar. 2022, Art. no. PC11941.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[11] A. G. Roy et al., "ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Exp.*, vol. 8, no. 8, pp. 3627–3642, 2017.

[12] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[13] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[15] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.

[16] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, "Patch-based output space adversarial learning for joint optic disc and cup segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2485–2495, Nov. 2019.

[17] J. Li, G. Gao, L. Yang, G. Bian, and Y. Liu, "DPF-Net: A dual-path progressive fusion network for retinal vessel segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.

[18] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[19] M. Astaraki, Ö. Smedby, and C. Wang, "Prior-aware autoencoders for lung pathology segmentation," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102491.

[20] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[21] B. Chen, T. Niu, W. Yu, R. Zhang, Z. Wang, and B. Li, "A-Net: An A-shape lightweight neural network for real-time surface defect segmentation," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[22] X. He et al., "Exploiting multi-granularity visual features for retinal layer segmentation in human eyes," *Frontiers Bioeng. Biotechnol.*, vol. 11, Jun. 2023, Art. no. 1191803.

[23] B. Wang, W. Wei, S. Qiu, S. Wang, D. Li, and H. He, "Boundary aware U-Net for retinal layers segmentation in optical coherence tomography images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3029–3040, Aug. 2021.

[24] X. Liu, J. Cao, S. Wang, Y. Zhang, and M. Wang, "Confidence-guided topology-preserving layer segmentation for optical coherence tomography images with focus-column module," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[25] K. Hu, D. Liu, Z. Chen, X. Li, Y. Zhang, and X. Gao, "Embedded residual recurrent network and graph search for the segmentation of retinal layer boundaries in optical coherence tomography," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–17, 2021.

[26] M. Gende et al., "Automatic segmentation of retinal layers in multiple neurodegenerative disorder scenarios," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 11, pp. 5483–5494, Nov. 2023.

[27] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," 2022, *arXiv:2209.08575*.

[28] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.

[29] 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.

[30] J. Li et al., "Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images," *Biomed. Opt. Exp.*, vol. 12, no. 4, pp. 2204–2220, 2021. [Online]. Available: http://opg.optica.org/boe/abstract.cfm?URI=boe-12-4-2204

[31] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiseNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 325–341.

[33] W. Song et al., "Visible light optical coherence tomography of peripapillary retinal nerve fiber layer reflectivity in glaucoma," *Transl. Vis. Sci. Technol.*, vol. 11, no. 9, p. 28, Sep. 2022.

[34] W. Song, S. Zhang, N. Sadlak, M. G. Fiorello, M. Desai, and J. Yi, "Visible and near-infrared optical coherence tomography (vnOCT) in normal, suspect, and glaucomatous eyes," *Proc. SPIE*, vol. 11623, Mar. 2021, Art. no. 1162311.

[35] J. Yi and W. Song, "Systems and methods for fiber-based visible and near infrared optical coherence tomography," U.S. Patent 10 732 354, Aug. 4, 2020.

[36] W. Song, L. Zhou, S. Zhang, S. Ness, M. Desai, and J. Yi, "Fiber-based visible and near infrared optical coherence tomography (vnOCT) enables quantitative elastic light scattering spectroscopy in human retina," *Biomed. Opt. Exp.*, vol. 9, no. 7, p. 3464, 2018.

[37] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Exp.*, vol. 6, no. 4, pp. 1172–1194, 2015.

[38] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract)," in *Proc. AAAI*, 2020, pp. 13943–13944.

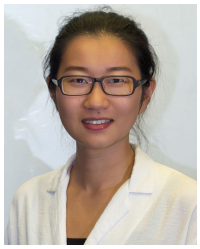
[39] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

Xiang He received the B.Sc. degree from Henan University of Science and Technology, Luoyang, China, in 2018. He is currently pursuing the Ph.D. degree from Shandong University, Jinan, China.

His research interests include deep learning and medical image analysis.

Weiye Song received the Ph.D. degree from Jilin University, Changchun, China, in 2010.

He is also a Postdoctoral Fellow with Boston University, Boston, MA, USA/Boston Medical Center, Boston, and Harvard University, Cambridge, MA, USA/Massachusetts General Hospital, Boston. He is currently a Distinguished Professor with the School of Mechanical Engineering, Shandong University. His research interests include opto electromechanical integrated medical equipment based on weak coherent optical imaging, spectroscopic analysis and other technologies, endoscopes, early medical diagnosis, and intraoperative navigation equipment.



Yiming Wang (Member, IEEE) received the Ph.D. degree in electronic engineering from the Queen Mary University of London, London, U.K., in 2018.

She is a Researcher with the Deep Visual Learning (DVL) Unit, Fondazione Bruno Kessler (FBK), Trento, Italy. She works on topics related to active robotic perception and 3-D scene understanding. She has organized a couple of workshops on related topics. She is actively serving as a reviewer for top-tier conferences and journals in both the computer vision and robotics domains.

Dr. Wang is a member of ELLIS.

Fabio Poiesi received the Ph.D. degree in electronic engineering and computer science from the Queen Mary University of London, London, U.K., in 2014.

He is the Head of the Technologies of Vision Laboratory, Fondazione Bruno Kessler, Trento, Italy. His research interests primarily include computer vision and deep learning, particularly on topics related to 2-D and 3-D scene understanding.

Dr. Poiesi has served as an Area Chair at conferences, such as ECCV and ICPR, and a Guest Editor for journals, including *ACM Transactions on Multimedia Computing, Communications, and Applications*. He is a member of ELLIS.

Ji Yi received the Ph.D. degree in biomedical engineering from Northwestern University, Evanston, IL, USA, in 2012.

He joined the Department of Biomedical Engineering, Johns Hopkins, Baltimore, MD, USA, as an Assistant Professor in September 2020. He mainly studies visible light optical coherence tomography (vis OCT) for high-resolution imaging, reliable retinal blood oxygen measurement, and spectroscopy. He also studies the interaction between vascular perfusion and other tissue components observed in a 3-D native environment.

Manishi Desai received the Medical degree from the School of Medicine, Indiana University, Indianapolis, IN, USA, in 2004.

She is an Assistant Professor of ophthalmology with Chobanian and Avedisian School of Medicine, Boston, MA, USA. She specializes in glaucoma. She went on to complete her ophthalmology residency with the University of Illinois at Chicago, Chicago, IL, USA, and then completed her fellowship in glaucoma with the Bascom Palmer Eye Institute, Miami, FL, USA, (affiliated with the University of Miami, Coral Gables, FL, USA). She is board-certified ophthalmologist with Boston University, Boston, affiliated Boston Medical Center, Boston. She currently serves as the Glaucoma Fellowship Director and the Director for international education. She has overseen the initiation of collaboration with the Lesotho Boston Health Alliance (LeBoHA) to work toward improving eye care. Her research interests include medical and surgical treatment of glaucoma, compliance, improving health care access/operations, and resident education.

Quanqing Xu received the B.Sc. degree from Shandong University of Science and Technology, Qingdao, China, in 2022. He is currently pursuing the master's degree with Shandong University, Jinan, China.

His research interests include artificial intelligence and super-resolution reconstruction.

Kongzheng Yang received the B.Sc. degree from Qingdao University, Qingdao, China, in 2022. He is currently pursuing the Ph.D. degree with Shandong University, Jinan, China.

His research interests include the application of OCT in the industrial field and embedded system development applications.

Yi Wan received the Ph.D. degree from Shandong University, Jinan, China, in 2006.

Since 2015, he has been the Dean and the Doctoral Supervisor of the School of Mechanical Engineering, Shandong University. He is a project letter evaluation expert from the National Natural Science Foundation of China, as well as a peer reviewer for domestic and foreign journals, such as *International Journal of Advanced Manufacturing Technology*, *Materials Science and Engineering C*, *Mechanical Systems and Signal Processing*, *Journal of South China University of Technology Natural Edition*, and *Journal of Southern Airlines*. His main research interests include deep learning, 3-D printing, and control of robot.