Wen Y, Wu YL, Bi L *et al.* A Transformer-assisted cascade learning network for choroidal vessel segmentation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(2): 286-304 Mar. 2024. DOI: 10.1007/s11390-024-3679-2

A Transformer-Assisted Cascade Learning Network for Choroidal Vessel Segmentation

Yang Wen[1] (温　阳), Yi-Lin Wu[1] (吴依林), Lei Bi[2] (毕　磊), Wu-Zhen Shi[1,*] (石武祯)
Xiao-Xiao Liu[3] (刘潇骁), Yu-Peng Xu[3] (许毓鹏), Xun Xu[3] (许　迅)
Wen-Ming Cao[1] (曹文明), *Senior Member, IEEE*, and David Dagan Feng[4] (冯大淦), *Fellow, IEEE*

[1] *Guangdong Provincial Key Laboratory of Intelligent Information Processing, School of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China*

[2] *Institute of Translational Medicine, Shanghai Jiao Tong University, Shanghai 200240, China*

[3] *Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China*

[4] *School of Information Technologies, The University of Sydney, Sydney 2006, Australia*

E-mail: wen_yang@szu.edu.cn; 2021280264@email.szu.edu.cn; lei.bi@sjtu.edu.cn; wzhshi@szu.edu.cn; kevinxu@sjtu.edu.cn drxuxun@sjtu.edu.cn; wmcao@szu.edu.cn; dagan.feng@sydney.edu.au

Received August 15, 2023; accepted January 20, 2024.

Abstract　As a highly vascular eye part, the choroid is crucial in various eye disease diagnoses. However, limited research has focused on the inner structure of the choroid due to the challenges in obtaining sufficient accurate label data, particularly for the choroidal vessels. Meanwhile, the existing direct choroidal vessel segmentation methods for the intelligent diagnosis of vascular assisted ophthalmic diseases are still unsatisfactory due to noise data, while the synergistic segmentation methods compromise vessel segmentation performance for the choroid layer segmentation tasks. Common cascaded structures grapple with error propagation during training. To address these challenges, we propose a cascade learning segmentation method for the inner vessel structures of the choroid in this paper. Specifically, we propose Transformer-Assisted Cascade Learning Network (TACLNet) for choroidal vessel segmentation, which comprises a two-stage training strategy: pre-training for choroid layer segmentation and joint training for choroid layer and choroidal vessel segmentation. We also enhance the skip connection structures by introducing a multi-scale subtraction connection module designated as MSC, capturing differential and detailed information simultaneously. Additionally, we implement an auxiliary Transformer branch named ATB to integrate global features into the segmentation process. Experimental results exhibit that our method achieves the state-of-the-art performance for choroidal vessel segmentation. Besides, we further validate the significant superiority of the proposed method for retinal fluid segmentation in optical coherence tomography (OCT) scans on a publicly available dataset. All these fully prove that our TACLNet contributes to the advancement of choroidal vessel segmentation and is of great significance for ophthalmic research and clinical application.

Keywords　choroidal vessel segmentation, optical coherence tomography (OCT), Transformer-assisted cascade learning, retinal fluid segmentation, multi-scale feature extraction

1　Introduction

The choroid is a vascular layer between the retina and the sclera, notable for being the most richly vascularized tissue in the human body. Due to its dense vasculature, it plays a crucial role in maintaining the health and function of the human eye, particularly the outer retina. The choroidal blood flow, one

Regular Paper
Special Section of CGI 2023
This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 62301330 and 62101346, the Guangdong Basic and Applied Basic Research Foundation under Grant Nos. 20231121103807001, 2022A1515110101, and the Guangdong Provincial Key Laboratory under Grant No. 2023B1212060076.

*Corresponding Author

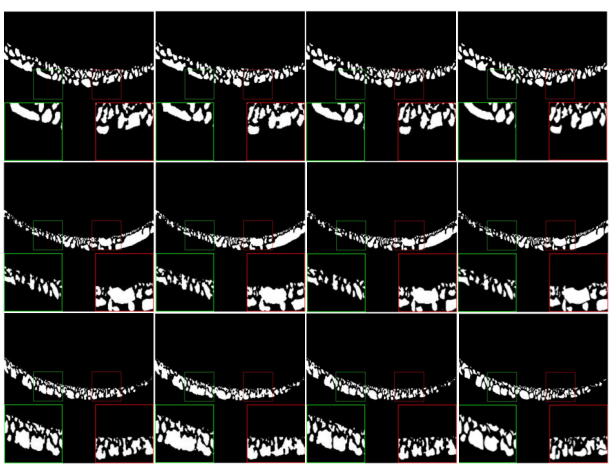
of the highest among organs, contributes to the thermal regulation of the retina. Intrinsic choroidal neurons modulate choroidal blood flow and control nonvascular smooth muscle cells within the choroid, especially behind the fovea[1].

Several studies used specific choroid biomarkers to assess the choroid quantitatively. These biomarkers[2], including features such as the choroidal thickness and vascularity index, are crucial in the diagnosis, prognosis, and treatment of a diverse range of ophthalmic diseases or pathological conditions. While quantifying biomarkers provides insight into the choroid's structure, these measures alone do not capture all the intricate details of the choroid layer and its complex vascular network. For instance, the shape, positioning, and branching patterns of individual choroidal vessels could offer valuable insights into revealing potential abnormalities related to specific ophthalmic conditions[3]. Thus, our study focuses on exploring an effective way to perceive the distribution of choroidal vessels.

In this paper, we concentrate on the segmentation of choroidal vessels in optical coherence tomography[4] (OCT) images. OCT is a non-invasive technique producing cross-sectional retina images, including the choroid, with broad applications in neurology, ophthalmology, gastroenterology, and cardiology. While earlier time-domain OCT (TD-OCT) struggled to image the choroid due to a low signal-to-noise ratio, newer versions like spectral-domain OCT (SD-OCT)[5], swept-source OCT (SS-OCT), and enhanced depth imaging OCT (EDI-OCT) improved in resolution and depth penetration, enhancing choroid imaging[6]. Although advances in the OCT technology significantly improve resolution and depth penetration, these advances do not fully address the complexity of choroidal imaging.

In OCT B-scans, the choroid presents three key challenges. First, the choroid lacks contrasting reflective properties, and the borders of its vessels often appear indistinct, making extracting discriminative features exceedingly difficult[7]. Second, the choroid layer is characterized by densely distributed vessels with irregular shapes, significantly increasing the complexity of their identification and segmentation. Lastly, the current method of delineating the choroid layer and choroidal vessels in OCT images primarily relies on manual annotation by experienced clinical professionals. This procedure is not only labor-intensive but also susceptible to potential inaccuracies. These challenges severely hinder us from training an efficient model for choroidal vessel segmentation.

In response to these challenges, recent methodologies grounded in deep learning have been gaining attention. Among the existing methods, direct vessel segmentation models[8] often grapple with noise and retinal shadows[9] in OCT scans, leading to suboptimal performance. In the synergistic method[10] a sharing encoder can significantly reduce the computing workload since it mainly extracts the commonalities of the choroid layer and choroidal vessels. However, for optimal results, the sharing encoder needs to concurrently and effectively extract the unique features of both tasks, including different boundaries and intricate details, which may not always align with the features common to both tasks. Thus, insufficient extraction of unique features for a specific task in the encoder could impose additional challenges on the specific and sharing decoders, leading to inferior segmentation results. For example, the synergistic method CUNet[10] shows over-segmentation between adjacent vessels in Fig.1(a). Meanwhile, in the common cascaded segmentation methods[11, 12], the choroidal vessel segmentation network (VSN) primarily relies on the intersection between the input OCT slice and the result of choroid layer segmentation backbone (LSB). Thus, the calculation error of the choroid layer segmentation backbone will affect the segmentation performance of the choroidal vessel network. For instance, ChoroidNET[11] unusually deviates from the



(a) (b) (c) (d)

Fig.1. Three comparative examples of the results of choroidal vessel segmentation by various methods. Each row is an individual OCT case. The green and red rectangles depict magnified images taken from two representative regions, emphasizing the enhanced performance of our TACLNet over methods CUNet and ChoroidNET in a fair comparison. (a) Result of CUNet[10]. (b) Result of ChoroidNET[11]. (c) Result of our TACLNet. (d) Ground truth.

ground truth (Fig.1(d)) in Fig.1(b), suggesting error propagation during the training phase. Nonetheless, compared with direct segmentation, the latter two methods provide a notable advantage by incorporating prior knowledge. Since all choroidal vessels reside between the Bruch's membrane (BM) and the choroid-sclera interface (CSI), integrating choroid layer segmentation within the workflow could exploit this significant relationship. Given the success of cascaded learning strategies in addressing noise and retinal shadows outside the choroid layer and their facilitative role for the following VSN, it is necessary to investigate a refined cascaded network for choroidal vessel segmentation.

In this study, we propose Transformer-Assisted Cascade Learning Network (TACLNet) for choroidal vessel segmentation. To address the issue of error propagation commonly encountered in cascaded structures, our method introduces a cascade pre-training strategy for the cascaded segmentation workflow, including pre-training and joint training stages. Additionally, we propose an auxiliary Transformer branch named ATB, which contains a sequence of shunted Transformer blocks[13]. These blocks obtain Key and Value of different sizes in self-attention, significantly enhancing the ability to capture the OCT slice's global and multi-scale features. The use of the ATB not only enhances the performance of our method by effectively capturing the global and multi-scale features but also significantly mitigates the impact of error propagation. Finally, acknowledging that existing cascaded and synergistic methods often overlook the importance of multi-scale feature extraction in dealing with the complexity of choroidal vessels, we introduce a multi-scale subtraction connection module named MSC. MSC adeptly negotiates the differences between shallow and deep convolution feature maps while preserving essential details during the segmentation phase.

Our main contributions can be summarized as follows.

• We propose Transformer-Assisted Cascade Learning Network (TACLNet) for choroidal vessel segmentation with a cascade pre-training strategy to train cascaded segmentation models more effectively. Vast experimental results validate the effectiveness and versatility of TACLNet and demonstrate its great potential in choroidal analysis.

• We present an auxiliary Transformer branch named ATB with the advantages of the shunted Transformer[13] to effectively utilize global information in the OCT slice to improve vessel segmentation performance.

• We introduce a novel multi-scale subtraction connection module named MSC, which can capture the differential information across multiple scales between feature maps while preserving the intricate details of local features for the choroidal vessel segmentation.

2 Related Work

2.1 Choroid Segmentation Method

2.1.1 *Conventional Methods for Choroid Segmentation*

Graph search algorithms play an extensive role in segmenting retinal layers in spectral-domain optical coherence tomography (SD-OCT). Zhang *et al.*[14] were the first to use the 3D graph search method for choroid surface detection. Chen *et al.*[15] used thresholding and graph's shortest path for choroid boundary and CSI detection, respectively. Hussian *et al.*[16] used Dijkstra's algorithm and a depth-based intensity normalization technique for layer segmentation while using a clustering method for vessel segmentation. However, it lacks robust testing. Despite reasonable results, these methods heavily rely on handcrafted features and are highly noise-sensitive.

2.1.2 *Deep-Learning Methods for Choroid Segmentation*

With the rise in interest in deep learning for medical image processing research, numerous deep learning models were newly developed specifically for segmenting the choroid layer and choroidal vessels.

End-to-End. Once trained, end-to-end structures are relatively straightforward, although they may not always achieve perfect results or require complex workarounds. Liu *et al.*[8] segmented the vessels using RefineNet[17]. Zheng *et al.*[18] segmented the choroid's upper and lower boundaries using ResUnet[19]. Zhu *et al.*[10] proposed a novel segmentation pipeline named CUNet for synergistically segmenting the choroid layer and vessels by treating these two tasks as a multi-task learning process. They employed a global encoder and global-specific decoders to manage the correlation and specifics of different tasks. They also proposed a new regularization term as an AMS loss func-

tion to prevent the model from favoring one specific task over the others. In the 3D choroidal vessel segmentation field, Huang *et al.*[20] introduced a shape-aware network that learns both the pixel and shape distributions of choroidal vessels, employing a relative distance map and a novel adversarial loss to optimize 3D choroidal vessel segmentation performance. Despite considerable efforts to develop effective straight out-of-the-box methods for vessel segmentation tasks, contemporary networks face challenges in simultaneously addressing the significant structural differences between the choroid layer and vessels.

Cascaded. The cascaded structure is commonly used in medical image segmentation tasks. It leverages prior knowledge and incorporates it into specific tasks, resulting in a multi-stage workflow. The output of each module in the cascaded structure is used as input for the subsequent module, allowing employing the prior knowledge. Chen *et al.*[21] implemented two SegNet[22] models to segment the Bruch's membrane (BM) and the Chorio-Scleral interface (CSI). Subsequently, they used a seam carving method to fill the area between the BM and CSI, achieving choroid layer segmentation. However, this final step's dependence on handcrafted features may limit the model's potential for adaptability and automation. Zhang *et al.*[23] developed a biomarker-infused global-to-local network (Bio-Net) for choroid layer segmentation. Bio-Net first employs a global multi-layer segmentation module to discern global structures. These are then concatenated with the original OCT images to serve as the input for the second stage's local LSB. Khaing *et al.*[11] proposed a two-stage cascaded pipeline named ChoroidNET. This deep-learning system initially employs a U-Net structure to segment the choroid layer. Once the choroid layer segmentation is complete, the intersection with the original image becomes the input for another U-Net designed for choroid vessel segmentation. Significantly, this work highlights the use of dilated convolution modules in both the choroid layer and vessel segmentation processes, marking it as the first to apply such a cascaded deep-learning approach for choroid vessel segmentation. Although these cascaded techniques are prone to error propagation, we draw inspiration from a pre-training strategy proposed by Bai *et al.*[24] to address this issue. As a result, we introduce a cascade pre-training strategy that effectively helps mitigate the existing problems related to error propagation.

2.2 Multi-Scale Feature Extraction

Techniques for multi-scale feature extraction primarily fall into inter-layer and intra-layer multi-scale methods. In inter-layer designs, a U-shaped architecture[25] aggregates different scale features extracted from high level to low level during decoding. Intra-layer methods[26, 27] often employ parallel multi-branch convolution layers to generate a range of receptive fields. Combining these two designs, M^{2}S-Net[28]'s Multi-Scale in Multi-Scale Subtraction Module (MMSM) generates a rich and complementary multi-scale feature set across different levels. MMSM achieves this by defining a multi-scale subtraction unit (MSU) that combines the feature maps of different filter sizes and then aggregates these features to generate a complementarity enhanced feature.

For our work, we utilize these principles by extracting differential data from the features of adjacent convolution layers at the inter-layer level, and we improve upon the MSU module's utility by preserving skip connection concatenations at the intra-layer level. This method allows us to maintain detailed specifics in the feature maps while obtaining multi-scale differential information within the same feature map layer.

2.3 Adaptation of Transformer Variants for Image Segmentation

The use of auxiliary branches in semantic segmentation tasks is quite prevalent. Li *et al.*[29] and Xu *et al.*[30] both introduced auxiliary task branches to assist with the main task. Ding *et al.*[31] introduced lightweight Transformers as an auxiliary branch to enhance the global context of image features.

Among multiple variants[13, 32, 33] of the Vision Transformer[34], a notable variant is shunted Transformer[13], which employs the shunted self-attention mechanism. This technique reduces the spatial scales of K and V to multiple sizes, not only reducing the computational cost but also obtaining multi-scale feature capturing ability. Thus, the shunted Transformer has the potential for comprehensive global information capture.

In the context of OCT scan choroidal vessel segmentation, it is critical to capture global information and preserve original structural information to guide and rectify the vessel segmentation process. Additionally, the choroidal vessels exhibit a multi-scale nature, making capturing features at multi-scale valuable. To

this end, we incorporate a sequence of shunted Transformer blocks into our pipeline as an auxiliary attention branch, adapting it for our TACLNet. This adaptation enables it to rectify potential errors and improve the overall performance of choroidal vessel segmentation.

3 Methods

3.1 Problem Definition

To provide a theoretical basis for our approach, we consider the i-th B-scan input, denoted as X_i, which is a part of a training set $\mathbb{X}$. For the tasks of choroid layer segmentation T^1 and choroidal vessel segmentation T^2, we define the output distributions of this B-scan input as P_i^1 and P_i^2, respectively. The training set, constructed using n B-scans, is represented by $\mathbb{X} = \{X_1, X_2, \ldots, X_n\}$. Each B-scan in $\mathbb{X}$ is associated with a set of choroid layer labels $\mathbb{Y}^1 = \{Y_1^1, Y_2^1, \ldots, Y_n^1\}$ and choroidal vessel labels $\mathbb{Y}^2 = \{Y_1^2, Y_2^2, \ldots, Y_n^2\}$. We represent the training loss term $\mathcal{L}^1(P_i^1, Y_i^1)$ for the choroid layer and loss term $\mathcal{L}^2(P_i^2, Y_i^2)$ for the vessel, where i represents the current slice number.

Given that all choroidal vessels are located within the choroid layer, it is logical to follow the task of choroid layer segmentation for the segmentation of these vessels. Thus, we employ two distinct deep convolutional neural networks (DCNNs) for two different tasks. Each of the DCNNs consists of a feature extractor $f(\cdot)$ to catch the discriminate feature and a pixel-wise classifier $g(\cdot)$ to obtain the pixel-wise classification probabilities. Both $f(\cdot)$ and $g(\cdot)$ are parametric functions that can be approximated by DCNNs, with learnable parameters denoted as W_{f}^j for $f(\cdot)$ and W_{c}^j for $g(\cdot)$, respectively. To enhance clarity, we consolidate the feature extraction and pixel-wise classification tasks into a unified composite function, denoted as $h(\cdot) = g \circ f$. The optimization objectives of the two DCNNs can be formulated as follows:

$$\mathbb{W}^1 = \underset{W_{\mathrm{f}}, W_{\mathrm{c}}}{\operatorname{argmin}} \sum_i \mathcal{L}^1_{\text{pretrain/joint}} \left(h\left(X_i\right), Y_i^1 \right),$$

$$\mathbb{W}^2 = \underset{W_{\mathrm{f}}, W_{\mathrm{c}}}{\operatorname{argmin}} \sum_i \mathcal{L}^2_{\text{joint}} \left(h\left((1-\alpha)(P_i^1 \cap X_i) + \alpha X_i\right), Y_i^2 \right),$$

$$\text{s.t. } \mathbb{W}^1 = \{W_{\mathrm{f}}^1, W_{\mathrm{c}}^1\}, \mathbb{W}^2 = \{W_{\mathrm{f}}^2, W_{\mathrm{c}}^2\},$$

$$\alpha \in [0, 1], P_i^1 = h(X_i | \mathbb{W}^1),$$

where $\mathbb{W}^1$ and $\mathbb{W}^2$ denote the optimal parameter spaces for T^1 and T^2, respectively. The configuration jointly combines two distinct optimization procedures into a singular process. In the second step, the input is derived from the intersection of X_i and the first step's choroid layer result, P_i^1, which is then scaled by $(1-\alpha)$. We add αX_i, representing a portion of the original information determined by α, supplementing the first output. To prevent an undue reliance on the output from the first segmentation module, which could potentially provide an erroneous prior, and to preserve crucial global structures that the modules can identify, we set α empirically to a modest constant, typically 0.001.

The choroid layer can be used as a practical preliminary guide, offering an approximate baseline that aids in accurately pinpointing the location of choroidal vessels within the B-scan. However, the variance in difficulty between the two tasks (relatively simple choroid layer segmentation and more complex vessel segmentation) can lead to error propagation. Thus, we introduce a novel training strategy and ATB to counteract such error propagation, as shown in Fig.2.

3.2 Network Design

As illustrated in Fig.2, our proposed network incorporates two integral components: the choroid layer segmentation backbone (LSB) and the choroidal vessel segmentation network (VSN). We utilize M^{2}S-Net[28] as a "plug and play" LSB, integrating it into our TACLNet without any modifications. Its sole function within our model is to provide the essential choroid layer segmentation needed for the subsequent vessel segmentation task carried out by VSN. The efficacy of M^{2}S-Net is corroborated through our ablation study on LSB, as evidenced in Table 1. Within VSN, we employ various strategies, including using the MSC and deriving five levels of features. Each encoder block in VSN (excluding the first one) separately applies up-sampling and a 3×3 convolution on the feature maps to ensure that the channel quantity matches the preceding encoder block's output. The resulting feature and the previous encoder block's output feature are seamlessly integrated into our MSC. Each MSC generates a complementary differential feature, individually serving as a skip connection in alignment with the feature maps of the decoder pathway, following the traditional U-Net skip connection scheme. Moreover, we incorporate ATB within our VSN to augment the global feature compensations in OCT B-scans. ATB is essential, considering the input of VSN comprises the intersection between the

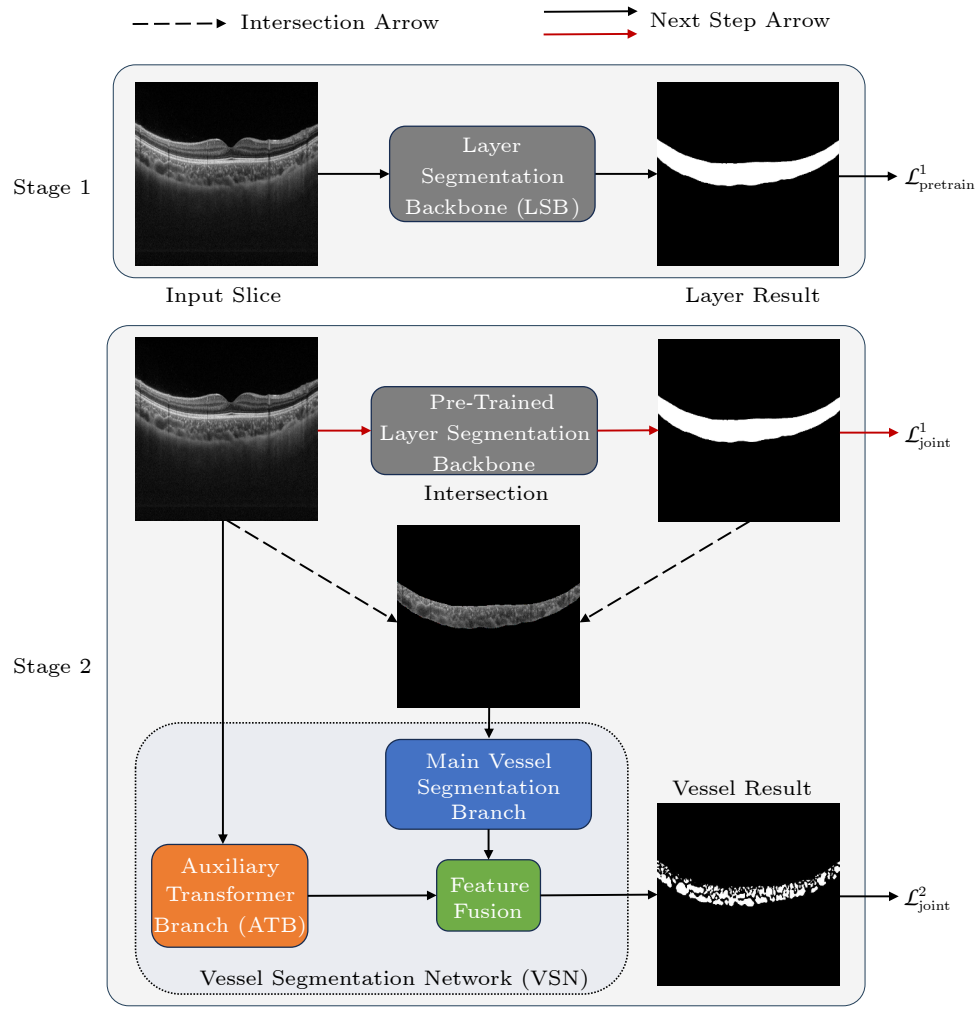


Fig.2. Outline of our proposed method, which consists of two stages. In stage 1, we pre-train LSB. In stage 2, we perform joint training of LSB and VSN. As indicated by the red arrows, the pre-trained LSB first segments the choroid layer result. We then intersect the choroid layer result with the B-scan input, providing input to the main vessel segmentation branch. Meanwhile, the B-scan input also goes through ATB, providing valuable global features to the main vessel segmentation branch through feature fusion.

Table 1. Ablation Study Results for LSB: Mean(%) (Standard Deviation(%)) of 5-Fold Cross-Validation

Method	Vessel	Layer				
	IoU	ACC	IoU	DSC	SE	PC
Attention U-Net[35]	72.30 (4.46)	99.50 (0.10)	93.66 (1.53)	96.70 (0.84)	97.07 (0.45)	96.40 (1.28)
U-Net[36]	73.30 (4.76)	99.51 (0.11)	93.70 (1.79)	96.72 (0.99)	97.03 (0.90)	96.50 (1.21)
CVI-Net[12]	73.24 (4.18)	99.51 (0.04)	93.72 (0.88)	96.75 (0.47)	96.98 (0.91)	96.57 (0.44)
M^2S-Net[28]	**75.17 (4.77)**	99.62 (0.07)	95.00 (0.98)	97.43 (0.52)	**97.82 (0.24)**	97.08 (0.75)
ChoroidNET[11]	73.55 (4.71)	**99.63 (0.08)**	**95.06 (1.26)**	**97.45 (0.67)**	97.57 (0.41)	**97.37 (1.46)**

Note: We use Attention U-Net[35], U-Net[36], CVI-Net[12], M^2S-Net[28], and ChoroidNET[11] as LSB, respectively. For choroidal vessels, we use our TACLNet's VSN.

LSB's binary result and the input B-Scan. LSB has an inherent potential for erroneous segmentation outcomes, such as over-segmentation and under-segmentation, necessitating the incorporation of ATB. In subsequent sections, we will elucidate the functional principles of both MSC and ATB. The comprehensive pipeline of our proposed TACLNet is delineated in Fig.3.

3.3 MSC

Drawing inspiration from the multi-scale in multi-scale subtraction module (MMSM)[28], we integrate the multi-scale Subtraction Unit (MSU)[28] into our skip connection structure within our VSN, but we altered the implement usage of MSU. This adjustment was primarily made due to the distinct characteris-

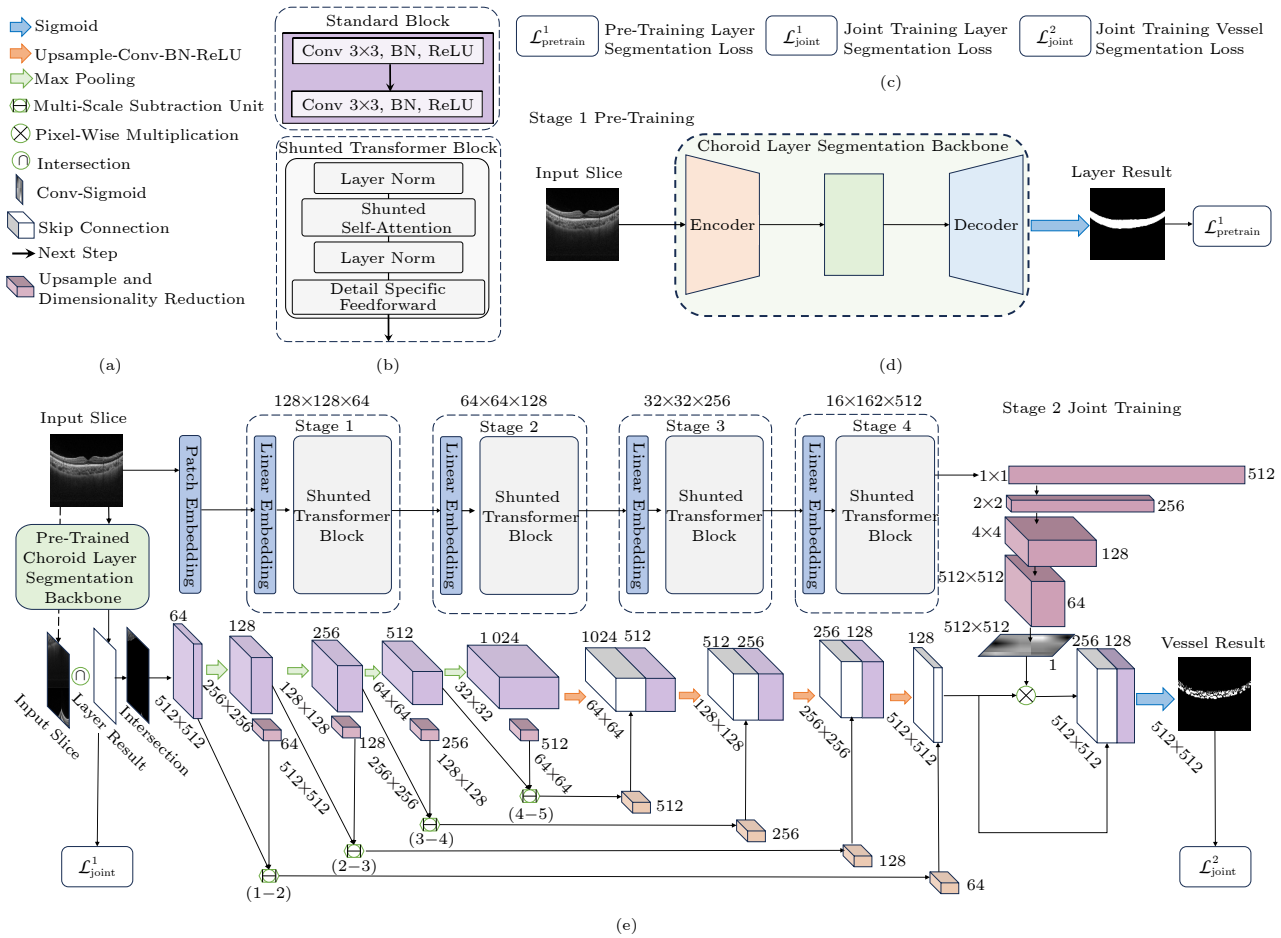


Fig.3. (a) Overview of computational operations and elements. (b) Structure of standard convolution block and shunted Transformer block[13]. (c) Descriptions of loss function. (d) Pre-training the LSB. (e) Implementing a dual-path approach to segment the choroidal vessel from the input slice. The second stage consists of a pre-trained LSB and a VSN to form a cascade learning network for fine vessel segmentation. We calculate the difference of features of adjacent layers to obtain multi-scale features. For instance, the notation "(1–2)" denotes the extraction of differential features between the feature maps of convolution layers 1 and 2. An ATB is applied to generate a global Sigmoid "attention" map. This map is then pixel-wise multiplied by the output features of the main branch to adaptively focus on important features. The weighted features are then concatenated with the original features, aggregating features from both paths. Finally, a 1×1 convolution and Sigmoid operation yield the vessel segmentation result.

tics of choroidal vessels, which are smaller and denser than the choroid layer.

While the standard MMSM excels at acquiring multi-scale differential information, it significantly sacrifices contextual data. This is mainly because of the initial dimension reduction from 64, 256, 512, 1 024, 2 048 to 64. As the encoder level increases, the context loss becomes more severe. Additionally, using a skip connection by adding complementary features to the decoder path instead of concatenation potentially contributes to the loss of detailed local data from the encoder side.

We introduce a multi-scale subtraction connection module named MSC to address these challenges. This bespoke solution is designed to segment vessels and other small objects. First, we extract five feature maps in the encoder path, denoted as $F_i, i \in \{1, 2, 3, 4, 5\}$. In particular, F_{i+1} $(i \neq 0)$ results from the convolution layer acting on F_i. Each F_{i+1} $(i \neq 0)$ is upsampled via bilinear interpolation and dimensionally reduced to align with the previous encoder block's feature F_i. These two features are then processed through MSU and a 3×3 convolution operation to yield multi-scale differential data, which can be formulated as:

$$M_{i+1} = Conv(MSU(F_i, Up(F_{i+1}))),$$

where $Up(\cdot)$ represents the upsample operation. We omit the Batch normalization (BN) and ReLU activation after $Conv(\cdot)$ for conciseness. Finally, this convoluted differential feature map is concatenated with the decoder path's feature D_j, which generates D'_j. Consequently, we can capture multi-scale differential information through MSU while preserving detailed

spatial hierarchical data via channel-wise concatenation, thereby achieving precise localizations. The depiction of MSC is illustrated in Fig.4. The second step of MSC can be formulated as follows:

$$D'_j = Concat(D_j, M_{i+1}),$$
$$\text{s.t. } 1 \leqslant i \leqslant (d-1), j = d - i,$$

where D_j is the feature map at the j-th level in the decoder path, and d is the total number of layers in the encoder. We use the variables i and j to index the encoder and decoder layers, respectively, with the constraint that $i + j = d$. $Concat(\cdot, \cdot)$ is concatenating two feature maps along the channel dimension.

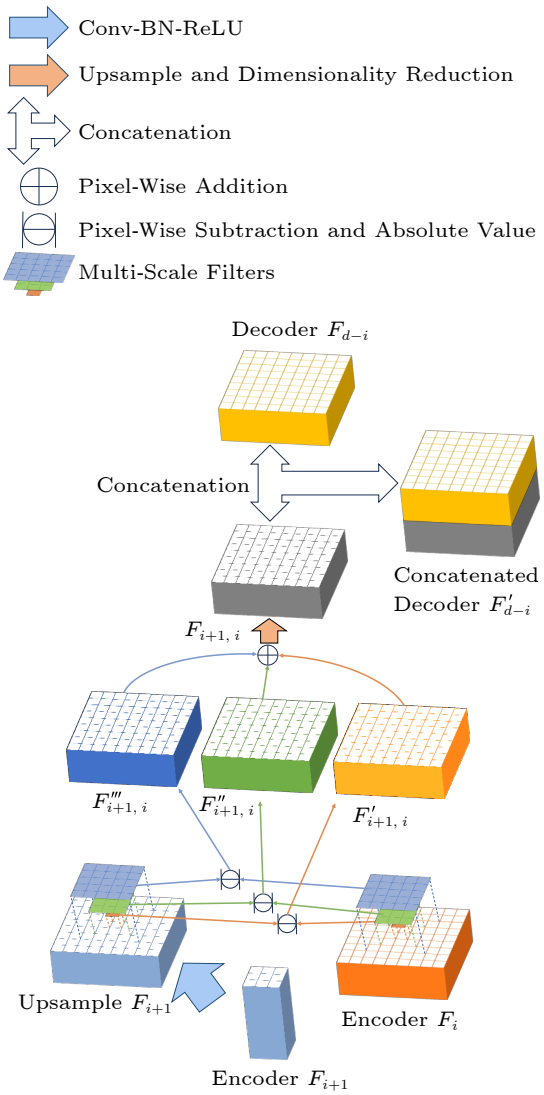


Fig.4. Outline of MSC. d denotes the total number of layers in the encoder, and i denotes the current layer.

3.4 ATB

Besides the main U-Net architecture, we introduce an auxiliary Transformer branch named ATB as a rectifier for potentially inaccurate prior information LSB might provide for VSN. This auxiliary branch comprises a four-stage sequence of shunted Transformer blocks and a feature fusion operation. Firstly, the original B-scan input is patch embedded to obtain a more informative token sequence with the length of $(H/4) \times (W/4)$ and the token dimension of C. There are four stages containing several shunted Transformer blocks[13]. In each stage, each block outputs feature maps of the same size. In the linear embedding step, we utilize a convolution layer with a stride of 2 to connect different stages. Before feeding the feature maps into the next stage, the size of the feature maps will be halved, but their dimensions will be doubled. Therefore, we have four feature maps, F_1, F_2, F_3, F_4, which are the output of each stage. And the F_i size is $(H/2^{i+1}) \times (W/2^{i+1}) \times (C \times 2^{i-1})$. The final stage's F_4 is up-sampled to align with the feature map's size from the last decoder block of VSN. A 1×1 convolution and Sigmoid activation are applied, resulting in a global attention map. This map is multiplied with the last decoder block's feature, enabling concentration on pivotal information. Concurrently, a skip connection is implemented from the last decoder block to the filtered feature map, thus ensuring the preservation of integral global knowledge. Finally, the conclusive feature map is established via a 1×1 convolution and Sigmoid activation, generating the ultimate vessel prediction mask. The implementation of our ATB can be formulated as follows:

$$S_x = \text{Sigmoid}(Up(ST(x))),$$

where the input x is first processed through the shunted Transformer, denoted as $ST(\cdot)$, then up-sampled, matching the original size of x, after a Sigmoid activation, resulting in a global attention map S_x. This map is then pixel-wisely multiplied with the main branch's last decoder feature map, F_{main}, resulting in a newly formed feature map, denoted as $F'_{\text{main}} = S_x \otimes F_{\text{main}}$. For a more efficient feature fusion, we concatenate the original F_{main} with F'_{main}, followed by applying a standard convolution block, forming F'''_{main}, which is now primed for pixel-wise classification of vessels. The feature fusion process can be formulated as:

$$F'''_{\text{main}} = Conv(Concat(F'_{\text{main}}, F_{\text{main}})).$$

3.5 Training Strategy

Inspired by the pre-training strategy proposed by

Bai *et al.*[24], to train the proposed network for better vessel segmentation, we develop a cascade pre-training strategy, mainly containing pre-training of LSB and joint training. The cascade pre-training strategy ensures that LSB and VSN can play their expected roles.

Pre-Training of LSB. The goal of LSB is to first segment the choroid layer from the OCT images, which serves as a solid prior for the VSN, as all choroidal vessels are resident in the choroid layer. The loss function we employ in this pre-training phase is designed to build a layer loss. We utilize the following equation:

$$\mathcal{L}_{\text{layer}} = \mathcal{L}_{\text{wBCE}} + \mathcal{L}_{\text{DICE}},$$

where $\mathcal{L}_{\text{wBCE}}$ and $\mathcal{L}_{\text{DICE}}$ signify the weighted binary cross-entropy (BCE) loss and Dice loss, respectively. The term $\mathcal{L}_{\text{wBCE}}$[37], a refined variant derived from the original $\mathcal{L}_{\text{BCE}}$, focuses more on hard pixels than its original form. In our layer loss function, while the $\mathcal{L}_{\text{wBCE}}$ enhances the pixel-level classification, the Dice loss, $\mathcal{L}_{\text{DICE}}$, is advantageous in refining the delineation of boundaries. Thus, we obtain $\mathcal{L}^1_{\text{pretrain}} = \mathcal{L}_{\text{layer}}$.

Joint Training. After pre-training LSB, we jointly train the LSB and VSN. At this step, the choroid loss is the same as in pre-training: $\mathcal{L}^1_{\text{joint}} = \mathcal{L}_{\text{layer}}$. We further introduce a vessel loss. Initially, we extract the vessel region based on the prediction from the VSN. Assuming a single OCT slice x as the input slice, we denote the output of the VSN as $o^{W\times H} = f(x')$ for simplicity, where x' is the intersection of LSB's layer result with the input slice x. Here, H and W represent the height and width of the input images, respectively. The process of extracting the predicted vessel region can be formulated as follows:

$$H_x = \text{Histogram}(T(\text{Sigmoid}(o), \lambda) \cap x),$$

where the function $T(a, \lambda)$ applies a threshold λ to a given input a, mapping all values $\geqslant \lambda$ to 1 and all values $< \lambda$ to 0. Since the Sigmoid activation is applied to the output o, which produces a probability ranging from 0 to 1, the threshold λ is typically set to 0.5, with scores exceeding this value interpreted as positive results indicative of the presence of a choroidal vessel. Then, we apply the threshold function T to generate a binary mask. This mask then intersects with the OCT slice x to get the vessel region. A histogram is computed from the predicted vessel region, resulting in H_x. The normalized and logarithmic form is represented as $H'_x = \log(H_x/(\sum H_x))$.

Similarly, the intensity probability of the ground-truth vessel region's histogram can be obtained by:

$$H_y = \text{Histogram}(\text{GT} \cap \text{x}).$$

The normalized H_y is formed in $H'_y = H_y/(\sum H_y)$. Subsequently, the vessel loss can be articulated as follows:

$$\mathcal{L}_{\text{vessel}} = \mathcal{L}_{\text{wBCE}} + \mathcal{L}_{\text{DICE}} + \lambda \mathcal{D}_{\text{KL}}(H'_x || H'_y),$$

where $\mathcal{L}_{\text{wBCE}}$, $\mathcal{L}_{\text{DICE}}$, and $\mathcal{D}_{\text{KL}}(H'_x||H'_y)$ represent the weighted BCE loss, Dice loss, and the Kullback-Leibler divergence (KL-divergence)[38], respectively. $\mathcal{D}_{\text{KL}}(H'_x||H'_y)$ measures the similarity between the predicted and the ground-truth vessel regions in the feature space. The vessel loss $\mathcal{L}_{\text{vessel}}$ is a linear combination of $\mathcal{L}_{\text{wBCE}}$, $\mathcal{L}_{\text{DICE}}$, and $\lambda \mathcal{D}_{\text{KL}}(H'_x||H'_y)$. Specifically, the two aforementioned loss partitions are optimized to encourage the segmentation results to align closely with the ground truth regarding spatial distribution, and $\lambda \mathcal{D}_{\text{KL}}(H'_x||H'_y)$ is introduced as a constraint regularization on the intensity distribution. Acknowledging the noticeable intensity contrast between the choroidal stroma and vessels, without accurate segmentation, the intensity discrepancy between the model's predictions and the ground truth could be considerable. We assign a value of $\lambda = 70$ to ensure the effectiveness of this regularization term. Thus, we obtain $\mathcal{L}^2_{\text{joint}} = \mathcal{L}_{\text{vessel}}$.

4 Experiments

4.1 Dataset

Our study uses images sourced from SS-OCT (model DRI OCT-1 Atlantis; Topcon). These images are produced using a radial scanning pattern of 12 lines, offering a resolution of 1 024 × 12. Each image represents an average of 32 consecutive, overlapped scans centered around the fovea, yielding a resolution of 1 024 × 992. This equates to a physical area of 12 mm × 2.6 mm. And when translated into a magnification ratio, it equates to 12.00 × 2.60/(1 024 × 992). For this research[①], we randomly select a total

[①]The research procedures were given the green light by the Institutional Review Board of Shanghai General Hospital, Shanghai Jiao Tong University, and complied with the stipulations of the Declaration of Helsinki. We ensure that all participants give informed consent before their involvement.

of 10 subjects, which are specifically made up of five emmetropes and five high myopes (refractive error $\leqslant -5.0$). In total, we conduct 120 OCT scans. Further in the process, two seasoned physicians separately annotate the choroid layer and choroidal vessel on each OCT slice and then jointly cross-verify them. To evaluate the effectiveness of our proposed method, in our research, we utilize a 5-fold cross-validation methodology to assess the robustness of our model. Our dataset comprises ten studies, each containing images and annotated regions. In each validation iteration, we utilize eight studies, including 96 images and approximately 8 740 annotated regions, as training data. These studies are employed to educate our model and adjust its parameters. The remaining two studies, consisting of 24 images and approximately 2 190 annotated regions, are used as testing data to evaluate the model's performance. This selection process is performed five times, with each iteration encompassing a unique combination of studies for training and testing.

To fully demonstrate the generalization of TACLNet in various medical tasks, we conduct a validation using the RETOUCH[39] dataset. The RETOUCH dataset comprises three training sets obtained from different OCT devices, totaling 70 volumes. Specifically, there are 24 volumes acquired with Cirrus (Zeiss), 24 volumes acquired with Spectralis (Heidelberg), and 22 volumes acquired with T-1000 and T-2000 (Topcon). Each volume from these devices contains a different number of B-scans: 128, 49, and 128, respectively. The resolutions of these scans are $512\times1\,024$, 512×496, and 512×885, respectively. In the RETOUCH dataset, three distinct fluid types, namely intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detachment (PED), are manually labeled and provided as ground truth. The RETOUCH organizers assess the submitted results from various research teams. As a result, the ground truth of the RETOUCH test data remains undisclosed to the public. We employ a 3-fold cross-validation strategy in our experiments on the RETOUCH training dataset, splitting the data at the case level. Like the choroidal vessel segmentation task, we label the retina layer labels for training. Specifically, we randomly select 711 slices from SD-OCT scans across 24 volumes acquired with the Spectralis device (Heidelberg) in the RETOUCH dataset. These slices originally contain official valid fluid labels. Since our TACLNet is cascaded, the retinal layer in 711 SD-OCT slices is further annotated and verified by two experienced doctors.

4.2 Experimental Setting

The whole framework is built on PyTorch 1.12.1. In the choroidal vessel segmentation, due to the limited GPU memory, we reduce the size of the input image instead of using the full resolution. Thus, the input size is resized to 512×512 without data augmentation. In the RETOUCH dataset's fluid segmentation task, the input size is resized to 256×256 with no data augmentation. We optimize the network with an SGD optimizer on randomly drawn OCT samples from the dataset. For the hyper-parameters, the channel numbers 64, 128, 256, 512, 1 024 are set for the five encoder stages of VSN's U-Net backbone, respectively, while the setting of decoder stages is symmetrical. In the first training step, the learning rates for LSB are set to 5.0×10^{-2}, and in the second step, the learning rates for the LSB are set to 5.0×10^{-3}, and the learning rates for VSN and ATB are set to 9.0×10^{-3}. The momentum and weight decay are set as 0.9 and 5.0×10^{-4}, respectively. For the Retinal Fluid Segment task, we replace $\mathcal{L}_{\mathrm{DICE}}$ in $\mathcal{L}_{\mathrm{vessel}}$ with $\mathcal{L}_{\mathrm{Cross-Entropy}}$ due to the multi-class nature of this task. The second difference is the number of epochs due to different convergence speeds. Specifically, the training epoch settings for different stages are as follows: LSB undergoes a pre-training phase of 25 epochs. Following this, the joint training of both LSB and VSN is conducted for another 25 epochs. For the RETOUCH dataset, the fluid segmentation is trained for 30 epochs. The network's training time is 1 hour, and the inference time is 7.116 seconds per validation fold with 24 slices. During inference, we do not use any post-processing operations. The accuracy (ACC), intersection over union (IoU), dice similarity coefficient (DSC), sensitivity (SE), and precision (PC) are used to evaluate our method's effectiveness in choroid layer segmentation, choroidal vessel segmentation, and retinal fluid segmentation. The metrics' mean value and standard deviation on the five splits (three splits in fluid segmentation) are reported to evaluate different methods. This validation strategy is adopted to guarantee the statistical significance of the experimental results.

4.3 Comparison with State-of-the-Art Methods

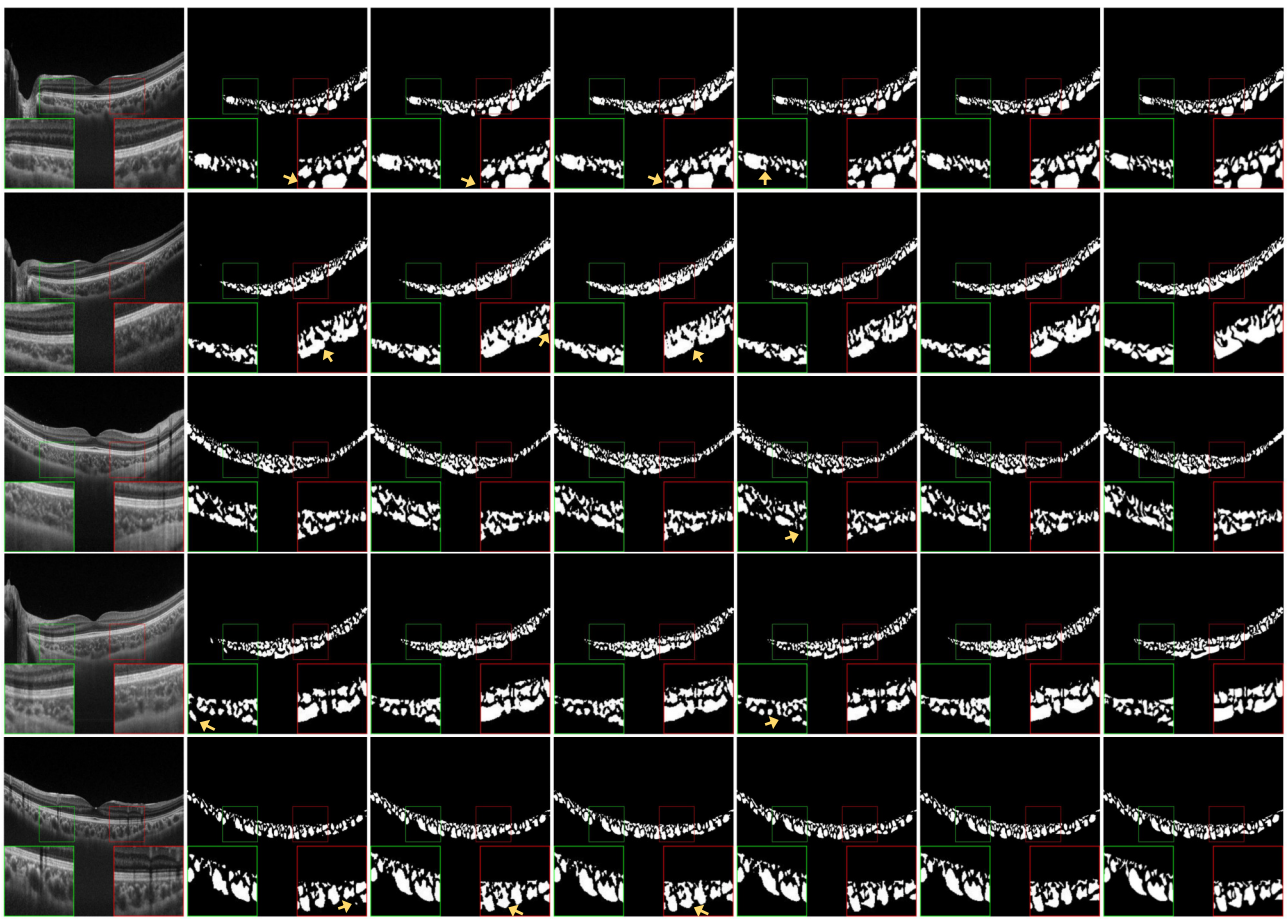
We compare TACLNet with state-of-the-art choroidal vessel segmentation methods and current

general medical segmentation networks. In our study, we carefully compare the results produced by our network and four established choroidal vessel segmentation methods (RefineNet[8, 17], ChoroidNET[11], CVI-Net[12], CUNet[10]), as well as the baseline models of U-Net[36] and Attention U-Net[35]. Through both qualitative and quantitative evaluation of the segmented images relative to their corresponding ground truths, our proposed approach yields promising results. Fig.5 provides illustrative examples of vessel segmentation outputs. A detailed comparison of choroidal vessel segmentation performance across all tested networks is presented in Table 2. The experimental results demonstrate that all compared methods except RefineNet improve the vessel segmentation baseline. Our TACLNet outperforms the other models in terms of vessel segmentation performance, comprising the highest ACC (99.27 ± 0.17), as well as IoU (76.26 ± 5.82), SE (87.55 ± 2.91), and PC (85.46 ± 5.16).

4.4 Ablation Experiments

To explore the contribution of each part of TACLNet, we investigate the impact of our cascade pre-training strategy, MSC, and ATB. The ablation models are trained and validated using the same training and test sets using the same LSB network (M^2S-Net). Fig.6 shows the choroidal vessel segmentation results for all ablation models. Table 3 compares the performance of TACLNet's ablation models. Every model, except for Ablation-5, uses a non-cascade strategy.

Effect of MSC. As shown in Fig.6, Ablation-1's sole U-Net hardly segments the vessel with appropriate extension, shape, and boundary. With the MSC added in the skip connection, the segmented vessel with different scales gains a better extension and shape, which improves in avoiding unexpected vacant holes in vessels. As a result, Ablation-2 improves the



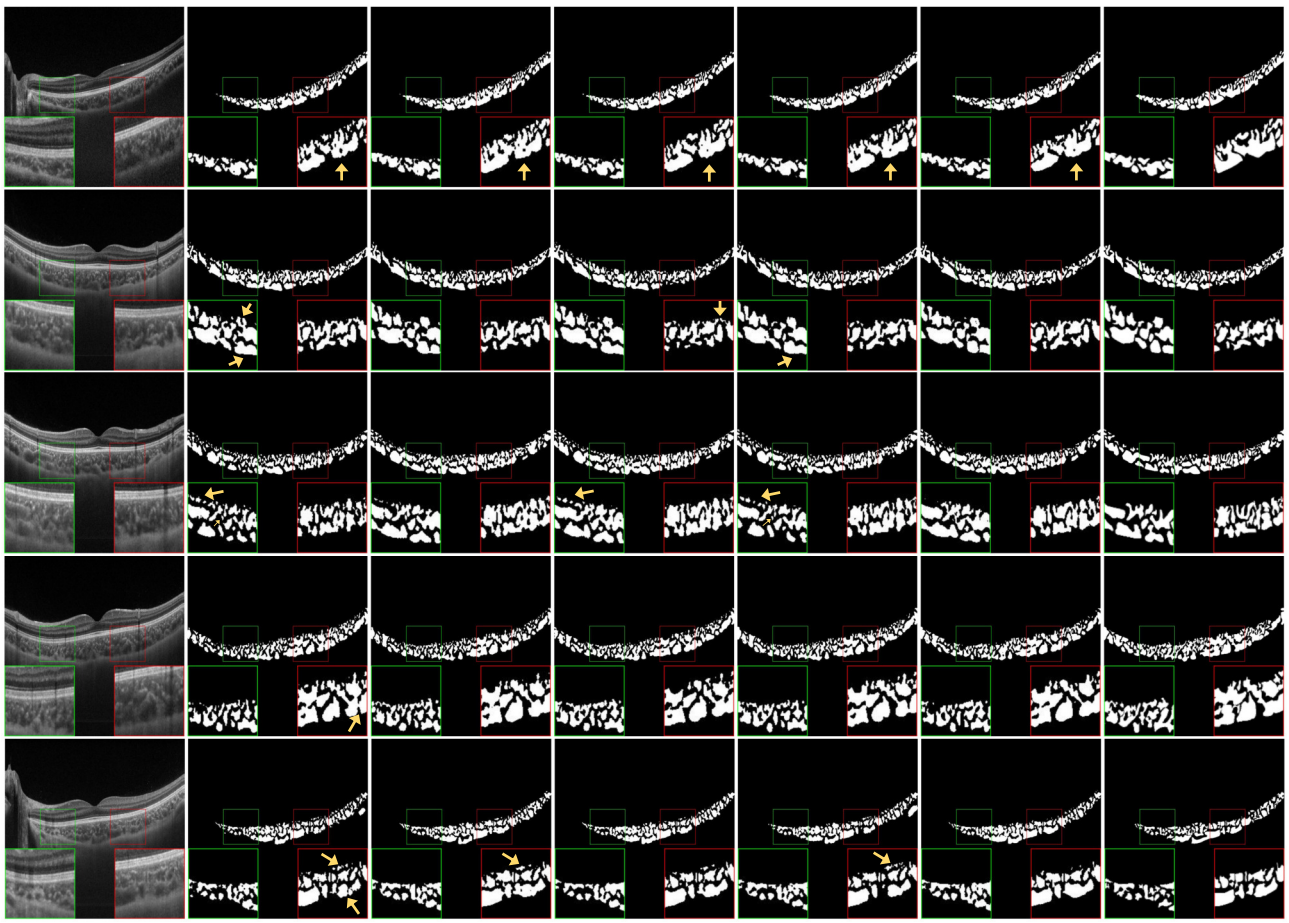
(a) (b) (c) (d) (e) (f) (g)

Fig.5. Visualized choroidal vessel segmentation comparisons of various methods. The five rows exhibit the visualized segmentation results of five individual OCT cases with different segmentation methods. The yellow arrows indicate the inaccuracies of the segmentation methods. The green and red rectangles depict magnified images taken from two representative regions. (a) Input slice. (b) Attention U-Net[35]. (c) ChoroidNET[11]. (d) CVI-Net[12]. (e) CUNet[10]. (f) Ours. (g) Ground truth.

Table 2. Segmentation Result Comparisons Between Various Methods on Clinical Choroidal Vessel Dataset: Mean(%) (Standard Deviation(%)) of 5-Fold Cross-Validation

Method	ACC	IoU	DSC	SE	PC
RefineNet[8, 17]	98.34(0.36)	54.07(7.08)	69.55(6.35)	78.79(2.67)	63.75(8.25)
U-Net[36]	98.91(0.14)	66.87(3.96)	79.74(3.14)	84.25(2.16)	78.47(5.92)
Attention U-Net[35]	98.94(0.12)	67.60(3.88)	80.29(3.04)	85.15(1.67)	78.36(5.09)
ChoroidNET[11]	99.01(0.18)	70.13(4.19)	82.22(3.02)	86.53(1.33)	78.91(4.97)
CVI-Net[12]	99.02(0.18)	70.35(4.50)	82.37(3.24)	86.91(1.69)	78.91(4.95)
CUNet[10]	99.18(0.21)	72.91(5.26)	84.09(3.61)	83.55(2.53)	85.15(5.46)
TACLNet (ours)	**99.27(0.17)**	**76.26(5.82)**	**86.34(3.88)**	**87.55(2.91)**	**85.46(5.16)**

Note: The highest mean results in this table are highlighted in bold. This format is consistent in all the following tables.



(a) (b) (c) (d) (e) (f) (g)

Fig.6. In the ablation study, samples of choroidal vessel segmentation using various VSNs are visualized. The five rows exhibit the visualized segmentation results of five individual OCT cases with varying VSNs. (a) Input slice. (b) Result of U-Net[36], as shown in Ablation-1. (c) Result of adding the MSC, as shown in Ablation-2. (d) Result of adding the ATB, as shown in Ablation-3. (e) Result of adding both MSC and ATB, as shown in Ablation-4. (f) Result of adding MSC, ATB, and cascade pre-training strategy, as shown in Ablation-5. (g) Ground truth. The yellow arrows indicate the inaccuracies of the segmentation network. The green and red rectangles depict magnified images taken from two representative regions.

Table 3. Ablation Study Results for Our Proposed TACLNet: Mean(%) (Standard Deviation(%)) of 5-fold Cross-Validation

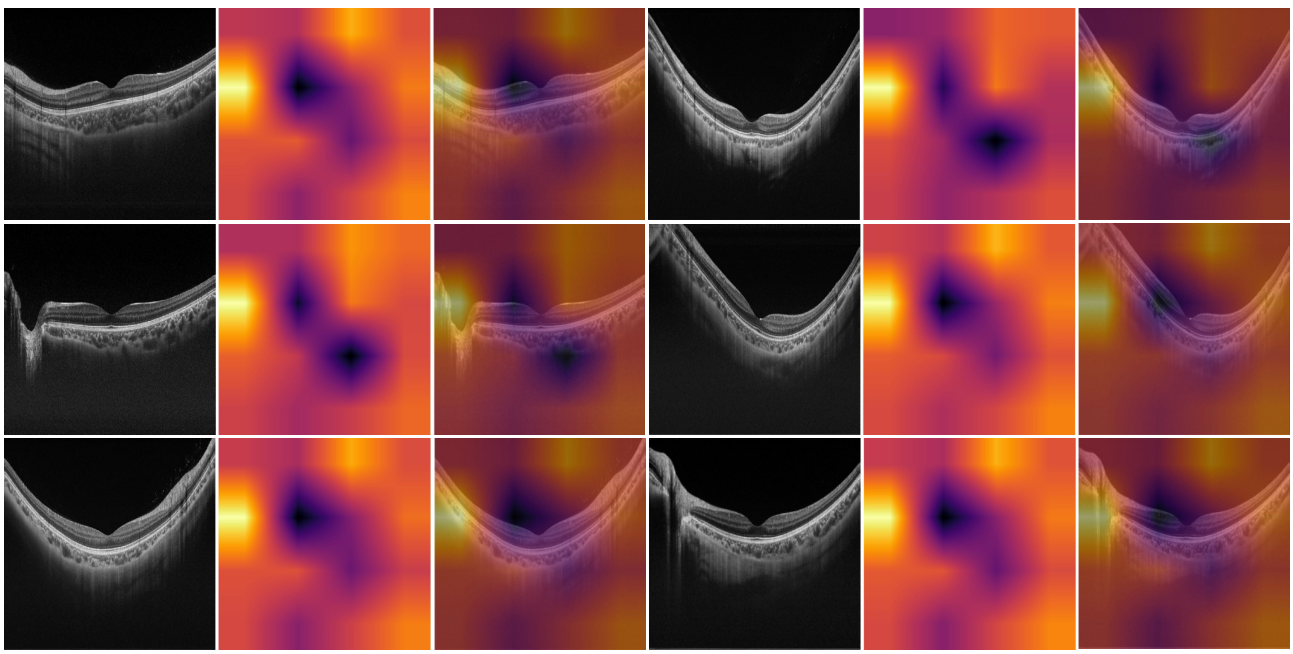
Method	ACC	IoU	DSC	SE	PC
Ablation-1	99.04(0.17)	70.43(3.75)	82.44(2.76)	85.36(0.88)	80.53(4.23)
Ablation-2	99.16(0.16)	73.62(4.17)	84.63(2.85)	86.24(1.84)	83.44(4.27)
Ablation-3	99.15(0.14)	73.10(3.78)	84.28(2.62)	85.89(1.41)	83.19(4.34)
Ablation-4	99.22(0.16)	75.17(4.77)	85.68(3.20)	86.45(2.06)	85.15(4.16)
Ablation-5	**99.24(0.15)**	**75.61(4.73)**	**85.93(3.17)**	**86.75(2.38)**	**85.47(4.69)**

average IoU score from 70.43% to 73.62%

Effect of ATB. Although multi-scale differential information between adjacent convolution feature maps is learned, the nature limitation of cascaded training is not solved (i.e., the second stage's VSN only processes the intersection of choroid layer prediction and OCT scan, which may lead to the segmentation of the wrong region, thereby misleading the VSN, or it may result in overlooking the real choroid layer part, failing to provide vital priors for the VSN). We try to add ATB to learn helpful global features from the OCT scan, which forms Ablation-3. As a result, the auxiliary branch further enhances the handling with vacant holes in vessel prediction. The auxiliary branch also enables the model to avoid predicting non-existing choroidal vessel regions. Singularly adding the ATB helps gain an IoU score improvement of 2.67%. When combining the MSC and ATB, formed as Ablation-4, the improvement in the IoU score further increases to 4.74%, integrating the strengths. The visualization of our Transformer auxiliary branch's captured feature map can be found in Fig.7.

Effect of Cascade Pre-Training Strategy. To unlock the potential of our MSC and ATB. We try to avoid the initial stage's error propagation in common cascaded training methods. We first pre-train LSB to a "plug-and-play" degree. Then, during the joint training stage, we reduce the learning rate to fine-tune LSB while setting a relatively higher learning rate for VSN. As a result, we successfully activate the latent potential of our network. After utilizing our cascade pre-training strategy, Ablation-5's result's boundary, shape, and extensions are even more comparable with the ground truth. Specifically, the only difference between Ablation-5 and Ablation-4 is that Ablation-5 applies the cascade pre-training strategy, and the gain is 0.44% in terms of IoU. Compared with Ablation-1, Ablation-5 achieves higher performance on average ACC, IoU, DSC, SC, and PC from 99.04%, 70.43%, 82.44%, 85.36%, and 80.53% to 99.24%, 75.61%, 85.93%, 86.75% and 85.47%, respectively, on our clinical choroidal vessel dataset.

Effect of Different Terms in Vessel Loss. To evaluate the effectiveness of different terms in $\mathcal{L}_{\text{vessel}}$, we first focus on the vanilla $\mathcal{L}_{\text{DICE}}$, resulting in an IoU of 75.85%. We then incrementally integrate $\mathcal{L}_{\text{wBCE}}$ and $\mathcal{D}_{\text{KL}}$ to assess their respective contributions to the model performance. Table 4 demonstrates that the addition of $\mathcal{L}_{\text{wBCE}}$ alone yields an increase in IoU from 75.85% to 76.15%. Similarly, integrating $\mathcal{D}_{\text{KL}}$ independently results in an IoU enhancement to 76.22%.



(a) (b) (c) (d) (e) (f)

Fig.7. The Visualization of our ATB's Sigmoid output using the "inferno" colormap. (a) and (d) showcase the input slices. (b) and (e) exhibit the Sigmoid output from the ATB, visually represented as a normalized graphical illustration using the "inferno" color map. (c) and (f) combine the input slice with the visualized Sigmoid output. The intensity of colors, with brighter, more vivid hues signifying higher values and darker shades indicating lower values, effectively conveys the numerical information. This graphic representation efficiently highlights our ATB's proficiency in identifying and integrating additional global features.

Table 4. Ablation Study Results for $\mathcal{L}_{\text{vessel}}$: Mean(%) of 5-Fold Cross-Validation

$\mathcal{L}_{\text{DICE}}$	$\mathcal{L}_{\text{wBCE}}$	$\mathcal{D}_{\text{KL}}$	IoU
✓			75.85
✓	✓		76.15
✓		✓	76.22
✓	✓	✓	**76.26**

Note: We set vanilla $\mathcal{L}_{\text{DICE}}$ as the main loss terms and explore the incremental impact of adding $\mathcal{L}_{\text{wBCE}}$ and $\mathcal{D}_{\text{KL}}$. The presence of these additional loss terms is indicated by a "✓" symbol.

When all three terms are applied together, the IoU is further increased to 76.26%. Besides, we additionally study the impact of the weight λ for the loss term $\mathcal{D}_{\text{KL}}$, which ranges from 0 to 100 with an increment of 10. In this ablation study, the main loss term $\mathcal{L}_{\text{DICE}}$ is with a constant weight of 1. Fig.8 shows that the optimal IoU is achieved when the weight λ is set to 70. Thus, when the weight λ is properly set, the proposed loss term $\mathcal{D}_{\text{KL}}$ may improve the performance of choroidal vessel segmentation.

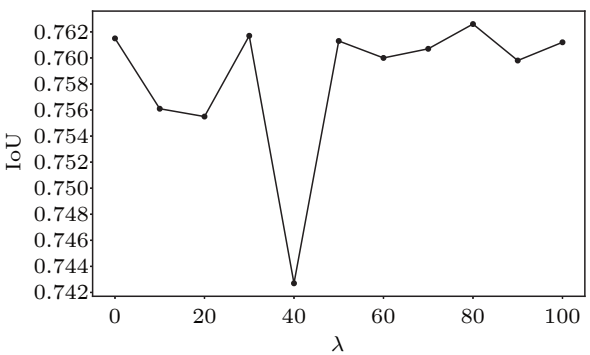


Fig.8. Relationship between $\mathcal{D}_{\text{KL}}$'s λ and IoU.

4.5 Versatility of Proposed Network

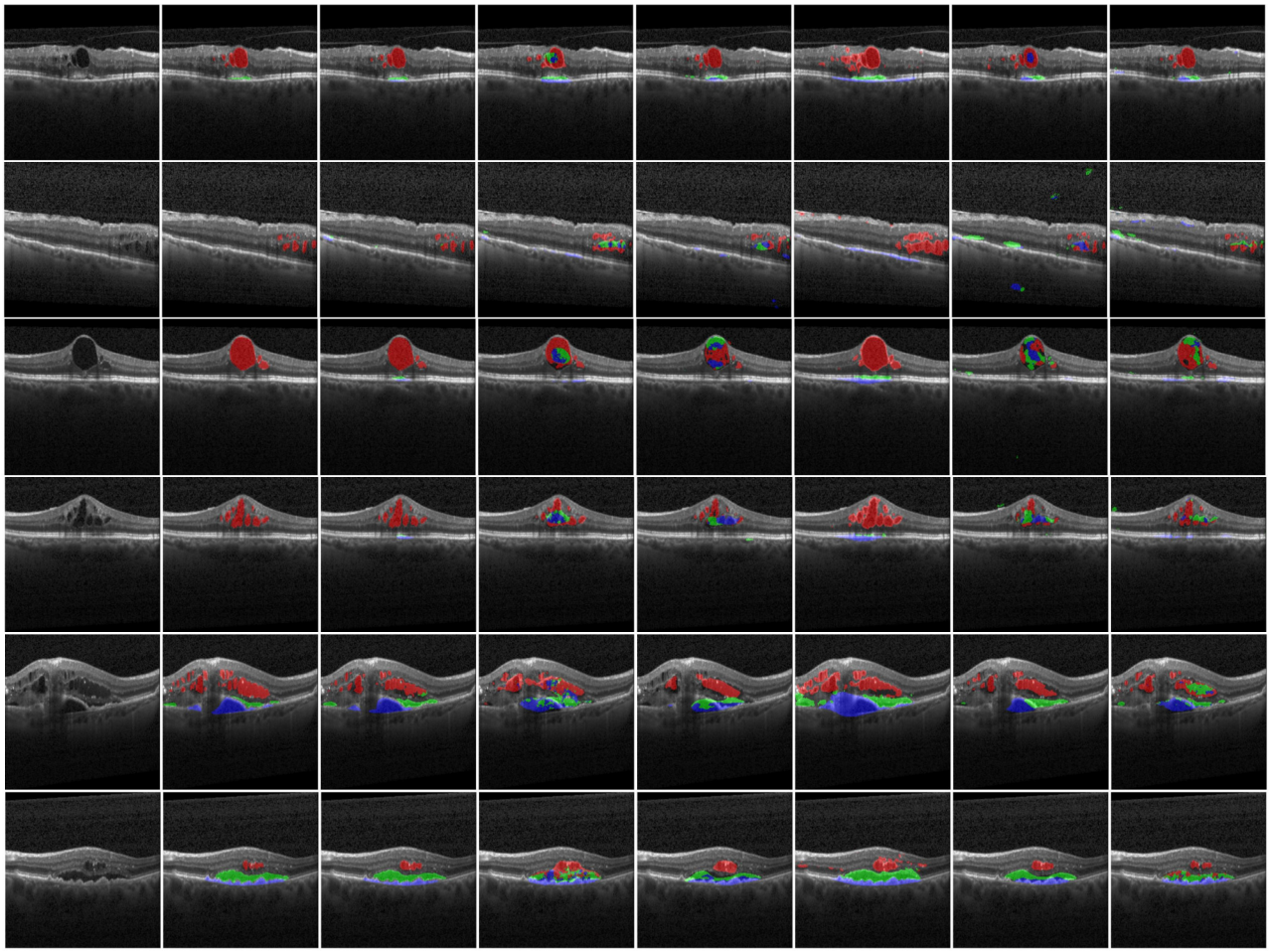
To underscore the versatility of our TACLNet, we train and validate it using an entirely new dataset, the RETOUCH[39] dataset. This poses a different challenge for our network: the segmentation of multi-class retinal fluid from spectral domain optical coherence tomography (SD-OCT) scans. We maintain consistency with the original training procedures. The only modification is replacing the dice-loss function with a cross-entropy loss function to better suit the new task's nature. To comprehensively understand our model's performance, we juxtapose it with three recent choroidal vessel segmentation methodologies, CUNet[10], ChoroidNET[11], and CVI-Net[12], and two baseline models, U-Net[36] and Attention U-Net[35]. Fig.9 provides a qualitative comparison, from which we can see that our TACLNet segments accurate fluid regions and inhibits the misclassification that often happens to the compared methods. Table 5 offers a detailed breakdown of comparative results. Together, they reinforce the assertion of the strength and versatility of our proposed model. TACLNet performs better than other methods across all metrics in most cases, except for the Dice coefficient of SRF, which shows a slight inferiority of 0.50%. Compared with the second-best model CUNet[10], TACLNet exhibits better results in terms of ACC, IoU, and DSC for all classes, as well as DSC for IRF and PED, achieving gains of 0.70%, 1.44%, 2.34%, 4.88%, and 1.67%, respectively.

5 Discussions

The choroid layer provides significant prior information for the ultimate segmentation when employing the cascaded method for choroidal vessel segmentation. This includes positioning the choroidal vessel and filtering out unrelated information in OCT scans. In this section, we challenge the commonly held belief that better choroid layer segmentation will always lead to improved segmentation of the vessels, which could guide future research efforts.

Importance of Choroid Layer Prior. As shown in Table 2, RefineNet[17], U-Net[36], and Attention U-Net[35] all directly learn from choroidal vessel labels, yielding inferior outcomes in vessel segmentation. Those methods that utilize choroid layer labels are all improved compared with the baseline. Thus, we can assume that applying the choroid layer prior can elevate the performance in the choroidal vessel segmentation task.

Influence of VSN Structure. In our initial ablation study for TACLNet, we utilize M^{2}S-Net[28] as the LSB. All ablation models incorporate cascaded training. As shown in Table 3 and Table 6, although our TACLNet (Ablation-5) achieves the best choroidal vessel segmentation performance, its performance in choroid layer segmentation is inferior to its ablation models. Our comparative study with State-of-the-Art methods also observes the same trend. As per Table 7, the choroid layer segmentation of our TACLNet trails behind two cascaded methods (CVI-Net[12] and ChoroidNET[11]) and one end-to-end method, CUNet[10]. These results potentially imply the influence of the VSN's structure is more significant than the prior information from the choroid layer segmen-



(a) (b) (c) (d) (e) (f) (g) (h)

Fig.9. Visualized retinal fluid segmentation samples of different segmentation networks. The six rows exhibit the visualized segmentation results of six individual OCT cases with different segmentation methods. Different colors indicate different fluid types (i.e., blue refers to SRF, red to IRF, and green to PED). (a) Input slice. (b) Ground truth. (c) Ours. (d) CVI-Net[12]. (e) Attention U-Net[35]. (f) CUNet[10]. (g) U-Net[36]. (h) ChoroidNET[11].

Table 5. Segmentation Result Comparisons Between Various Methods on the RETOUCH[39] Retinal Fluid Dataset: Mean(%) (Standard Deviation(%)) of 3-Fold Cross-Validation

Method	ACC	IoU	DSC	IRF	SRF	PED
CVI-Net[12]	98.45(0.27)	15.22(0.97)	22.72(1.18)	30.54(2.11)	20.15(0.99)	17.72(1.96)
Attention U-Net[35]	98.36(0.20)	18.91(3.68)	26.62(4.11)	32.95(2.44)	26.24(5.71)	20.69(4.78)
ChoroidNET[11]	98.81(0.18)	19.55(3.49)	27.56(3.74)	27.56(4.18)	**35.87(5.34)**	25.10(4.54)
U-Net[36]	98.38(0.16)	20.63(2.04)	28.64(2.16)	34.05(1.77)	28.79(3.32)	22.94(3.47)
CUNet[10]	98.23(0.36)	26.86(1.96)	33.84(1.95)	38.97(1.85)	35.68(3.54)	27.86(2.81)
TACLNet (ours)	**98.93(0.15)**	**28.30(1.63)**	**36.18(1.54)**	**43.85(1.97)**	35.37(3.51)	**29.53(2.32)**

Table 6. Ablation Study Results for Choroid Layer Segmentation During the Choroidal Vessel Segmentation: Mean(%) (Standard Deviation(%)) of 5-Fold Cross-Validation

Method	ACC	IoU	DSC	SE	PC
Ablation-1	99.54 (0.07)	94.01 (1.26)	96.88 (0.69)	97.70 (0.39)	96.20 (1.38)
Ablation-2	99.58 (0.06)	94.68 (0.94)	97.20 (0.50)	97.58 (0.33)	96.87 (0.72)
Ablation-3	**99.64 (0.05)**	**95.19 (0.94)**	**97.53 (0.50)**	**98.00 (0.36)**	**97.26 (0.45)**
Ablation-4	99.62 (0.07)	95.00 (0.98)	97.43 (0.52)	97.82 (0.24)	97.08 (0.75)
Ablation-5	99.50 (0.02)	93.40 (0.70)	96.56 (0.38)	95.93 (0.38)	97.23 (0.79)

Table 7. Choroid Layer Segmentation Comparisons During the Choroidal Vessel Segmentation: Mean(%) (Standard Deviation(%)) of 5-Fold Cross-Validation

Method	ACC	IoU	DSC	SE	PC
TACLNet (ours)	99.50 (0.01)	93.47 (0.67)	96.60 (0.37)	96.02 (0.22)	97.24 (0.55)
CUNet[10]	99.61 (0.08)	94.87 (1.14)	97.36 (0.60)	97.50 (0.84)	97.25 (1.00)
ChoroidNET[11]	99.70 (0.06)	96.04 (0.81)	97.97 (0.43)	98.01 (0.22)	97.96 (0.79)
CVI-Net[12]	**99.71 (0.05)**	**96.11 (0.73)**	**98.01 (0.38)**	**98.01 (0.22)**	**98.04 (0.58)**

tation.

Influence of LSB. We proceed to carry out an additional ablation study for LSB. For this particular study, we fix our proposed TACLNet's VSN. Then, we use U-Net[36], Attention U-Net[35], CVI-Net[12], M^{2}S-Net[28], and ChoroidNET[11] as LSB, all without pre-training. The results are shown in Table 1. From the result of M^{2}S-Net[28] and ChoroidNET[11], we can see that the LSB's performance in choroid layer segmentation doesn't decide the best vessel segmentation, since the model with better choroid layer segmentation is significantly inferior in the vessel segmentation task.

Cascade Pre-Training Strategy: Maximizing the Use of Prior Information. Although our data suggest that excellent choroid layer segmentation does not necessarily result in superior choroidal vessel segmentation, the importance of choroid layer segmentation should not be overlooked. Compared with methods lacking prior information, methods with decent choroid layer segmentation likely yield better results. This motivates us to devise a cascade pre-training strategy to use the prior information more efficiently. Pre-training can help establish a stronger foundation for our VSN, addressing error propagation and enhancing overall outcomes. During the joint training stage, combining fine-tuning choroid layer segmentation and training vessel segmentation could further render better performance. This can be observed in Table 3 and Table 6. Even when the choroid layer results are inferior, our Ablation-5 (TACLNet) achieves better segmentation than Ablation-4, which uses the same architecture but without applying the cascade pre-training strategy.

Model Complexity. From Table 8, our TACLNet has a similar floating-point computational workload compared with other cascaded methods[11, 36], but our TACLNet significantly outperforms in choroidal vessel segmentation, as illustrated in Table 2. Compared with non-cascade methods[10, 8, 17], TACLNet is larger than some methods but shows significant performance improvements over these methods. To reduce model complexity, in our future research, we aim to use the choroid layer priors with a more streamlined module without compromising the performance of the choroidal vessel segmentation.

Table 8. Complexity Comparisons between Various Segmentation Methods.

Method	Flops (10^9)	Size (10^6)
TACLNet (ours)	920.22	99.48
ChoroidNET[11]	770.36	57.40
CVI-Net[12]	2 886.88	234.59
U-Net[36]	929.52	69.72
U-Net[36](Non-Cascade)	464.62	34.86
CUNet[10] (Non-Cascade)	575.10	50.20
Attention U-Net[35] (Non-Cascade)	473.50	35.21
RefineNet[8, 17] (Non-Cascade)	186.62	177.42

6 Conclusions

In this work, we proposed Transformer-Assisted Cascade Learning Network (TACLNet) to improve the performance in choroidal vessel segmentation. The proposed TACLNet mainly uses a cascade pre-training strategy that can effectively learn the valuable prior from the pre-trained LSB and then jointly train LSB and VSN. The designed MSC can provide spatial differential information from adjacent feature maps while keeping the local details. Besides, ATB compensates for the lost global information from cascaded training. Experimental results on the choroidal vessel and retinal fluid segmentation tasks demonstrate that our TACLNet outperforms other well-known choroidal vessel segmentation methods in terms of accuracy and versatility. Moreover, our segmentation method can help ophthalmologists accurately predict choroidal vessel regions, reducing the burden of manual quantitative analysis of choroidal-related retinal diseases.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Nickla D L, Wallman J. The multifunctional choroid.

Progress in Retinal and Eye Research, Mar. 2010, 29(2): 144–168. DOI: 10.1016/j.preteyeres.2009.12.002.

[2] Singh S R, Vupparaboina K K, Goud A, Dansingani K K, Chhablani J. Choroidal imaging biomarkers. *Survey of Ophthalmology*, 2019, 64(3): 312–333. DOI: 10.1016/j.survophthal.2018.11.002.

[3] Arrigo A, Bordato A, Romano F, Aragona E, Grazioli A, Bandello F, Parodi M B. Choroidal patterns in retinitis pigmentosa: Correlation with visual acuity and disease progression. *Translational Vision Science & Technology*, 2020, 9(4): 17. DOI: 10.1167/tvst.9.4.17.

[4] Spaide R F, Fujimoto J G, Waheed N K, Sadda S R, Staurenghi G. Optical coherence tomography angiography. *Progress in Retinal and Eye Research*, May 2018, 64: 1–55. DOI: 10.1016/j.preteyeres.2017.11.003.

[5] Fercher A F, Hitzenberger C K, Kamp G, El-Zaiat S Y. Measurement of intraocular distances by backscattering spectral interferometry. *Optics Communications*, May 1995, 117(1/2): 43–48. DOI: 10.1016/0030-4018(95)00119-S.

[6] Lavinsky F, Lavinsky D. Novel perspectives on swept-source optical coherence tomography. *International Journal of Retina and Vitreous*, 2016, 2(1): Article No. 25. DOI: 10.1186/s40942-016-0050-y.

[7] Sezer T, Altınışık M, Koytak İ A, Özdemir M H. The choroid and optical coherence tomography. *Turkish Journal of Ophthalmology*, 2016, 46(1): 30–37. DOI: 10.4274/tjo.10693.

[8] Liu X X, Bi L, Xu Y P, Feng D G, Kim J, Xu X. Robust deep learning method for choroidal vessel segmentation on swept source optical coherence tomography images. *Biomedical Optics Express*, 2019, 10(4): 1601–1612. DOI: 10.1364/boe.10.001601.

[9] Qiu B, Huang Z Y, Liu X *et al*. Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function. *Biomedical Optics Express*, 2020, 11(2): 817–830. DOI: 10.1364/boe.379551.

[10] Zhu L, Li J M, Zhu R L *et al*. Synergistically segmenting choroidal layer and vessel using deep learning for choroid structure analysis. *Physics in Medicine & Biology*, 2022, 67(8): 085001. DOI: 10.1088/1361-6560/ac5ed7.

[11] Khaing T T, Okamoto T, Ye C *et al*. ChoroidNET: A dense dilated U-Net model for choroid layer and vessel segmentation in optical coherence tomography images. *IEEE Access*, 2021, 9: 150951–150965. DOI: 10.1109/ACCESS.2021.3124993.

[12] Wang X H, Li R, Chen J Y *et al*. Choroidal vascularity index (CVI)-Net-based automatic assessment of diabetic retinopathy severity using CVI in optical coherence tomography images. *Journal of Biophotonics*, Jan. 2023, 16(6): e202200370. DOI: 10.1002/jbio.202200370.

[13] Ren S C, Zhou D Q, He S F, Feng J S, Wang X C. Shunted self-attention via multi-scale token aggregation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.10843–10852. DOI: 10.1109/CVPR52688.2022.01058.

[14] Zhang L, Lee K, Niemeijer M, Mullins R F, Sonka M, Abràmoff M D. Automated segmentation of the choroid from clinical SD-OCT. *Investigative Ophthalmology & Visual Science*, 2012, 53(12): 7510–7519. DOI: 10.1167/iovs.12-10311.

[15] Chen Q, Fan W, Niu S J, Shi J J, Shen H L, Yuan S T. Automated choroid segmentation based on gradual intensity distance in HD-OCT images. *Optics Express*, 2015, 23(7): 8974–8994. DOI: 10.1364/oe.23.008974.

[16] Hussain M A, Bhuiyan A, Ishikawa H, Theodore Smith R, Schuman J S, Kotagiri R. An automated method for choroidal thickness measurement from enhanced depth imaging optical coherence tomography images. *Computerized Medical Imaging and Graphics*, Jan. 2018, 63: 41–51. DOI: 10.1016/j.compmedimag.2018.01.001.

[17] Lin G S, Milan A, Shen C H, Reid I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.5168–5177. DOI: 10.1109/CVPR.2017.549.

[18] Zheng G, Jiang Y F, Shi C *et al*. Deep learning algorithms to segment and quantify the choroidal thickness and vasculature in swept-source optical coherence tomography images. *Journal of Innovative Optical Health Sciences*, 2021, 14(01): 2140002. DOI: 10.1142/S1793545821400022.

[19] Zhang Z X, Liu Q J, Wang Y H. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 2018, 15(5): 749–753. DOI: 10.1109/LGRS.2018.2802944.

[20] Huang K, Su N, Ma X, Li M C, Yang J D, Yuan S T, Liu Y, Chen Q. Choroidal vessel segmentation in SD-OCT with 3D shape-aware adversarial networks. *Biomedical Signal Processing and Control*, 2023, 84: 104982. DOI: 10.1016/j.bspc.2023.104982.

[21] Chen M, Wang J C, Oguz I, VanderBeek B L, Gee J C. Automated segmentation of the choroid in EDI-OCT images with retinal pathology using convolution neural networks. In *Proc. the 4th International Workshop on Ophthalmic Medical Image Analysis*, Sept. 2017, pp.177–184. DOI: 10.1007/978-3-319-67561-9_20.

[22] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2017, 39(12): 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[23] Zhang H H, Yang J L, Zhou K *et al*. Automatic segmentation and visualization of choroid in OCT with knowledge infused deep learning. *IEEE Journal of Biomedical and Health Informatics*, 2020, 24(12): 3408–3420. DOI: 10.1109/JBHI.2020.3023144.

[24] Bai H R, Cheng S S, Tang J H, Pan J S. Learning a cascaded non-local residual network for super-resolving blurry images. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2021, pp.223–232. DOI: 10.1109/CVPRW53098.2021.

00031.

[25] Zhao X Q, Zhang L H, Lu H C. Automatic polyp segmentation via multi-scale subtraction network. In *Proc. the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Sept. 2021, pp.120–130. DOI: 10.1007/978-3-030-87193-2_12.

[26] Yang M K, Yu K S, Zhang C, Li Z, Yang K. DenseASPP for semantic segmentation in street scenes. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.3684–3692. DOI: 10.1109/CVPR.2018.00388.

[27] Zhao X P, Pang Y W, Zhang L H, Lu H C, Zhang L. Suppress and balance: A simple gated network for salient object detection. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.35–51. DOI: 10.1007/978-3-030-58536-5_3.

[28] Zhao X Q, Jia H P, Pang Y W *et al.* M^{2}SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. arXiv: 2303.10894, 2023. https://arxiv.org/abs/2303.10894, Mar. 2024.

[29] Li J, Fan J S, Zhang Z X. Towards noiseless object contours for weakly supervised semantic segmentation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.16835–16844. DOI: 10.1109/CVPR52688.2022.01635.

[30] Xu L, Ouyang W L, Bennamoun M *et al.* Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.6964–6973. DOI: 10.1109/ICCV48922.2021.00690.

[31] Ding M Y, Lian X C, Yang L J, Wang P, Jin X J, Lu Z W, Luo P. HR-NAS: Searching efficient high-resolution neural architectures with lightweight transformers. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.2981–2991. DOI: 10.1109/CVPR46437.2021.00300.

[32] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S, Guo B N. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.9992–10002. DOI: 10.1109/ICCV48922.2021.00986.

[33] Wang W H, Xie E Z, Li X *et al.* Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.548–558. DOI: 10.1109/ICCV48922.2021.00061.

[34] Dosovitskiy A, Beyer L, Kolesnikov A *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. the 9th International Conference on Learning Representations*, May 2021.

[35] Oktay O, Schlemper J, Folgoc L L *et al.* Attention U-Net: Learning where to look for the pancreas. arXiv: 1804.03999, 2018. https://arxiv.org/abs/1804.03999, Mar. 2024.

[36] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Oct. 2015, pp.234–241. DOI: 10.1007/978-3-319-24574-4_28.

[37] Wei J, Wang S H, Huang Q M. F^{3}Net: Fusion, feedback and focus for salient object detection. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Feb. 2020, pp.12321–12328. DOI: 10.1609/aaai.v34i07.6916.

[38] Kullback S, Leibler R A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, 22(1): 79–86. DOI: 10.1214/aoms/1177729694.

[39] Bogunovic H, Venhuizen F, Klimscha S *et al.* RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. *IEEE Trans. Medical Imaging*, 2019, 38(8): 1858–1874. DOI: 10.1109/TMI.2019.2901398.

Yang Wen received her Master's degree in electronics and communication engineering from Xidian university, Xi'an, in 2015, and her Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, in 2021. She is currently an assistant professor with the College of Electronics Information Engineering, Shenzhen University, Shenzhen. Her research interests include image processing, computer vision, medical image analysis and application, and image/video coding and transmission.

Yi-Lin Wu is an undergraduate student at Shenzhen University, Shenzhen. He is currently pursuing his B.S. degree in electronic information engineering. His research interests include image processing, computer vision, and medical image analysis.

Lei Bi is an associate professor with the Institute of Translational Medicine at the Shanghai Jiao Tong University. He received his Ph.D. degree from The University of Sydney, Sydney, in 2018. His current research focus is on multi-modality medical image analysis and visualization, and he collaborates along with hospital and industry partners to translate the research outputs into clinical applications.

Wu-Zhen Shi received his Bachelor's degree in information and computing science from Shenyang Agricultural University, Shenyang, his Master's degree in agricultural informatization from Northwest A&F University, Yangling, and his Ph.D. degree in computer applied technology from Harbin Institute of Technology, Harbin, in 2012, 2014, and 2020, respectively. He is currently an assistant professor with the College of Electronics Information Engineering, Shenzhen University, Shenzhen. His research interests include image processing, computer vision, and image/video coding and transmission.

Xiao-Xiao Liu received his Bachelor degree of medicine and Ph.D. degree in ophthalmology from Shanghai Jiao Tong University, in 2013 and 2019, respectively. He is currently an attending ophthalmologist in Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai. His research interests include application of artificial intelligence in pediatric ophthalmology, such as choroid segmentation of myopia children and amblyopic children.

Yu-Peng Xu received his MBBS degree (Bachelor of Medicine/Bachelor of Surgery) and Ph.D. degree in ophthalmology from Shanghai Jiao Tong University, Shanghai, in 2012 and 2017, respectively. He is currently an attending doctor in Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai. His research interests include disease model and image analysis in fundus disease, children visual function development, and fundus disease-related glaucoma.

Xun Xu is the director of the National Clinical Research Center for Eye Diseases, director of the Department of Ophthalmology, Shanghai Jiao Tong University School of Medicine, Shanghai, vice chairman of the Chinese Ophthalmology Society, and president of the Chinese Retinal Society. He was awarded the Second Prize of National Science and Technology Progress and the First Prize of Shanghai Science and Technology Progress.

Wen-Ming Cao received his M.S. degree in communication and information systems from the System Science Institute Science Academy, Beijing, in 1991, and his Ph.D. degree in communication and information systems from the School of Automation, Southeast University, Nanjing, in 2003. He is currently a professor with Shenzhen University, Shenzhen. His research interests include pattern recognition, image processing, and visual tracking.

David Dagan Feng received his M.E. degree in electrical engineering and computing science (EECS) from Shanghai Jiao Tong University, Shanghai, in 1982, his M.S.c degree in biocybernetics and his Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), in 1985 and 1988, respectively. He is currently a professor at the School of Information Technologies, Faculty of Engineering and Information Technologies, The University of Sydney, Sydney. He is also the director of Biomedical and Multimedia Information Technology (BMIT) Research Group and director (Research) of Institute of Biomedical Engineering and Technology (BMET), School of Information Technologies, The University of Sydney, Sydney.