Biocybernetics and Biomedical Engineering 45 (2025) 340–356

Contents lists available at ScienceDirect

Biocybernetics and Biomedical Engineering

journal homepage: www.elsevier.com/locate/bbe

Research Paper

Multi-scale CNN-Swin transformer network with boundary supervision for multiclass biomarker segmentation in retinal OCT images

Zhanpeng Fan [a,b], Xiaoming Liu [a,b,*], Ying Zhang [c], Jia Zhang [a,b]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430070, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430070, China
[c] Wuhan Aier Eye Hospital of Wuhan University, Wuhan 430064, China

ARTICLE INFO

Keywords:
Optical Coherence Tomography (OCT)
Retinal Biomarker Segmentation
Convolutional Neural Network
Multi-class Segmentation

ABSTRACT

Retinal biomarker morphology is closely associated with a variety of chronic ophthalmic diseases, in which biomarker localization and segmentation in optical coherence tomography (OCT) play a key role in the diagnosis of retina-related diseases. Although great progress has been made in deep learning based OCT biomarker segmentation, several challenges still exist. Due to issues such as image noise or class imbalance, retinal biomarkers affect the model's recognition of other biomarkers. Moreover, small biomarkers are prone to lose accuracy during downsampling. And most existing methods rely on convolutional neural networks, which make it challenging to obtain the global context due to locality of convolution. Benefiting from the Swin Transformer with powerful modeling capabilities, we propose MSCS-Net (Multi-scale CNN-Swin Network), a network for OCT biomarker segmentation, which effectively combines CNN and Swin Transformer and integrates them in parallel into a dual-encoder structure. Specifically, an edge detection path is added alongside to enhance the localization of biomarkers at the edges. For the Swin Transformer branch, considering the irregular distribution of most OCT biomarkers, a new windowing partition is performed in the Swin Transformer to capture the features more efficiently. Meanwhile, we design a Feature Dimensionality Reduction Module to extensively collect the information of small-scale biomarkers. To effectively integrate information from two scales, we design a Transformer Cross Fusion Module to finely fuse the global and local feature information from the two-branch encoders. We validate the proposed approach on local and public datasets, and the experimental results demonstrate the effectiveness of the proposed framework.

1. Introduction

The retina is an important structure inside the human eye, and retinal diseases are receiving increasing attention in the 21st century. In 2023, according to the World Health Organization, at least 2.2 billion people globally will have impaired near or distance vision, with at least 1 billion of these individuals experiencing vision impairment that could have been prevented or has yet to be addressed [1]. The retina is susceptible to ophthalmic diseases and therefore contains a wealth of diagnostic information for ophthalmic diseases. Initial diagnosis and disease activity are imaged by optical coherence tomography (OCT), a tool for assessing specific retinal morphology and subretinal changes associated with visual function and disease progression [2]. OCT is a non-invasive diagnostic method that uses infrared light at wavelengths from 800 to 1400 nm to generate high-quality and high-resolution scans of the retinal structure [3]. The automatic processing of OCT images makes it easier to present retinal biomarkers, such as vascular segmentation [4], fluid segmentation [5] and layer segmentation [6], which is not only beneficial to disease diagnosis, but also helps ophthalmologists to monitor retinal diseases more accurately, and adjust the treatment plan in a timely manner.

Fig. 1 shows 6 OCT-B Scans, they include Geographic Atrophy (GA), Hyperreflective Foci (HF), Choroidal Neovascularization (CNV), Epiretinal Membrane (ERM), Pigment Epithelial Detachment (PED), Intraretinal Fluid (IRF), Subretinal Fluid (SRF), DRUSEN, and a healthy eye. Age-related macular degeneration (AMD) is the most common cause of visual impairment in developed countries [7] and is a disease that affects a person's central vision. Drusen is a deposition of yellow lipid particles that form between the retinal pigment epithelium and the Bruch's membrane, and is one of the main features of early AMD [8]. Soft (early)

* Corresponding author.
E-mail address: lxmspace@gmail.com (X. Liu).

https://doi.org/10.1016/j.bbe.2025.05.002
Received 9 August 2024; Received in revised form 23 February 2025; Accepted 6 May 2025
Available online 20 May 2025

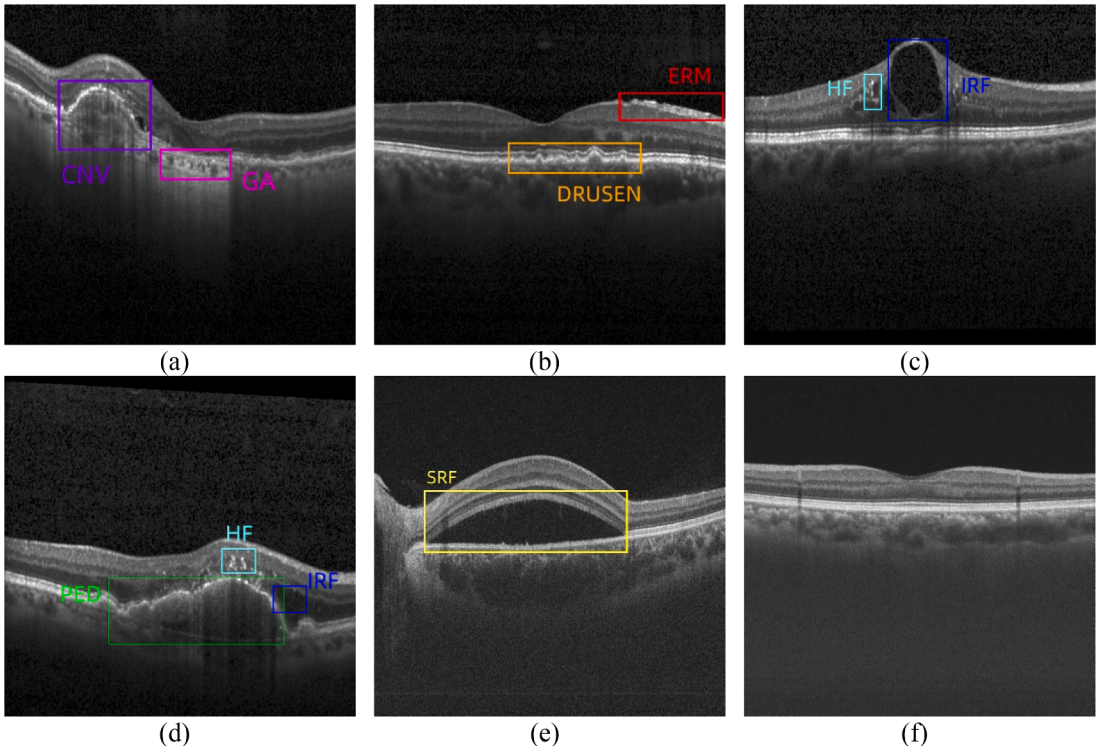

Fig. 1. Eight classes of OCT biomarkers and a healthy OCT image.

Drusen is more likely to progress further to wet AMD (CNV or PED) or dry AMD (GA) and these forms of AMD can lead to severe vision loss or blindness [9]. These can be scanned and monitored by OCT. Meanwhile diabetic macular edema (DME) is a serious complication of diabetic retinopathy and is one of the leading causes of blindness in middle-aged and elderly populations [10]. HF is usually due to abnormal lipid deposition in the macular region and is one of the complications of DME. Severe HF may also lead to inflammation and cellular damage in the macular region, which can further exacerbate the condition of DME [11]. Meanwhile, IRF and SRF are also often accompanied in patients with DME, further aggravating the edema and visual impairment in the macular region [9]. Retinal biomarkers are important in the treatment of ophthalmic diseases and can help physicians achieve the goals of early diagnosis, personalized treatment, therapeutic monitoring, and prediction of risk to improve patient outcomes and quality of life.

In recent years, with the rise of deep learning, especially the deep learning network with U-Net [12] as the basic network framework, CNN has gradually dominated the medical image segmentation. Compared with some traditional methods such as thresholding methods [13], tracking methods [14,15], deep networks can automatically learn features with a large number of learnable parameters and can segment objects more accurately. For biomarker segmentation, in recent years, there are also some deep learning methods [16,17], either based on CNN, or based on transformer. In [18], a two-branch U-Net and Y-Net are proposed for automatic GA segmentation in multimodal Fundus Autofluorescence (FAF) and Near-Infrared (NIR) images. In [19], Pham et al. propose two cascade networks for the small and fuzzy Drusen segmentation in early AMD, and the segmentation probability maps are combined using image-level and patch-level networks to obtain the final results. For CNV segmentation, Xi et al. [20] propose a new information-attentive convolutional neural network (IA-Net) in OCT images, targeting the two characteristics of small objects in CNV and its complexity.

In the past few years, there have been works such as the above to segment a single biomarker in some of the OCT images. But for most of the diseases, there may be multiple types of biomarkers in the same OCT image at the same time, and currently there are only a few investigations on multiclass biomarker segmentation on OCT images. Based on FCN [21], Xing et al. [22] performe fluid biomarker (IRF,SRF,PED) segmentation simultaneously, by employing attention gates and spatial pyramid pooling modules to improve the network's ability to extract multi-scale objects, and then introducing a novel curvature regularization term in the loss function to merge the priori shape information. Multi-class biomarkers are more complex and usually contain a large number of details and complex structures, while segmenting them requires accurate identification and classification of these details and structures, which is more difficult. Simply using FCN may not be able to handle biomarkers of different sizes efficiently. To address this problem, Rasti et al. [23] propose a novel Self-Adaptive Dual Attention module alongside multiple Skip connections for effective representation learning, encompassing texture, context, and edge features. However, it does not utilize multiscale information generated by hierarchical structures and fails to integrate low and high-level features cohesively while maintaining feature consistency.

We identify the following challenges in the existing biomarker segmentation work. First, for biomarkers in OCT images, existing work has rarely segmented more than four classes of biomarkers. However, by segmenting various biomarkers, diseases can be diagnosed more accurately, and the progression and treatment effects of diseases can be monitored. The combination of different biomarkers can provide more comprehensive information, helping doctors make more precise diagnostic and treatment decisions. Second, as shown in Fig. 1(d), different biomarkers exhibit significant differences in shape, position, and size, which makes it challenging for the model to recognize and segment these highly variable biomarkers. In our work, there are eight types of biomarkers in the OCT image, and the shape of different biomarkers such as IRF and CNV varies greatly, and some base models such as FCN or U-Net are not good at capturing representative differences and features of individual biomarkers. Third, small-sized biomarkers, like HF and certain fluid biomarkers (as shown in Fig. 1(c) and Fig. 1(e)), have discrete sizes, and they are generally distributed in a vertical direction. In the context of generic medical image segmentation networks such as U-Net, downsampling processes may dilute fine details of these small

objects, leading to information loss and hindering accurate localization and identification. Additionally, variations in biomarker characteristics among different patients further complicate segmentation tasks, potentially resulting in the loss of crucial details and structural information essential for semantic segmentation of dense, small-scale objects in OCT images. Fourth, neglecting boundary information. Extracting biomarker boundary is important to improve the quality of segmentation and help doctors more accurately identify and quantify the morphology and features of biomarkers. Biomarkers are mainly concentrated in the middle and lower positions of the image, and boundary information helps to obtain more accurate locations and boundaries of the OCT biomarker regions. For single-class segmentation within a B-scan, Liu et al. [24] propose a weakly supervised framework utilizing knowledge distillation for segmenting biomarkers.

In this study, we propose a new encoder-decoder framework based on CNN-Transformer, named MSCS-Net, to perform more accurate multi-class biomarker segmentation. To deal with the many shortcomings of CNN in global modeling, which make it difficult to adapt to biomarker at all scales, we take advantage of the powerful global modeling capability of Swin Transformer [25] to form a two-branch encoder to perform multi-scale interaction through a four-layer multi-scale pyramid. And to tackle the issue of small biomarkers losing details easily during downsampling, we propose a new small target extraction module Feature Dimensionality Reduction Module in Swin branch to effectively identify small biomarkers in OCT images. In order to address the issue of more accurately locating biomarkers using edge information, a boundary guided module (BGM) is designed alongside the CNN encoder to assist in supervising biomarker segmentation. Then by the designed feature fusion module, the resulting minimum and maximum pyramid level representations are then fed into the proposed Transformer Cross Fusion Module (TCF). The newly proposed TCF module integrates a multi-scale vision transformer that combines two feature maps via cross-attention. These recalibrated feature maps are subsequently fed into the decoder block to generate the final segmentation mask. Our main contributions are as follows:

1. A new biomarker automatic segmentation framework MSCS-Net is proposed, which integrates the long-range context interaction of Transformer and the local semantic information of CNN. And design a boundary extraction module to assist in supervising the segmentation results, while focusing on the edge information.
2. In order to better extract small biomarkers, a Feature Dimensionality Reduction Module is proposed, meanwhile we modify the window strategy of Swin Transformer to target the distributional features of small biomarkers.
3. A new feature fusion module is proposed to achieve coarse-grained and fine-grained fusion, and optimize the training and improve the performance of the model by combining the Transformer with the cross-attention.
4. We evaluate the proposed method on both local and public datasets. The results demonstrate the superiority of our proposed method.

2. Related work

2.1. OCT biomarker segmentation based on CNN

With the application of deep learning to medical image segmentation, the demand for segmentation in retinal OCT images is increasing. Recently, various deep learning techniques have been applied in the segmentation of retinal diseases and layers. U-Net [12], FCN [21], SegNet [26], Deeplabv3+ [27], and U-Net++ [28] are popular and well-liked networks in the field of medical segmentation. For the segmentation of biomarker in OCT images, Lachinov et al. [29] proposed an improved U-Net-like projective skip connection network for automatic GA segmentation. Meng et al. [30] designed a new multi-scale adaptive-aware deformation module in the information fusion network, which focuses on the specific shapes of CNVs and aggregates contextual information with attention to more semantic details. Gende et al. [31] proposed three fully automated multi-task end-to-end approaches for ERM screening and segmentation. In order to take more account of the retinal layer structure, Liu et al. [32] introduced an attention-based fluid segmentation approach, incorporating multiscale inputs, side outputs, and attention mechanisms into U-Net++ to enhance performance. Liu et al. [16] introduced the Local-Global Transformer module and Contrastive Learning Enhancement module to address challenges in biomarker segmentation.

Although CNN-based methods have achieved good results in most specific biomarker segmentation, there are still some problems. First, previous work primarily focuses on segmenting a specific class of biomarker (CNV, HF, etc.), or three classes of fluid biomarker, but in ophthalmic diseases, which are often accompanied by multiclass lesions, multiclass biomarkers are important, and previous work did not consider this aspect. As for the few-class biomarker, previous CNN-based methods have focused on the region related to that lesion, which is not very comprehensive for the global information, and also seldom consider the effect of multi-scale information. The scale of OCT retinal images varies greatly, which makes it difficult to deal with at a single scale. Therefore, methods that segment multiple classes of biomarkers simultaneously are both necessary and challenging.

2.2. Medical segmentation using boundary features

Extracting boundary features is greatly beneficial for medical image segmentation. Since there are regions in medical images with subtle grayscale variations or complex textures, the boundary information can more accurately locate the target object, resulting in a more precise segmentation result. Hata et al. [33] propose end-to-end boundary aware CNNs for medical image segmentation, and consider organ boundary information through special network edge branches and edge-aware loss terms. However, due to the lack of modeling capability for complex shapes and fine details of target boundaries, CNN-based methods often produce inaccurate segmentation masks of target boundaries. Therefore, Wang et al. [34] propose a boundary-aware context neural network for 2D medical image segmentation, utilizing a pyramid edge extraction module to obtain edge information with multiple granularities. Lee et al. [35] propose the boundary key point selection algorithm. In which, key points on the structural boundary of the target object are estimated. Subsequently, a boundary preserving block with the boundary key point map is applied to predict the structural boundary of the target object. Lin et al. [36] combine Transformer to construct a boundary-aware local transformer module, which adopts an adaptive window partitioning scheme under the guidance of entropy, reducing computational complexity and preserves shape integrity. Different from these works, we design an edge-guided module, which introduces the Scharr operator to enhance boundary features.

2.3. Transformer for medical segmentation

Unlike the locally-focused CNN, the globally-focused Transformer has achieved great success in NLP [37], and is widely used in various fields therein. Dosovitskiy et al. [38] propose Vision Transformer (ViT), which was the first attempt to pre-train a purely Transformer-based architecture on large datasets such as ImageNet-22 K, which results in SOTA performance in image recognition. Subsequently, Chen et al. [39] explore a generalized Transformer-based pre-training method for image processing tasks. Nevertheless, ViT still suffers from a huge training cost for intensive prediction tasks. Therefore, some works modify the ViT architecture to accommodate intensive prediction tasks such as semantic segmentation and object detection. SETR [40] treats the Transformer as an encoder that models the global context in each layer, and by using the Transformer as an example of an implementation of such a sequential model, SETR improves the segmentation capability of ViT. In order to

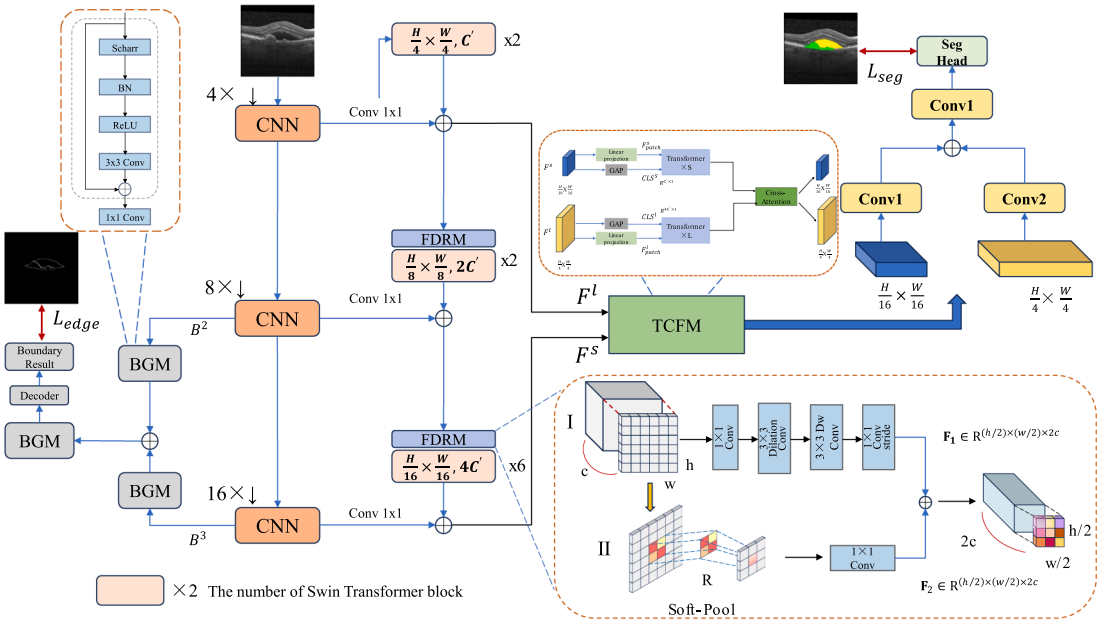

Fig. 2. Overall framework of the Proposed MSCS-Net.

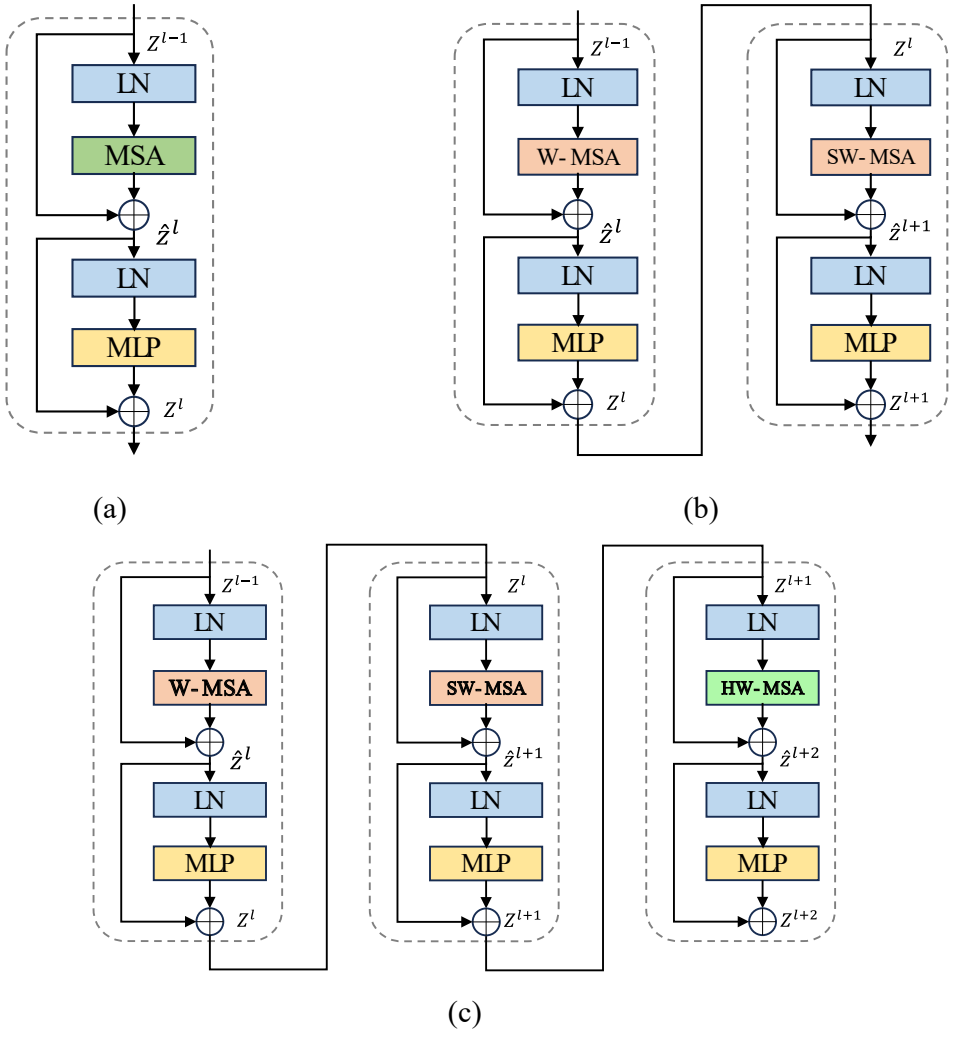

Fig. 3. (a) Architecture of a standard transformer block. (b) Schematic of a Swin Transformer block. (c) The Hybrid shifted window-based MSA we propose.

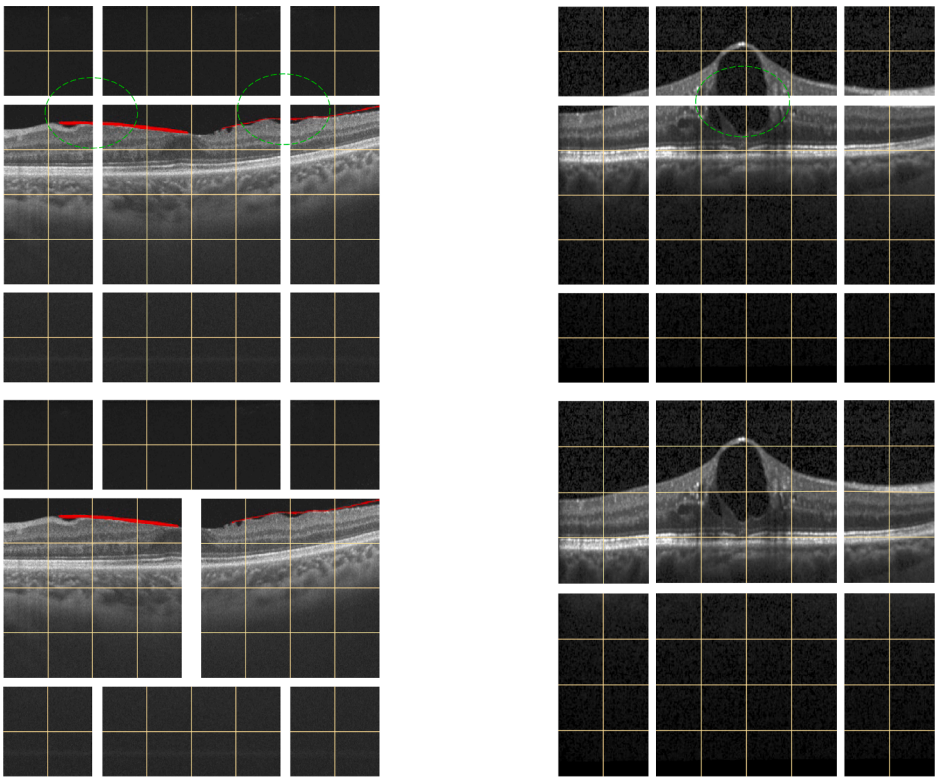

Fig. 4. The first line is the SW-MSA of the Swin Transformer, and the second line shows the two additional windowed partitioning methods we added to the Swin block module. (a) represents the biomarker ERM; (b) represents the biomarker IRF.

improve the computational efficiency of ViT, Swin Transformer [25] introduces a window attention mechanism, which decomposes the input feature map into multiple chunks, restricts the attention computation to a single window, and the shift-window based MSA (SW-MSA) has linear computational complexity, which is flexible, and is useful in a variety of areas including image classification, object detection and semantic segmentation, achieving state-of-the-art performance.

In the field of medical images, with transformer as the backbone, Chen et al. [41] first propose TransU-Net, which is a U-shaped transformer, proving that transformer can also be used as a powerful encoder in the field of medical image segmentation. However, TransFuse [42] points out that a segmentation network based entirely on the Transformer focuses only on global modeling and ignores localization capabilities, leading to unsatisfactory segmentation results. To solve this problem, they stack the CNN and the transformer sequentially in a parallel fashion to form a new encoder that fuses the two features while executing the encoder in parallel. Inspired by the above method, we propose a new CNN-Transformer framework. We design a Feature Dimensionality Reduction Module to better capture features of small targets. Through the Transformer Cross Fusion Module, we exchange multiscale information and introduce an edge boundary-guided pathway to supervise the segmentation results.

3. The proposed method

Fig. 2 shows the detailed flowchart of our proposed multi-class biomarker segmentation network MSCS-Net, which roughly consists of a CNN-Transformer two-branch encoder, a fusion module, an edge detection path and a decoder. Given a training dataset $D_{\text{train}} = \{(X_n, Y_n)\}_{n=1}^{N}$ consists of N samples, where X_n is the input image set, $X_n \in R^{H\times W}$, $Y_n = \{0, 1, \cdots, 8\}^{H\times W}$ represents 9 categories, where the normal region is 0, and the rest of the 1–8 is the biomarker images respectively (ERM, PED, SRF, IRF, GA, HF, DRUSEN, and CNV). The test dataset with M testing samples $D_{\text{test}} = \{X_m\}_{m=1}^{M}$ is used to evaluate the trained model capability. The image first goes through the CNN encoder, and on the right side, it is input to the Swin Transformer encoder through a skip connection. In the Swin Transformer block, based on the original sliding window, our proposed Targeted Shifted Window-based MSA is applied to adapt to the distribution characteristics of OCT images. To enhance focus on small biomarkers and mitigate the loss of details during downsampling, we design the Feature Dimensionality Reduction Module (FDRM). The FDRM enhances the detection and extraction capabilities of small targets. After downsampling by the proposed two-branch encoder, the features of the first and third layers are fused using our designed Transformer Cross Fusion Module (TCFM), which can fuse features while preserving localization information. The segmentation results are then obtained by the decoder. To better utilize boundary information, we add a boundary detection pathway on the left side of the CNN. To enhance boundary features, the features of each convolutional block are refined by our proposed Boundary Guided Module (BGM). The resulting boundary map is used to supervise the segmentation results.

3.1. Encoder of MSCS

In the CNN encoder branch, we first use ResNet [43] as a CNN feature extractor to construct a four-layer CNN feature pyramid at different scales. For an OCT b-scan image, it is first fed into a CNN module containing three layers, each with a 1×1 convolution connected to a flanking three-layer Swin Transformer block, which compensates for the disadvantage of the Transformer that the information is lost in the lower levels and at the same time recovers the local spatial information.

As shown in Fig. 3(a), a traditional Transformer block [37] consists of Multi-head Self-Attention (MSA), Multi-Layer Perceptron (MLP), and

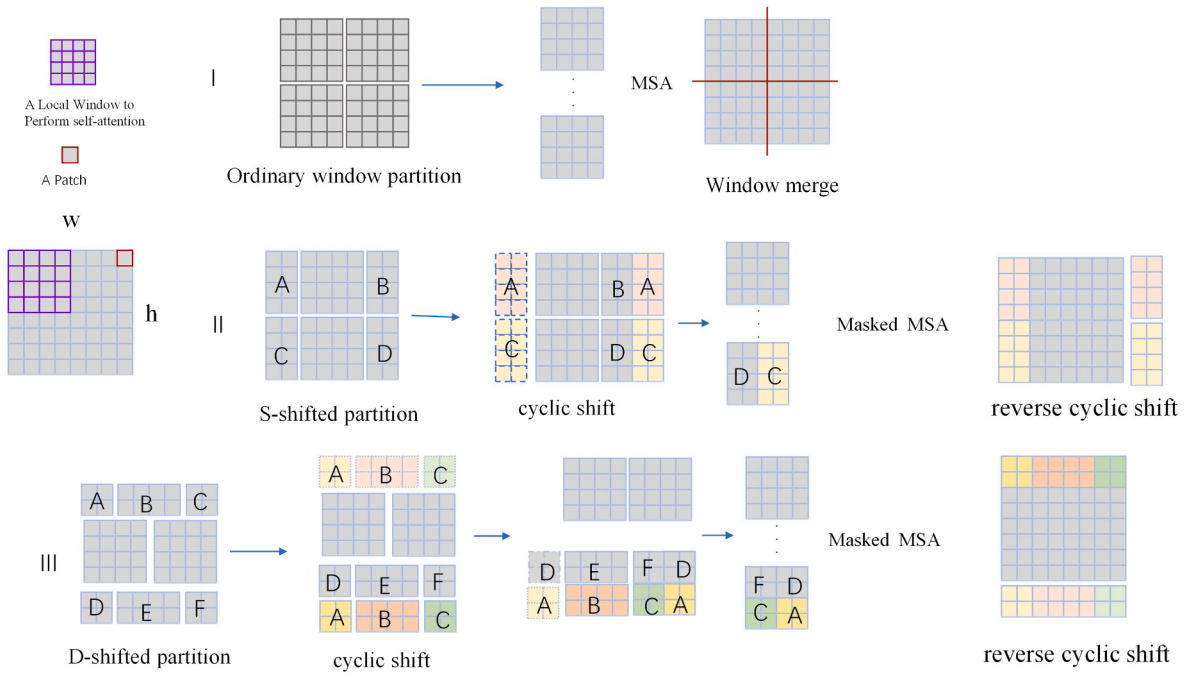

Fig. 5. Graphical representation of efficient computation of self-attention in Targeted window partitioning (S-shifted and D-shifted) partitions. I represent the standard W-MSA process. II and III represent two proposed windowing strategies.

Layer Normalization (LN). Thus, the output $\mathbf{z}^l$ of layer l can be expressed as:

$$\begin{aligned} \hat{\mathbf{z}}^l &= MSA\left(\mathrm{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \\ \mathbf{z}^l &= MLP\left(\mathrm{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l. \end{aligned} \quad \# \tag{1}$$

However, in the ViT model, each token requires computation based on its relationship with all others, resulting in a quadratic computational complexity. This is not feasible for many intensive prediction and high-resolution image tasks. In order to overcome this problem and improve the modeling, Swin Transformer proposes Window based Multi-Head Self-Attention(W-MSA) and Shifted Window based Self-Attention (SW-MSA) as an alternative to ordinary MSA. As shown in Fig. 3(b), the output of lth can be expressed as [25]:

$$\begin{aligned} \hat{\mathbf{z}}^l &= W\text{-}MSA\left(\mathrm{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \\ \mathbf{z}^l &= MLP\left(\mathrm{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l, \\ \hat{\mathbf{z}}^{l+1} &= SW - MSA\left(\mathrm{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l, \\ \mathbf{z}^{l+1} &= MLP - \left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1}. \end{aligned} \quad \# \tag{2}$$

where $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ denote the output features of the (S)WMSA module and the MLP module for block l, respectively; W-MSA and SW-MSA denote window based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

Different from the conventional fully-connected self-attention mechanism, W-MSA leverages windowing to handle large images, applying self-attention within a local window of size M × M. This approach offers linear complexity, reducing computational demands and memory usage. However, the partitioning of shifted windows in the consecutive blocks based on window-based self-attention modules lacks inter-window connections. And the modeling capability is limited. To address this limitation, Swin introduces SW-MSA, enhancing the model's capability to capture global image information by incorporating window offsetting and overlapping. This innovative approach enables the model to capture both local details and broader contextual information more effectively, enhancing its overall performance in capturing complex dependencies and global image features.

3.2. Hybrid shifted window-based MSA

SW-MSA solves the problem of interaction between blocks for biomarker images in OCT, and we find that only this one window shifting strategy cannot fully adapt to the distributional characteristics of OCT images. The first row in Fig. 4 ((a) for ERM and (b) for IRF) shows the shifting methods of Swin Transformer, SW-MSA. It can be observed from the green dashed boxes that using only Swin Transformer's shifting method disperses ERM and IRF into different windows. This limitation restricts the modeling capability of the model and may overlook some crucial features, making it difficult for the model to learn from the entire region. Therefore, inspired by [44], two new window shifting methods were added on top of the original SW-MSA to adapt to the different distributions of OCT biomarkers. As shown in the second row of Fig. 4, these two new shifting methods gather the biomarkers into one window and then perform self-attention calculations on these newly combined regular windows, thus increasing the receptive field and connectivity. As shown in Fig. 5 I, the division method of regular W-MSA divides 4x4 patches into one window and does the self-attention computation within each window, and Fig. 5 II, III shows two improved windowed sub-shifting strategies, which increase the number of windows by re-dividing the windows, and thus may lead to an increase in computational complexity within small windows. Inspired by Swin transformer, the smaller windows can be filled with M × M (e.g., 8 × 8 in the figure) by shifting, and shifting is equivalent to reorganizing the smaller windows into regular windows, which can control the number of computations of attention to 4 on the whole, which maintains consistency with the regular windows and masks out filled values during attention computation.

As shown in Fig. 5 II, after dividing the windows, we can shift all the windows to the left in an overall loop, at which time windows A and C are shifted to the rightmost part of the initial window. Then we merge windows A and C with windows B and D respectively, forming two new windows AB and CD. When computing the attention to AB, the spliced window, we should let the parts that were originally in the same window perform the attention calculation, and ignore the attention calculation between the two windows of AB, so if we need to get the correct result under the original window, we have to add a mask to the result of the

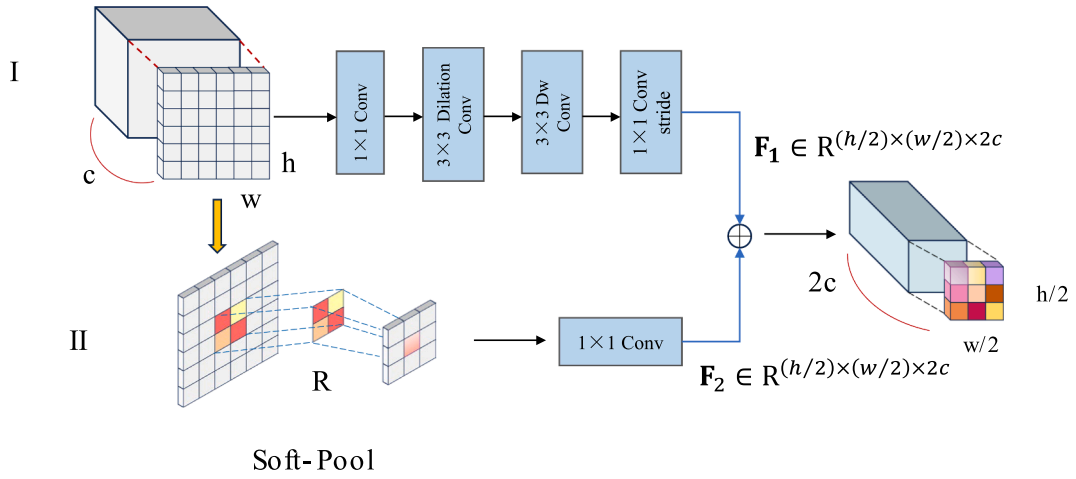

Fig. 6. Diagram of FDRM.

calculated attention to ignore the unnecessary part, the way of mask is similar to the way of mask in Swin, the mask is added to the result of attentions and softmaxed. The value of the mask is set to −100, and the corresponding value will be ignored after softmaxing. After that, we restore the feature map to the state before shifting for the next round of computation. Fig. 5 III is similar to Fig. 5 II, except that the downward shift is followed by a rightward shift in the local region. Our self-attention calculation method is consistent with the Swin Transformer.

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V\# \tag{3}$$

where $Q, K, V \in R^{M^2\times d}$ are the query, key and value matrices; d is the query/key dimension, and M^2 is the number of patches in a window. B is the relative position bias, $B \in R^{M^2\times M^2}$.

3.3. Feature Dimensionality Reduction module

The proposed new hybrid windowing strategy adjusts window partitioning based on the distribution characteristics of biomarkers. However, for different patients or different stages of AMD disease in the same patient, small biomarkers in OCT images may exhibit different sizes and shapes, presenting challenges in accurate recognition and segmentation. The downsampling process may dilute shallow fine details, further complicating segmentation tasks. In practical scenarios, there are complex relationships between foreground and background elements, as well as among foreground elements themselves. For instance, in OCT images, biomarkers often interact rather than exist in isolation [45].

Some self-attention-based methods [32,46] overlook small biomarker segmentation, risking detail and structural information loss. So, we propose FDRM (Feature Dimensionality Reduction Module) to enhance segmentation of small biomarkers, thereby improving results for dense and small-scale objects in OCT images. This module is tailored to tackle the challenges associated with segmenting small biomarkers, ultimately enhancing the segmentation performance for objects of varying scales.

As shown in Fig. 6, there are two branches in FDRM, the first involves a bottleneck block utilizing depth convolution and dilated convolution to gather comprehensive feature and structural details of small-scale objects. In the bottleneck block, the channel is firstly increased by a 1 × 1 convolution, and then a 3 × 3 dilated convolution layer is passed to obtain extensive structural information in the image, meanwhile, in order to keep the depth information of the feature map and effectively reduce parameters and computation, a 3 × 3 depth-wise convolution (Dw-conv) is added afterward, which allows for more efficient feature extraction. The output of this branch $\mathbf{F}_1 \in \mathrm{R}^{(h/2)\times(w/2)\times 2c}$. In another branch, in order to obtain finer downsampling results, we use soft-pooling [47], whereas many existing pooling methods employ combinations of max pooling and average pooling in different configurations, Soft Pooling utilizes softmax weighting to retain the fundamental characteristics of inputs while enhancing the activation of more intense features Unlike maxpooling, softpool is differentiable, so the network obtains a gradient for each input during backpropagation, which facilitates improved training. For the boxed local region R with dimension C × H × W, where C represents the number of channels, H denotes the height, and W signifies the width of the activation map in this branch, each activation $\mathbf{a}_i$ with index i is applied a weight $\mathbf{w}_i$,

$$\mathbf{w}_i = \frac{\exp(\mathbf{a}_i)}{\sum_{j\in\mathbf{R}}\exp(\mathbf{a}_j)}\# \tag{4}$$

Then multiply the corresponding feature values with the weights and do the total addition operation,

$$\tilde{\mathbf{a}} = \sum\nolimits_{i\in\mathbf{R}} \mathbf{w}_i{}^{*}\mathbf{a}_i\# \tag{5}$$

After soft pooling, the size of the image is reduced to half of the original size. The resulting feature map is then input into a 1x1 convolution to increase the dimension, which ultimately yields the output of this branch $\mathbf{F}_2 \in \mathrm{R}^{(h/2)\times(w/2)\times 2c}$.

We obtain the features of small-scale biomarker of OCT image by branch 1, and retain the details in it by branch 2, which can improve the detection and extraction ability of small targets. The two branches work together, and their results are ultimately combined by element-level addition to get the final output F of FDRM, which is functionally equivalent to the patch Merging module in the Swin Transformer block, but our module is more specific to the multi-class biomarker in OCT, which can improve the overall performance of the model.

$$F = \mathbf{F}_1 \oplus \mathbf{F}_2\# \tag{6}$$

3.4. Boundary Guided module

Supplementing the edge information of biomarkers in OCT images can improve the model's segmentation performance [48]. High-level features carry significant semantic information, aiding in biomarker localization and identification. On the other hand, low-level features contain intricate spatial details crucial for image refinement and segmentation. By integrating high-level semantic features with low-level spatial details, a more holistic understanding and processing of biomarker images can be achieved. As depicted in the left segment of Fig. 2, we leverage layers 2 and 3 of the CNN branch as inputs ($B^2 and B^3$) to the Boundary Guided Module (BGM). This approach provides a blend

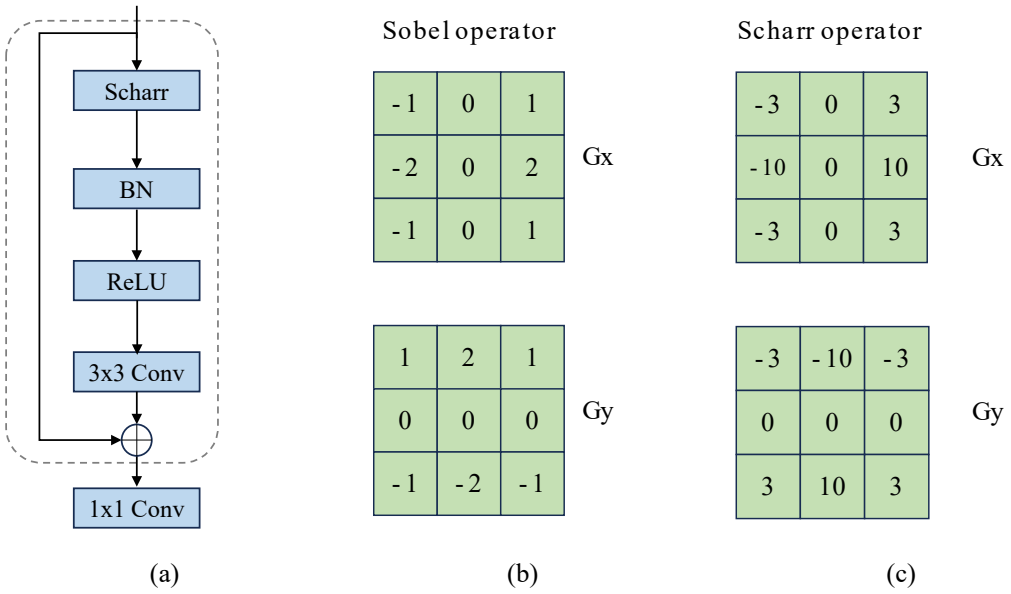

Fig. 7. (a) The architecture of the Boundary Guided Module (BGM) designed for edge feature enhancement. (b) Sobel operator. (c) Scharr operator.

of low-level detailed features and high-level semantic features, facilitating a synergistic fusion for comprehensive biomarker image comprehension and processing.

The architecture of the BGM is shown in Fig. 7(a). To enhance edge features, the BGM refines the features of each convolutional block. The Scharr operator is used in the BGM, which calculates the edge strength and orientation of each pixel point in the image based on a convolution operation on the first-order derivatives of the image. The Sobel operator (as shown in Fig. 7(b)) achieves edge detection by convolving the above templates with the image. For each pixel in the image, Sobel_x and Sobel_y are convolved with the 3x3 neighborhood around that pixel, and then the gradient values in two directions are combined to obtain the edge gradient of that pixel. Gx and Gy extract horizontal and vertical gradients, respectively. The Scharr operator (as shown in Fig. 7 (c)) is similar to the Sobel operator but is smoother in calculating the weights of the convolution kernel, which allows for a better handling of noise in the image and better detection of edges in the diagonal direction. The 1 × 1 convolution reduces the feature map to a single channel representation to obtain output features. Additionally, the output features from the 2-stage are combined with those from the 3-stage to incrementally enhance the richness of edge features. Ultimately, the output features are interpolated to restore the original input dimensions, thereby reconstructing the edge detection results.

3.5. *Transformer cross fusion module*

Feature fusion after two different encoders is an important issue, and ensuring that features from different models or different levels can be combined effectively is one of the keys to improving model performance. Fusing multi-scale features is common in many works [49,50], our main challenge is to combine two different scales of features from CNN and Transformer while maintaining consistency. The simplest approach is to just add the feature results directly. However, this approach may lead to feature inconsistency. Typically, shallow features contain more local details and localization information, while deeper features contain more abstract semantic information, and mid-level

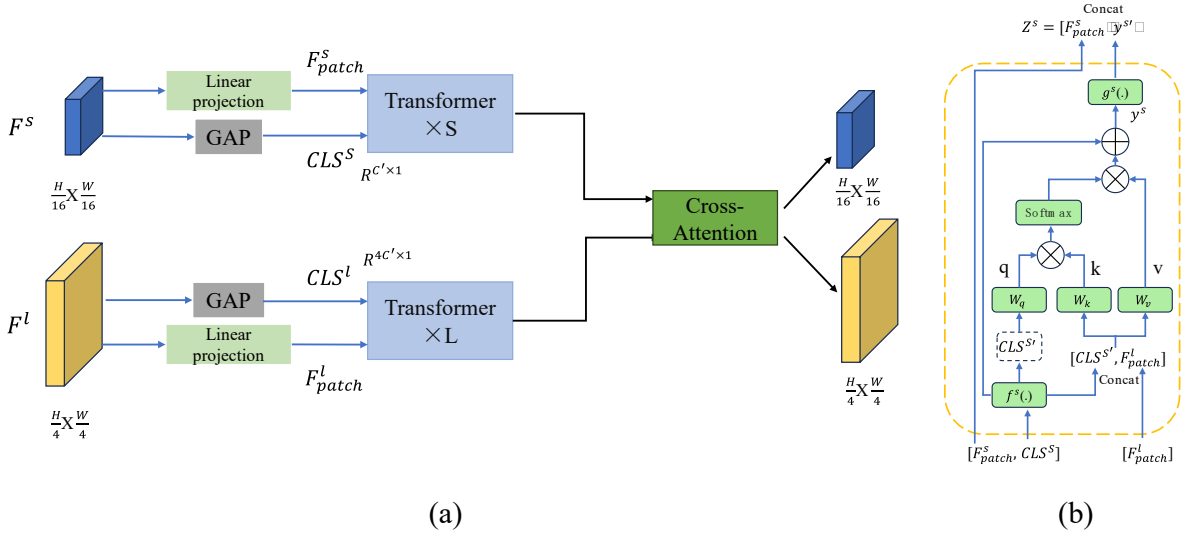

Fig. 8. (a) Our proposed TCFM. (b) Cross-attention module (Small branch). In the Cross-attention module for the small branch, the CLS token of the small branch acts as a query token, engaging in attention-based interactions with the patch tokens from the large branch. The functions $f^s(.)$ and $g^s(.)$ represent projections aimed at aligning dimensions. The large branch follows a similar process, with the CLS and patch tokens from another branch interchanged.

computational costs are huge and have little impact on model accuracy [51]. This distribution of features makes it crucial to choose the appropriate feature layers in the feature fusion and decoder part of the model, and inspired by [49], we note that the MSA mechanism can be utilized to perform multiscale feature processing on the first and third layers of features, preserving the localization information while fusing the features.

The proposed Transformer Cross Fusion Module (TCFM) is shown in Fig. 8 (a). The outputs of the first and third stages are denoted as large level F^l and small level F^s with sizes $H/4 \times W/4$ and $H/16 \times W/16$, respectively. Then obtain F^l_{patch} and F^s_{patch} through linear projection. We assign a class token to each of them by global average pooling, which summarizes all the information of the input features and has an important role.

$$\begin{aligned} CLS^l &= GAP(\mathrm{Norm}(F^l)) \\ CLS^s &= GAP(\mathrm{Norm}(F^s)) \end{aligned} \# \tag{7}$$

Where $CLS^l \in R^{C' \times 1}$, $CLS^s \in R^{4C' \times 1}$. Subsequently, the class tokens need to be associated with the corresponding level of embedding before being input into the Transformer. We chose the standard Transformer rather than the Swin Transformer block because the latter primarily operates on a grid-based feature map. Furthermore, we conduct single-level self-attention operations twice at each stage, ensuring manageable computational complexity. Additionally, we introduce learnable position embeddings for each marker at both levels, which are then incorporated into the Transformer to capture positional information effectively.

After embedding through the Transformer encoder, the cross-attention module is used to fuse the features of each level (small and large). Prior to fusion, we interchange two levels of class tokens, linking labels from one level to those of another. Subsequently, each new embedding undergoes fusion within modules before being projected back to its original level. This cross-level interaction allows class tokens to access and exchange rich information across different levels, facilitating enhanced information exchange and collaboration. In particular, Fig. 8(b) shows details of the cross-attention for the small level. $f^s(.)$ and $g^s(.)$ are the projection and back-projection function for dimension alignment, respectively. $f^s(.)$ first projects the CLS^s into the dimension F^l_{patch} and the output is denoted as $CLS^{s'}$.

$CLS^{s'}$ is concatenated with F^l_{patch} as keys and values, and is executed independently as a query to compute attention, the process is denoted as cross-attention (CA). The cross-attention mechanism runs in linear time since we only use CLS in the query. Additionally, just like self-attention, multiple heads are also used in the cross-attention, denoted as MCA. The final output Z^s can be written mathematically as follows:

$$\begin{aligned} y^s &= f^s(CLS^s) + MCA\left(LN\left(\left[f^s(CLS^s) \middle\| F^l_{patch}\right]\right)\right) \\ Z^s &= \left[F^s_{patch} \middle\| g^s(y^s)\right] \end{aligned} \# \tag{8}$$

3.6. Decoder

In order to combine two feature maps of different sizes obtained via TCFM into a unified feature mask, we design a decoder to fuse the low-resolution feature map F^s (H/16, W/16) with the high-resolution feature map F^l (H/4, W/4). The feature map F^s is first passed through the Conv1 block, which is used to recover the feature map size to (H/4, W/4), which consists of two stages of 3 × 3 Conv, 2 × bilinear upsampling, Group Norm, and ReLU. Group Normalization is applied to normalize the feature map by dividing the feature channel into multiple groups and normalizing each group, which helps to reduce the internal covariate bias and improves the training stability of the model. For F^l, we do not need to perform feature recovery, it passes through 3 × 3 Conv, Group Norm and ReLU and maintains (H/4, W/4) resolution. Then the processed F^s and F^l are concatenated, and passed through a Conv1 block to obtain the final H × W feature map. The obtained final feature map is fed into the segmentation head to obtain the final segmentation result. The process is shown as:

$$F_{seg} = Seghead(Conv1(Concat(Conv1(F^s), Conv2(F^l)))) \# \tag{9}$$

3.7. Loss function

The proposed segmentation network consists of two loss functions L_{seg} and L_{edge}.L_{seg} represents the loss function used in the semantic segmentation module to learn semantic features. In OCT images, the foreground area is usually small, and simply using dice loss would make the model more inclined to predict background pixels. To address this, we combine Focal loss and Dice loss and define L_{seg} as the

$$L_{Focal} = -\sum_{i=1}^{i} \varepsilon_i (1 - p_t)^{\gamma} \log(p_t) \# \tag{10}$$

$$L_{Dice} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{2\sum_{j=1}^{N} G_{ij} P_{ij}}{\sum_{j=1}^{N} G_{ij} + \sum_{j=1}^{N} P_{ij}} \# \tag{11}$$

$$L_{seg} = \lambda_{focal} L_{Focal} + \lambda_{dice} L_{Dice} \# \tag{12}$$

where ε_i is a hyperparameter that represents the weight assigned to the i-th class. γ is a hyperparameter used to adjust the weight between easy to classify and difficult to classify samples. p_t denotes the model's predicted confidence in sample classification. In equation (11), N denotes the count of elements within class i. Then G_{ij} and P_{ij} respectively represent the j-th element of class i in ground truth and predicted labels. This method can help the model better understand the semantic information in the images and more accurately segment biomarkers. L_{edge} represents the loss function used in the edge detection module for learning edge features. For the edge detection problem in biomarker, the category imbalance problem is important because for a class of biomarkers, most samples are negative. To solve this problem, we combine the dice loss and the edge loss Richer Convolutional Features (RCF) proposed by Liu et al [52], which combines deep learning and traditional edge detection methods, and can effectively distinguish the edge information in the extracted image.

$$L_{edge} = L_{Dice} + 0.5 L_{RCF} \# \tag{13}$$

Here is the definition of L_{RCF}:

$$L_i^j = \begin{cases} \alpha \cdot \log\left(1 - \hat{y}_i^j\right) & \hat{y}_i = 0 \\ 0 & 0 < \hat{y}_i \le \eta \\ \beta \cdot \log \hat{y}_i^j & other \end{cases} \# \tag{14}$$

where $\hat{y}_i^j$ denotes the predicted value for the i-th pixel in the j-th edge map, with η representing a predefined threshold. This means that pixels with an edge probability higher than η will be considered as positive samples, while pixels with an edge probability equal to 0 will be considered as negative samples. β represents the percentage of negative samples in the total sample. $\alpha = \lambda(1 - \beta)$, where λ is a hyperparameter that balances positive and negative samples. The total loss L is defined as:

Table 1
The distribution of the RB dataset and RK dataset. *Mix** indicates that in an OCT B-scan, there are more than two different biomarkers present.

Dataset	Biomarker Label	Label Color	OCT B-scans			
			Total	Training	Validation	Testing
RB	ERM		270	162	54	54
	PED		160	96	32	32
	IRF		290	174	58	58
	SRF		260	156	52	52
	CNV		160	96	32	32
	HF		290	174	58	58
	GA		180	108	36	36
	DRUSEN		190	114	38	38
	*Mix**	*	1000	600	200	200
	All	*	2800	1680	560	560
RK	PED		530	318	106	106
	IRF		650	390	130	130
	SRF		630	378	126	126
	CNV		480	288	96	96
	HF		530	318	106	106
	DRUSEN		550	330	110	110
	*Mix**	*	1130	678	226	226
	All	*	4500	2700	900	900

$$L = L_{seg} + \lambda_{edge} L_{edge} \# \tag{15}$$

The overall process of our method is shown in Algorithm 1.

Algorithm 1 The proposed algorithm.

Input: X_n, Y_n; // X_n *represents the training set for OCT B-scan biomarkers*; Y_n *is manually labeled annotation*;
Output: $S(\cdot)$ *with parameters* θ; // $S(\cdot)$ *is the proposed network model*;
for *all* $e \in \{1, 2, \cdots, epoch\}$ ***do***; // *Perform epoch traversal*;
for *all* $b \in \{1, \cdots, batch\}$ ***do***;
get data from $\{(X_n, Y_n)\}_1^B$; // *get input B-scan and label*;
$B^2, B^3 = CNN(X_n)$; // *input the image into the MSCS, obtain the edge outputs* $\{B^2, B^3\}$;
$F^l, F^s = CNN - Transformer(X_n)$; // *Same as the previous line, obtain the outputs* $\{F^l, F^s\}$;
(*continued on next column*)

(*continued*)

$F^s, F^l = TCFM(F^l, F^s)$; // *input* F^i *into Transformer Cross Fusion Module, get outputs* $\{F^s, F^l\}$;
$P = Decoder(F^s, F^l)$; // *input the output* F^i *into the decoder, obtain image prediction map* P;
$B = BGM(B^2, B^3)$; // *input the output* B^i *into the BGM, get the boundary map* B;
$g_\theta^{seg} = -\nabla_\theta L_{seg}(P, Y_n)$; // L_{seg} *refers to the segmentation loss, see equation* (12);
$g_\theta^{Dice} = -\nabla_\theta L_{Dice}(P, Y_n)$; // L_{Dice} *refers to the segmentation loss, see equation* (11);
$g_\theta^{edge} = -\nabla_\theta L_{edge}(B, Y_n)$; // L_{edge} *refers to the boundary loss, see equation* (13) *and* (14);
$\theta \leftarrow \theta + lr \cdot Adam\left(\theta, \lambda_1 g_\theta^{seg} + \lambda_2 g_\theta^{edge}\right)$; // *update parameters, as shown in equation* (15);
end for;
end for;

4. Experimental results

4.1. Dataset

Retinal Biomarker (RB) dataset. The retinal biomarker dataset is obtained by the Wuhan Aier Eye Hospital using a DRI OCT Triton scanner with a resolution of 1024 × 992.This dataset contains 2800 OCT B-scans, which includes 8 different retinal biomarkers (i.e., PED, IRF, SRF, CNV, ERM, GA, HF, and DRUSEN). Each OCT biomarker image contains one or more types of biomarkers. When using the RB dataset, we partition it into training, validation, and testing sets based on ratio of 60 %-20 %-20 %, i.e. And it contains 1680, 560, and 560 images, respectively. Table 1 shows the distribution of the RB dataset and we resized all the images to 512 × 512.

RETOUCH dataset [53]. RETOUCH contains a total of 112 individual data acquired from three different devices, including a training set containing 70 labeled and a test set containing 42 unlabeled, for a total of 6936 OCT B-scan slices. It has 3 different types of biomarker images, the IRF, SRF, and PED, all manually annotated by experts. **Kermany's dataset** [54]. Kermany's dataset compris108,312 OCT B-scans, (37,206 with choroidal neovascularization, 11,349 with diabetic macular edema, 8,617 with drusen, and 51,140 normal). The experts select images with the following five biomarkers from Kermany's dataset: PED, CNV, IRF, DRUSEN and HF, and with the help of experts, the required B-scans are annotated.

In the process of medical clinical diagnosis, there are differences in the OCT images obtained due to the different equipment used to acquire the images, for this reason, in our study, we construct a hybrid dataset based on two datasets, RETOUCH and Kermany. The hybrid dataset evaluates our model's capability across different devices and allows us to assess its accuracy and generalization. Table 1 shows the distribution of the hybrid dataset of RETOUCH and Kermany (**RK dataset**). We divide the hybrid dataset containing 4500 OCT images into three subsets according to the ratio of 6:2:2, which are used for training, testing and validation respectively. For the images of different sizes in the dataset, we resize them to the same size of 512 × 512.

AMD-SD dataset [55]**.** The dataset contains 3,049 OCT B-scan images and their corresponding labels. All images have a resolution of 570 × 380 pixels. These images are from 156 eyes of 138 patients, including 61 females and 77 males. The average age of the patients is 66.7 years, with a standard deviation of 9.1 years. In addition to the three types of biomarker labels (IRF, PED and SRF), the dataset also includes SHRM. SHRM may include various constituents such as exudates, fibrosis, blood, scars, or choroidal neovascularization. The dataset is divided into three subsets for training, validation, and testing in a ratio of 6:2:2. The image size is resized to 512 × 512.

4.2. Comparison methods and evaluation Protocol

In our experiments, the performance of the models is compared with common CNN segmentation models including U-Net [12], Deeplabv3+

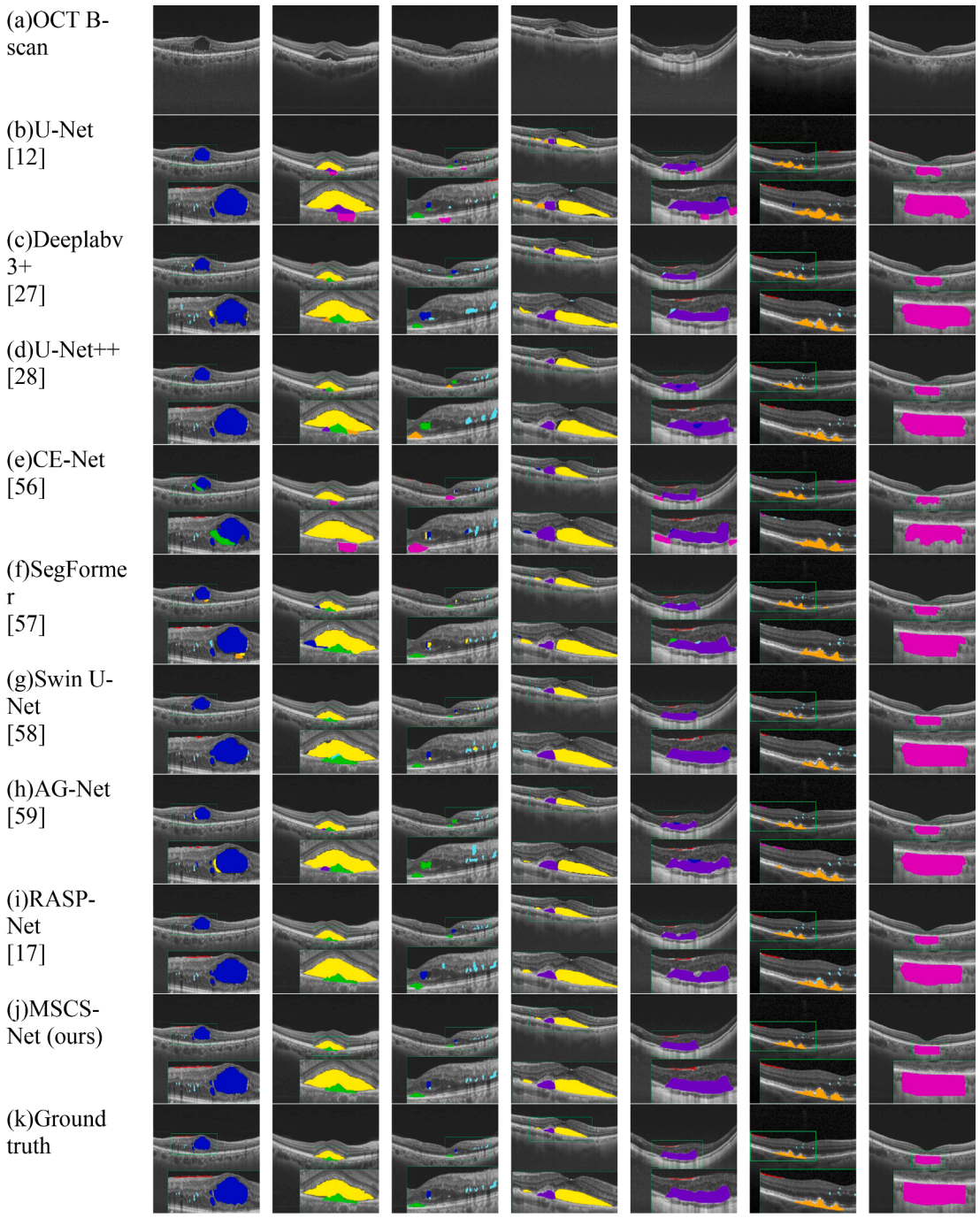

Fig. 9. Comparison of qualitative results between MSCS-Net and existing models on local datasets. The corresponding colors for each biomarker are shown in Table 1. ■ ERM, ■ PED, ■ IRF, ■ SRF, ■ CNV, ■ HF, ■ GA, ■ DRUSEN.

[27], and medical image segmentation models including U-Net++ [28], CE-Net [56], and models of Transformer including SegFormer [57], Swin U-Net [58], and some improved medical segmentation methods including RASP-Net [17], AG-Net [59]. We use PyTorch to build all the models required for the experiments and train them on NVIDIA RTX 3090 GPUs. In our experiments, the proposed MSCS model employ data augmentation techniques, including flipping and rotation. Specifically, images and their corresponding labels are randomly rotated by 0, 90, 180, or 270 degrees (rotation by multiples of 90 degrees) and flipped along a randomly chosen axis (horizontal or vertical). The learning rate is set to 1e-4, momentum to 0.9, weight decay to 1e-5, and the batch size is set to 8. We train the model for a total of 250 epochs. To ensure a fair comparison, we apply the same settings to all baseline models, and each model uses the loss functions as described in the original papers.

In this study, two evaluation metrics, Dice similarity coefficient (DSC) and Jaccard index (Jac) are utilized to evaluate the segmentation performance of various methods. They are both used to evaluate the similarity between predicted samples and ground truth according to the following equations:

$$DSC = \frac{2 \times TP}{2 \times TP + FN + FP}\# \tag{16}$$

$$Jac = \frac{TP}{TP + FN + FP}\# \tag{17}$$

Table 2
Localization and segmentation performance (Jac and Dice) compared with different methods on the RB dataset (unit: %). The best results are indicated in bold, and the second-best results are indicated in italics. * and ★ indicates the p-value of the paired *t*-test between each method with the proposed MSCS-Net. (*: $p < 0.01$, ★: $p < 0.05$).The 95 % confidence intervals are represented in parentheses.

Networks (Params)	ERM		PED		IRF		SRF		CNV		HF		GA		DRUSEN		Overall	
	Jac		Jac		Jac		Jac		Jac		Jac		Jac		Jac		Jac	Dice
U-Net (4.42 M)	64.93*(63.64, 66.22)		57.13*(55.84, 58.42)		58.07*(56.78, 59.36)		62.38*(60.09, 64.67)		63.29*(61.50, 65.08)		47.90*(45.61, 50.19)		61.07*(59.28, 63.76)		62.54*(60.65, 64.43)		59.66*(57.64, 61.68)	74.89*(73.14, 76.64)
Deeplabv3+ (58.50 M)	71.14	(69.85,72.43)	62.11*(60.32, 63.90)		*68.16*(66.06, 70.26)		73.02(71.03, 75.11)		73.66★	(72.17,75.15)	55.12*(53.33, 56.91)		75.98*(74.83, 77.13)		69.37(67.82, 70.92)		68.57★	*79.45**
U-Net++ (32.15 M)	68.01*	(66.51,69.51)	58.78*	(56.38,61.18)	59.75*	(57.50,62.00)	62.25*	(59.94,64.56)	74.06	(72.69,75.43)	44.10*	(41.72,46.48)	77.32	(75.81,78.83)	65.63*	(64.20,67.06)	*(67.32,69.82)*	*(78.11,80.79)*
CE-Net (190.04 M)	67.65*	(66.23,69.07)	57.32*	(55.27,59.37)	61.93*	(59.77,64.09)	65.37*	(62.79,67.95)	63.46*	(61.68,65.24)	49.11*	(47.10,51.12)	54.66*	(53.00,56.32)	61.79*	(59.96,63.62)	60.16*	(58.17, 62.15)	73.18*	(71.38, 74.98)
SegFormer (46.73 M)	66.70*	(65.74,67.66)	64.22*	(63.17,65.27)	62.60*	(61.40,63.80)	66.01*	(65.06,66.96)	67.65*	(66.92,68.38)	49.90*	(47.16,52.64)	70.16*	(69.28,71.04)	66.20*	(64.84,67.56)	64.18*	(62.84, 65.52)	76.05*	(74.87, 77.23)
Swin U-Net (27.17 M)	68.63*	(67.69,69.57)	61.03*	(59.87,62.19)	62.19*	(61.26,63.12)	66.34*	(65.32,67.36)	**77.18**	(76.36,78.00)	49.86*	(46.87,52.85)	**79.38**	(78.78,79.98)	64.05*	(63.03,65.07)	66.08*	(64.93, 67.23)	77.10*	(75.99, 78.21)
AG-Net (30.98 M)	71.30	(69.96,72.64)	65.25*	(63.81,66.69)	64.03*	(62.70,65.36)	*73.33*	(71.65,75.01)	75.21	(73.95,76.47)	56.99*	(55.51,58.47)	73.87*	(72.53,75.21)	68.71*	(67.37,70.05)	68.59*	(67.06, 70.12)	79.38*	(78.02, 80.74)
RASP-Net (88.21 M)	*71.54*	(70.33,72.75)	64.45*	(62.82,66.08)	68.15	(67.25,70.05)	72.71★	(71.16,74.26)	74.60	(73.19,76.01)	53.27*	(51.61,54.93)	75.11*	(73.83,76.39)	*69.43*★	(68.02,70.84)	68.66★	(67.34, 69.98)	79.96*	(78.92, 80.00)
MSCS-Net (ours) (27.67 M)	**72.52**	(71.49,73.55)	**70.11**	(68.99,71.23)	**69.76**	(68.69,70.83)	**75.84**	(74.54,77.14)	*76.65*	(75.47,77.83)	**60.23**	(58.92,61.54)	*79.12*	(78.04,80.20)	**72.10**	(70.96,73.24)	**72.04(70.96, 73.12)**		**82.53(81.58, 83.48)**	

4.3. Comparison with other methods on the RB dataset

Fig. 9 and Table 2 show the qualitative results and quantitative results of our proposed MSCS-Net on the RB dataset compared with the other eight state-of-the-art methods, respectively. By observing and analyzing the results we can see that for U-Net in Fig. 9(b), there are many mis-segmentation and under-segmentation problems, which indicates that U-Net has a weak ability to capture small targets and poor anti-interference. For example, for the segmentation of HF (column 3), U-Net is prone to mistaking some HF for other biomarkers, or failing to recognize HF. In contrast, such as Fig. 9 (d) U-Net++ adopts a better upsampling strategy and introduces a multi-level, dense feature-connecting path, which allows the network to better fuse feature information from different levels, which improves the feature characterization and achieves better results than U-Net. As shown in Fig. 9 (f) and (g), both SegFormer and Swin U-Net are networks constructed with different Transformers as the basic components, which have more powerful global modeling capability than convolutional operations, and also retain a part of the CNN's ability to extract features efficiently, and effective segmentation results are achieved for some larger biomarkers (CNV, GA, etc. column 5 and 7). However, SegFormer is deficient in capturing small-scale targets, resulting in poor segmentation results for small objects (Such as in Fig. 9 (f) column 1, some middle parts of HF are ignored, and smaller IRF is mis-segmented as HF). Although Swin Block adopted by Swin U-Net uses a better windowing strategy to reduce the loss of details, it is still deficient in extracting small targets. As shown in Fig. 9 (c), DeepLabv3 + combines a multi-scale feature fusion and spatial pyramid strategy, which alleviates the problem of mis-segmentation to a certain extent by fusing feature maps of different scales for multiple classes of biomarkers in a single map, but it mainly focuses on capturing a wider range of contextual information in which the detailed information of the underlying features may be ignored. RASP-Net also combines spatial pyramids and incorporates novel extended residual blocks, but it fails to take into account the interactions between layers, which ultimately leads to less fine-grained segmentation results (shown in Fig. 9 (i) column 3 and 5).AG-Net utilizes an attention mechanism to suppress image noise and background interference, thereby enhancing the network's capability to learn complex features. This results in superior performance compared to U-Net (shown in Fig. 9 (h) column 1 and 3). Our segmentation results are closer to GT compared to other methods. These observations demonstrate the merits of the proposed MSCS-Net.

In deep learning, multiple convolutions generate features of different scales. The proposed CNN and Transformer framework is designed to capture features of different scales for better segmentation of different-sized biomarkers in OCT images. Furthermore, we have also introduced TCFM to fuse features of different scales, thereby enhancing the effectiveness of the model. As shown in Fig. 9 (j), our method does not mis-segment biomarkers compared to other segmentation methods and have higher overlap with ground truth. Moreover, our proposed FDRM module and new target shifting strategy enhance the ability of segmenting small biomarkers without under-segmentation problem for HF, DRUSEN, etc (as shown in the 1st, 3rd, and 6th columns in Fig. 9 (j)). Our method demonstrates superior accuracy compared to others in segmenting adjacent fluid biomarkers. This is attributed to our BGM module, which prioritizes edge region pixels, thus mitigating challenges posed by low image contrast and blurred edges. Overall, the results shown in Fig. 9 effectively demonstrate the efficacy of the strategy employed in our method.

Table 2 lists the quantitatively evaluated performance of our method and other methods, the best results are indicated in bold, and the second-best results are indicated in italics. As evident from the table, our method achieves superior results in the segmentation of most biomarkers, reflecting the advantages of our method. For fluid biomarker (PED, IRF, SRF), where the fluid region is similar to the background but different from the retinal layer, the proposed method has comparable

Table 3
Ablation results (Dice) of proposed blocks on the RB dataset. A: BGM. B: A + HW-MSA and FDRM. C: B + TCFM.

	ERM	PED	IRF	SRF	CNV	HF	GA	DRUSEN	Average
Baseline(U-Net)	77.19	68.92	70.24	74.61	76.10	63.33	74.69	74.06	72.39
Baseline + A	78.19	71.91	72.05	76.83	77.23	64.34	76.01	74.97	73.94
Baseline + B	82.54	77.56	77.68	81.27	83.29	71.21	82.47	79.63	79.45
Baseline + C	84.30	80.45	80.07	85.50	85.92	75.58	86.78	81.66	82.53

results in IRF segmentation compared to RASP-Net, but achieves superior results in PED and SRF segmentation, with Jac improved 5.66 points (from 64.45 to 70.11) and 3.13 points (from 72.71 to 75.84), respectively. It is worth noting that for smaller biomarker HFs, which are indistinguishable due to their similar contrast, the segmentation results for most networks are not particularly good, with Jac in the range of 47 % to 53 %, and the segmentation results are greatly improved as a result of our proposed FDRM for small biomarkers and the new windowing strategy, Jac reaches 60.23 %. For larger biomarkers (CNV, GA), we found that the methods applying multi-scale and attention mechanisms (Deeplabv3+, Swin U-Net, AG-Net and RASP-Net) achieved good results, among which, Swin U-Net achieved the best results, with two biomarkers' Jac is 77.18 % and 79.38 %, respectively, and our method is comparable with its results. For ERM segmentation results, our method also achieved the best performance, with Jac score reaching 72.52 %. This surpasses the best scores of comparative methods by absolute differences of 0.98 points and 1.30 points. The analysis of both qualitative and quantitative results shows that our method outperforms the other methods overall, further validating the effectiveness of our method.

4.4. Ablation experiments

In this section, we conduct ablation studies on RB dataset to assess the effectiveness of each component in our proposed MSCS-Net. The experimental findings are detailed in Table 3, with the Baseline being U-Net. A corresponds to the BGM for edge detection, B includes A plus modules focusing on small target features (HW-MSA and FDRM), and C integrates B with the final fusion module TCFM, forming the complete MSCS model structure.

Comparing the first and second rows of Table 3, it can be observed that the use of the edge detection path with BGM for edge-assisted segmentation has enhanced the localization of biomarkers and optimized edge segmentation. The metrics for all biomarkers in the segmentation network are higher than those of the baseline, showing an average Dice improvement of 1.55 points. However, U-Net's relatively simple structure relies solely on skip connections to integrate contextual information within the same layer, it lacks the ability to differentiate between biomarkers of different sizes. The improvement for small targets is not significant, with a Dice improvement of 1.01 points for HF and only 0.91 point for DRUSEN. The most significant improvement is for PED, with a Dice improvement of 2.99 points.

From the comparison between the Baseline + A and Baseline + B rows in Table 3, it can be seen that with the addition of the Swin Transformer module and the corresponding HW-MSA and FDRM after the second encoding path, the average Dice is improved by 5.51 points (from 73.94 to 79.45) compared to the U-Net with only the edge detection path. The segmentation results for all biomarkers have been greatly improved. This indicates that after adding the Swin Transformer branch, the network is able to combine the long-range contextual interaction capability of Swin Transformer and perform multi-scale information exchange through CNN-Transformer. This can further enhance the discriminative ability for biomarker features of different sizes. For large biomarkers such as CNV, the Dice is improved by 6.06 points (from 77.23 to 83.29), and for SRF, the Dice is improved by 4.44 points (from 76.83 to 81.27). In particular, with the addition of FDRM, through its two branches, the detection and extraction capabilities for small biomarkers are strengthened, highlighting the spatial information and remote dependence between biomarkers. This has led to significant improvements in the segmentation results for small biomarkers. For example, the Dice is improved by 6.87 points (from 64.34 to 71.21) for HF and 5.63 points (from 72.05 to 77.68) for IRF. In the last row of Table 3, we evaluate the impact of TCFM on the network, and the results reveal the non-negligible role of TCFM module in the encoding and decoding process. Specifically, the TCFM module brings a large improvement of 3.08 points (from 79.45 to 82.53) to Dice. The TCFM module leverages the cross-attention mechanism to aid the network in integrating global and local features, and exchanging and fusing information from different scales. The results demonstrate that the convenient combination of CNN and Transformer helps to segment the biomarker.

Fig. 10 shows the qualitative results of the ablation experiment. Fig. 10 (b) displays the segmentation results of U-Net, where in the second row, some HF are mis-segmented as SRF, and some foreground objects are mis-segmented as other biomarkers (GA, IRF). This is attributed to U-Net's challenge in effectively integrating detailed low-level information with high-level semantic information, both of which are critical for segmenting multiple classes of OCT biomarkers. Additionally, the segmentation of boundary areas is not accurate enough.

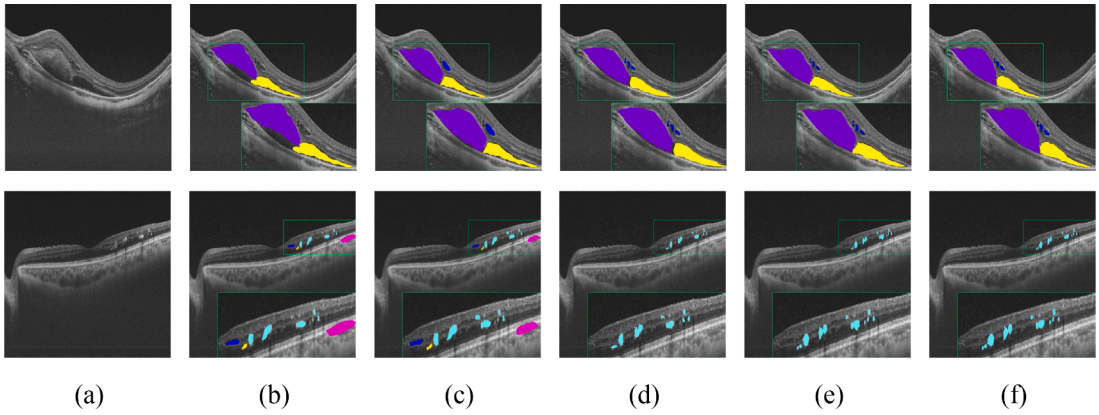

Fig. 10. Qualitative results of the ablation experiment of the proposed blocks on the RB dataset. (a) OCT B-scans; (b) Baseline. (c) Baseline + A (BGM for the edge detection path). (d) Baseline + B (HW-MSA and FDRM). (e) Baseline + C (TCF). (f) ground truth.

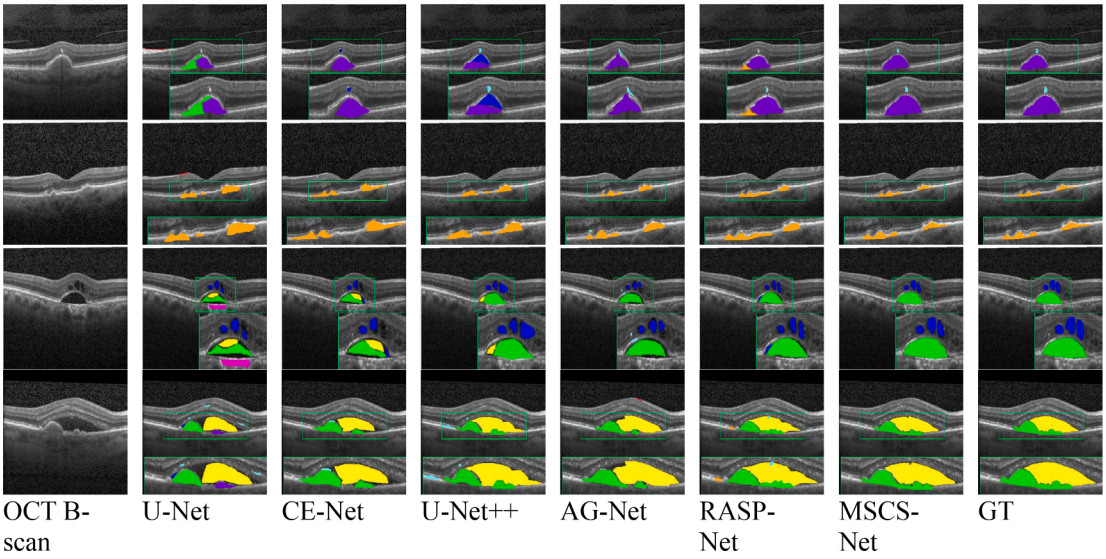

Fig. 11. Visualized results of comparative methods on the RK dataset. The corresponding colors for each biomarker are shown in Table 1.

Table 4
Performance comparison of our network with different methods on RK datasets. The best results are indicated in bold. (unit: %). * ★ indicates the p-value of the paired *t*-test between each method with the proposed MSCS. (*: $p < 0.01$, ★: $p < 0.05$). The 95 % confidence intervals are represented in parentheses.

Networks	PED	IRF	SRF	CNV	HF	DRUSEN	Overall	
	Jac	Jac	Jac	Jac	Jac	Jac	Jac	Dice
U-Net	56.19* (54.28,58.10)	49.61* (46.81,52.41)	60.47* (58.80,62.14)	70.68* (69.68,71.68)	45.27* (43.69,46.85)	53.67* (51.78,55.56)	55.98* (54.02,57.94)	65.25* (63.53,66.97)
CE-Net	58.36* (56.26,60.46)	51.36* (48.07,54.65)	63.26* (61.65,64.87)	70.95* (69.66,72.24)	48.64* (46.34,50.94)	50.10* (47.99,52.21)	57.11* (55.17,59.05)	66.82* (65.09,68.55)
U-Net++	64.48* (62.13,66.83)	49.87* (47.89,51.76)	70.11 (68.06,72.16)	71.32* (70.31,72.33)	46.55* (44.45,48.65)	55.54* (53.78,57.30)	59.64* (57.74,61.54)	68.06* (66.41,69.71)
AG-Net	66.42* (64.73,68.11)	56.77* (53.79,59.75)	66.39* (64.49,68.29)	73.36★ (72.59,74.13)	52.49* (51.20,53.78)	58.49* (56.91,60.07)	62.32* (60.95,63.69)	70.66* (69.47,71.85)
RASP-Net	68.79 (67.12,70.46)	60.20* (58.77,61.63)	68.55★ (66.82,70.28)	74.21* (70.79,71.63)	54.74★ (52.39,57.09)	61.22 (59.63,62.81)	64.62★ (63.10,66.14)	74.54* (73.30,75.78)
MSCS-Net	**70.10** (69.12,71.08)	**63.51** (62.36,64.66)	**72.09** (70.86,73.32)	**75.17** (74.30,76.04)	**58.16** (57.12,59.20)	**64.36** (62.63,66.09)	**67.23** (66.68,67.78)	**77.86** (76.82,78.90)

foreground as HF. On the other hand, RASP-Net accurately identifies DRUSEN but still has over-segmentation issues. Our network has more accurate localization and achieves better segmentation results of DRUSEN.

Fig. 10 (c) shows that by adding the edge detection pathway, the model first utilizes boundary supervision to determine the biomarker region, focusing more on the foreground region. The edge features are refined through BGM, which partially alleviates the problem of under-segmentation and improves the segmentation results of the edges. However, there are still issues with the identification and segmentation of small targets, as well as the lack of global information. Fig. 10 (d) shows that after integrating Swin Transformer and the segmentation module for small targets, the model effectively improves its classification ability for different biomarkers through multi-scale features. Additionally, with the help of the FDRM, the model highlights the features of small biomarkers, greatly improving the recognition rate of HF and other biomarkers, and significantly enhancing the segmentation effect of small biomarkers. Fig. 10 (e) demonstrates that after integrating the TCFM, the further fusion of different scales of information through cross-attention can enhance the model's ability to perceive different scales of information, and improve the model's performance and accuracy when dealing with multi-scale tasks. Based on the experimental results above, it can be concluded that all the designed components play essential roles in the OCT biomarker segmentation task.

4.5. Comparison with other methods on the mixed RK dataset

Our qualitative results and quantitative results on the public dataset are shown in Fig. 11 and Table.4. As shown in the first row of Fig. 11, for HF, none of the networks except ours have been able to recognize HF completely, U-Net recognizes the least regions. U-Net has the most serious under-segmentation problem for biomarkers with large biomarkers (such as CNV). CE-Net also has a mis-segmentation problem, AG-Net and RASP-Net have made some improvements to the under-segmentation problem. Ours has not missed HF and has a more accurate segmentation of the CNV. In the second row, U-Net mis-segments some of the foreground as ERM and has under-segmentation issues. CE-Net severely over-segments DRUSEN, and both CE-Net and AG-Net mistakenly segment a small portion of translucent.

As shown in the third and fourth rows of Fig. 11, all other networks have under-segmentation issues for PED and SRF, and U-Net and U-Net++ have mis-segmentation problems. For the small biomarker IRF, due to its similarity in color to the background, U-Net and CE-Net are unable to fully recognize and locate IRF. While RASP-Net identified IRF, it also had under-segmentation issues. In contrast, our network is capable of effectively recognizing and localizing biomarkers of different sizes, reducing the impact of foreground and background on the segmentation results. The segmentation accuracy is higher for smaller

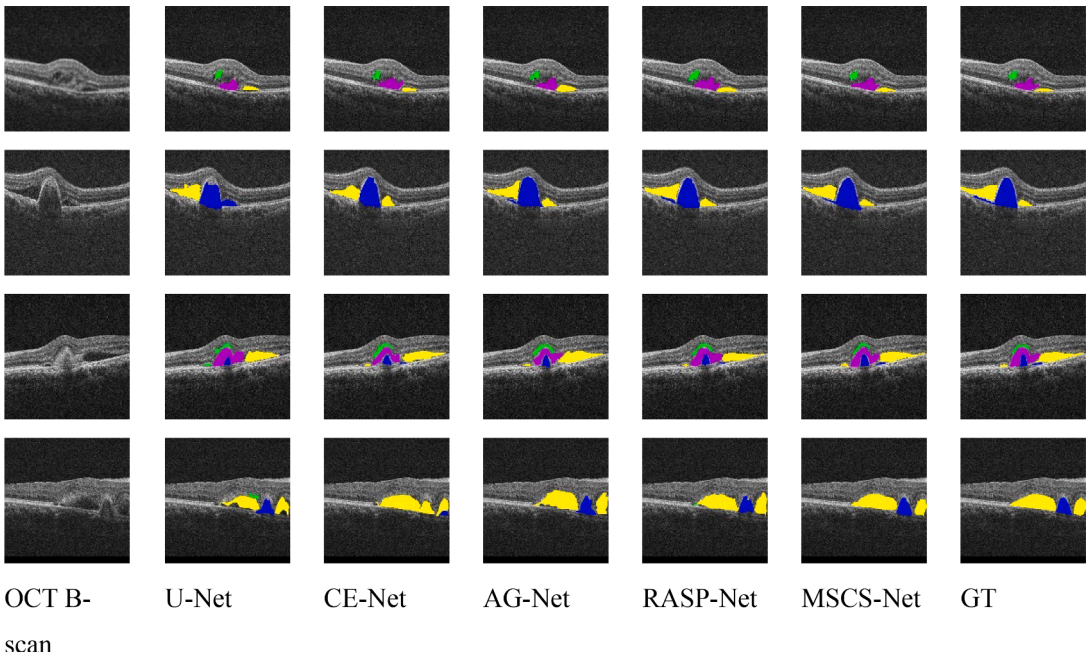

Fig. 12. Visualized results of comparative methods and our method on the AMD-SD dataset. ■ SRF, ■ PED, ■ IRF, ■ SHRM.

Table 5
Segmentation performance (Jac and Dice) compared with different methods on the AMD-SD dataset (unit: %). * and ★ indicate the p-value of the paired *t*-test between each method with the proposed MSCS-Net. (*: $p < 0.01$, ★: $p < 0.05$). The 95 % confidence intervals are represented in parentheses.

Networks	SRF		PED		IRF		SHRM		Overall	
	Jac	Dice	Jac	Dice	Jac	Dice	Jac	Dice	Jac	Dice
U-Net	60.36* (58.61, 62.11)	71.06* (69.48, 72.64)	51.37* (49.46, 53.28)	64.19* (62.34, 66.04)	61.31* (59.25, 63.37)	74.66* (72.82, 76.50)	50.06* (47.69, 52.43)	65.71* (63.10, 68.32)	55.78* (53.77,57.79)	68.90* (67.02,70.78)
CE-Net	63.21* (61.66, 64.76)	75.22* (73.81, 76.63)	54.75* (52.71, 56.79)	66.08* (64.19, 67.97)	64.26* (61.62, 66.90)	75.23* (72.27, 78.19)	58.44* (56.20, 60.68)	68.49* (66.18, 70.80)	60.16* (58.22,62.10)	71.50* (69.77,73.23)
AG-Net	66.99* (64.99, 68.99)	77.49* (76.10, 78.88)	60.91* (58.61, 63.21)	71.59* (69.75, 73.43)	67.49* (65.65, 69.33)	78.39* (76.83, 79.95)	59.62* (57.58, 61.66)	70.10* (68.10, 72.10)	63.70* (61.98,65.52)	74.4* (73.18,75.62)
RASP-Net	69.79 (68.11, 71.47)	80.43* (79.03, 81.83)	62.11* (60.48, 63.74)	75.72 (74.27, 77.17)	69.48* (68.11, 70.85)	80.93★ (79.46, 82.40)	61.25★ (59.43, 63.07)	73.36 (71.43, 75.29)	65.66* (64.22,67.10)	77.60★ (76.59,78.61)
MSCS-Net	**72.56** (71.38, 73.74)	**83.64** (82.47, 84.81)	**66.12** (65.19, 67.05)	**78.31** (77.05, 79.21)	**74.20** (72.86, 75.54)	**84.16** (82.84, 85.48)	**65.39** (64.48, 66.30)	**76.08** (75.07, 77.09)	**69.57** (68.56,70.57)	**80.54** (79.49,81.58)

biomarker such as IRF, indicating a potential for more precise segmentation of challenging small biomarkers.

Table 4 presents the quantitative results of our MSCS-Net on the RK dataset. Impressively, our model outperforms existing methods across all six biomarker segmentation types. Specifically, our MSCS-Net outperforms the state-of-the-art RASP-Net with an overall Dice improvement of 3.32 points (from 74.54 to 77.86) and a Jaccard improvement of 2.61 points (from 64.62 to 67.23). For the small biomarker HF, we achieve a Jaccard improvement of 3.42 points (from 54.74 to 58.16) surpassing the performance of the leading RASP-Net. For the biomarker IRF, compared to RASP-Net, we achieved a Jaccard improvement of 3.31 points (from 60.20 to 63.51). Additionally, our method significantly improves the Jac accuracy of PED from 66.42 % to 70.10 % compared to AG-Net. Furthermore, for the typically smaller DRUSEN biomarker, we achieve a Jaccard improvement of 3.14 points (from 61.22 to 64.36) compared to RASP-Net. These results demonstrate the effectiveness of our approach in accurate biomarker segmentation.

Our comprehensive analysis of qualitative and quantitative results demonstrates the superior performance of our method in biomarker segmentation on the RK dataset. Despite the significant variability in mixed datasets arising from diverse acquisition devices and time disparities, our approach exhibits robustness and consistently outperforms other methods. Notably, our method excels in precise localization and segmentation accuracy, surpassing existing techniques in handling the challenges posed by dataset heterogeneity.

4.6. Comparison with other methods on the AMD-SD dataset

Our qualitative results and quantitative results on the AMD-SD dataset are shown in Fig. 12 and Table 5. As shown in Fig. 12, the segmentation results appear unsatisfactory for U-Net. Due to insufficient attention to more details, there are still problems of mis-segmentation and under-segmentation, for example, in the second row, part of the SRF was mistakenly identified as IRF. In the fourth row, U-Net also does not completely segment the SRF. There are also issues of over-segmentation for the SHRM. Compared to U-Net, CE-Net shows some improvement in the situation, but problems of mis-segmentation and under-segmentation still exist. AG-Net and RASP-Net pay more attention

Table 6
Segmentation performance (Jac and Dice) of using the RB dataset as the training set and the Kermany dataset as the test set. (unit: %).

Networks	PED		IRF		CNV		HF		DRUSEN		Overall	
	Jac	Dice	Jac	Dice	Jac	Dice	Jac	Dice	Jac	Dice	Jac	Dice
U-Net++	51.25	62.04	38.10	45.39	56.83	64.27	33.05	42.72	43.22	54.61	46.57	49.09
AG-Net	56.71	68.36	48.46	52.60	65.07	69.22	42.76	52.48	47.15	58.59	52.95	61.23
RASP-Net	59.34	70.87	52.12	62.56	63.91	72.45	48.67	59.90	52.30	63.74	56.02	66.79
MSCS-Net	64.23	74.14.	57.61	67.48	67.93	76.11	53.37	63.31	57.03	69.47	60.51	70.86

Table 7
Comparison of different models based on inference time, FLOPs, and GPU memory usage.

Networks	U-Net	Deeplabv3+	CE-Net	U-Net++	SegFormer	Swin U-Net	AG-Net	RASP-Net	MSCS-Net
FLOPs(G)	22.68	40.58	76.37	42.57	44.33	60.12	31.64	39.87	27.19
Infer-time (s)	0.021	0.032	0.071	0.039	0.024	0.059	0.041	0.048	0.025
Memory (G)	9.74	8.41	16.71	20.54	13.78	18.34	17.31	15.42	13.67

to details than the previous two networks, improving the issue of mis-segmentation, but they still cannot accurately identify small portions of biomarkers. Our network has more accurate localization and achieves better segmentation results of biomarkers.

Table 5 presents the quantitative results of our MSCS-Net on the AMD-SD dataset. As can be seen from the last row of the table, our model outperforms the others in the task of segmenting the five types of biomarkers in this dataset. Specifically, our MSCS-Net outperforms the state-of-the-art RASP-Net with an overall Dice improvement of 2.94 points (from 77.60 to 80.54) and a Jaccard improvement of 3.91 points (from 65.66 to 69.57). For the biomarker IRF, compared to RASP-Net, we achieve a Dice improvement of 3.23 points (from 80.93 to 84.16) and a Jaccard improvement of 4.72 points (from 69.48 to 74.20). Additionally, our method significantly improves the Dice accuracy of PED from 71.59 % to 78.31 % compared to AG-Net. For the SHRM segmentation task, which is unique to this dataset, Dice is generally low. We achieve a Dice improvement of 2.72 points (from 73.36 to 76.08) and a Jaccard improvement of 4.14 points (from 61.25 to 65.39) compared to RASP-Net. Meanwhile, paired *t*-test also shows that our method outperforms other methods in this segmentation task. These results demonstrate the effectiveness of our approach in accurate biomarker segmentation.

Our comprehensive analysis of qualitative and quantitative results demonstrates the superior performance of our method in biomarker segmentation on the AMD-SD dataset.

4.7. Additional experimental supplements

4.7.1. Analysis of different devices

We analysis the model's potential performance under different OCT imaging devices and acquisition protocols. We used RB dataset as the source domain, which is collected using a Topcon device, and the Kermany dataset as the target domain, collected using a Heidelberg device. We use the RB dataset as the training set. The learning rate is set to 1e-4, momentum to 0.9, weight decay to 1e-5, and the batch size is set to 8. We train the model for a total of 250 epochs. Then we test with Kermany dataset, and the results are shown in Table 6. Quantitative results show that, compared to training and testing solely on the target domain, there is a significant performance gap when training on the source domain and testing on the target domain. This gap is primarily due to the domain shift caused by differences in imaging devices. However, despite the overall performance decline, our method still achieved the best results under these challenging conditions.

4.7.2. Efficiency analysis

We analyze the efficiency of major models compared in Table2, using three comprehensive metrics: inference time (seconds per image), floating-point operations per second (FLOPs), and GPU memory usage (Memory). We maintained consistency by employing the same critical hyperparameters (e.g., batch size set to 8, input size to 512 × 512 to ensure a fair comparison. As shown in Table 7, our model outperforms most models on three metrics while having better segmentation performance. With the ability to train economically effectively, future work will focus on further reducing training costs to enhance overall efficiency.

5. Conclusion

In this paper, we propose an MSCS-Net (Multi-scale CNN-Swin Network), a hybrid CNN-Transformer multi-scale framework for improving the quality of different kinds of biomarker segmentation in OCT images. Our MSCS-Net combines CNN-Transformer efficiently by combining the local features of CNN with the global features of Transformer to better focus on biomarkers of different sizes. We also incorporate an edge detection path and propose a Boundary Guided Module to better differentiate the biomarker regions and alleviate the challenges associated with difficult segmentation due to low image contrast and blurred edges. Meanwhile, we utilize the proposed Targeted shifted window-based MSA and Feature Dimensionality Reduction Module to adapt to the distributional features of OCT images and extensively collect the information of small biomarker land to accurately perform small biomarker land segmentation. Finally, a cross-attention based feature fusion module is proposed to fuse the first and last layers of features in a finer way to establish dependencies. Experiments on local and public datasets show that our proposed method improves the segmentation performance for multi-class biomarker. However, our work has some limitations, starting with the difficulty of data labeling, which is often time-consuming and labor-intensive for medical images that require specialized knowledge and experience. Fully supervised learning requires a large amount of well-labeled data, which can make data labeling difficult and costly. Second, there is a lack of generalization for other diseases of the retina. In our future work, we aim to employ semi-supervised or weakly-supervised methods for biomarker segmentation to address the high cost associated with pixel-level label annotations and alleviate the burden on physicians. And based on this, we will try to extend the network to further segment other retinal biomarkers.

CRediT authorship contribution statement

Zhanpeng Fan: Writing – original draft. **Xiaoming Liu:** Writing – review & editing. **Ying Zhang:** Resources. **Jia Zhang:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62176190.

References

[1] Tan TF, et al. *Artificial intelligence and digital health in global eye health: opportunities and challenges*. Lancet Glob Health 2023;11(9):e1432–43.
[2] Lee AY, et al. *UK AMD EMR USERS GROUP REPORT V: benefits of initiating ranibizumab therapy for neovascular AMD in eyes with vision better than 6/12*. Br J Ophthalmol 2015;99(8):1045–50.
[3] Trichonas G, Kaiser PK. *Optical coherence tomography imaging of macular oedema*. Br J Ophthalmol 2014;98(Suppl 2):ii24–9.
[4] Franklin, S.W. and S.E. Rajan, *Computerized screening of diabetic retinopathy employing blood vessel segmentation in retinal images.* biocybernetics and biomedical engineering, 2013**4**(2): p. 117-124.
[5] Chen Z, et al. *Automated segmentation of fluid regions in optical coherence tomography B-scan images of age-related macular degeneration*. Opt Laser Technol 2020;122:105830.
[6] Liu X, et al. *Confidence-guided topology-preserving layer segmentation for optical coherence tomography images with focus-column module*. IEEE Trans Instrum Meas 2020;70:1–12.
[7] Lim LS, et al. *Age-related macular degeneration*. Lancet 2012;379(9827):1728–38.
[8] Khan KN, et al. *Differentiating drusen: Drusen and drusen-like appearances associated with ageing, age-related macular degeneration, inherited eye disease and other pathological processes*. Prog Retin Eye Res 2016;53:70–106.
[9] Phadikar P, et al. *The potential of spectral domain optical coherence tomography imaging based retinal biomarkers*. International journal of retina and vitreous 2017;3:1–10.
[10] Bandello F, et al. *Diabetic macular edema*. Macular Edema 2017;58:102–38.
[11] Hou KK, et al. *Pseudoflow with OCT angiography in eyes with hard exudates and macular drusen*. Transl Vis Sci Technol 2019;8(3):50.
[12] Ronneberger O, Fischer P, Brox T. *U-net: convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer; 2015.
[13] Jiang X, Mojon D. *Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images*. IEEE Trans Pattern Anal Mach Intell 2003;25(1):131–7.
[14] Mendonca AM, Campilho A. *Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction*. IEEE Trans Med Imaging 2006;25(9):1200–13.
[15] Roychowdhury S, Koozekanani DD, Parhi KK. *Iterative vessel segmentation of fundus images*. IEEE Trans Biomed Eng 2015;62(7):1738–49.
[16] Liu X, et al. *Multi-scale local-global transformer with contrastive learning for biomarkers segmentation in retinal OCT images*. Biocybernetics and Biomedical Engineering 2024;44(1):231–46.
[17] Hassan B, et al. *Joint segmentation and quantification of chorioretinal biomarkers in optical coherence tomography scans: a deep learning approach*. IEEE Trans Instrum Meas 2021;70:1–17.
[18] Spaide T, et al. *Geographic atrophy segmentation using multimodal deep learning*. Transl Vis Sci Technol 2023;12(7):10.
[19] Pham QT, et al. *Automatic drusen segmentation for age-related macular degeneration in fundus images using deep learning*. Electronics 2020;9(10):1617.
[20] Xi X, et al. *IA-net: informative attention convolutional neural network for choroidal neovascularization segmentation in OCT images*. Biomed Opt Express 2020;11(11):6122–36.
[21] Long J, Shelhamer E, Darrell T. *Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
[22] Xing G, et al. *Multi-scale pathological fluid segmentation in oct with a novel curvature loss in convolutional neural network*. IEEE Trans Med Imaging 2022;41(6):1547–59.
[23] Rasti R, et al. *RetiFluidNet: a self-adaptive and multi-attention deep convolutional network for retinal OCT fluid segmentation*. IEEE Trans Med Imaging 2022.
[24] Liu X, et al. *TSSK-Net: weakly supervised biomarker localization and segmentation with image-level annotation in retinal OCT images*. Comput Biol Med 2023;153:106467.
[25] Liu Z, et al. *Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[26] Badrinarayanan V, Kendall A, Cipolla R. *Segnet: a deep convolutional encoder-decoder architecture for image segmentation*. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95.
[27] Chen L-C, et al. *Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV)*. 2018.
[28] Zhou Z, et al. *Unet++: redesigning skip connections to exploit multiscale features in image segmentation*. IEEE Trans Med Imaging 2019;39(6):1856–67.
[29] Lachinov D, et al. *Projective skip-connections for segmentation along a subset of dimensions in retinal OCT. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer; 2021.
[30] Meng Q, et al. *MF-Net: multi-scale information fusion network for CNV segmentation in retinal OCT images*. Front Neurosci 2021;15:743769.
[31] Gende M, et al. *End-to-end multi-task learning approaches for the joint epiretinal membrane segmentation and screening in OCT images*. Comput Med Imaging Graph 2022;98:102068.
[32] Liu W, Sun Y, Ji Q. *MDAN-UNet: multi-scale and dual attention enhanced nested U-Net architecture for segmentation of optical coherence tomography images*. Algorithms 2020;13(3):60.
[33] Hatamizadeh A, Terzopoulos D, Myronenko A. *End-to-end boundary aware networks for medical image segmentation. Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer; 2019.
[34] Wang R, et al. *Boundary-aware context neural network for medical image segmentation*. Med Image Anal 2022;78:102395.
[35] Lee HJ, et al. *Structure boundary preserving segmentation for medical image with ambiguous boundary. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[36] Lin X, et al. *Batformer: towards boundary-aware lightweight transformer for efficient medical image segmentation*. IEEE J Biomed Health Inform 2023.
[37] Vaswani A, et al. *Attention is all you need*. Adv Neural Inf Proces Syst 2017;30.
[38] Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.
[39] Yuan L, et al. *Tokens-to-token vit: Training vision transformers from scratch on imagenet. Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[40] Zheng S, et al. *Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[41] Chen, J., et al., *Transunet: Transformers make strong encoders for medical image segmentation.* arXiv preprint arXiv:2102.04306, 2021.
[42] Zhang Y, Liu H, Hu Q. *Transfuse: fusing transformers and cnns for medical image segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer; 2021.
[43] He K, et al. *Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[44] Wang H, et al. *Mixed transformer u-net for medical image segmentation. ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE; 2022.
[45] Mekjavić PJ, et al. *The burden of macular diseases in central and eastern Europe—implications for healthcare systems*. Value in health regional issues 2019;19:1–6.
[46] Wenxuan W, et al. *Transbts: Multimodal brain tumor segmentation using transformer. International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2021.
[47] Stergiou A, Poppe R, Kalliatakis G. *Refining activation downsampling with SoftPool. Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[48] Chen T, et al. *RBGNet: reliable boundary-guided segmentation of choroidal neovascularization. International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2023.
[49] Lin A, et al. *Ds-transunet: dual swin transformer u-net for medical image segmentation*. IEEE Trans Instrum Meas 2022;71:1–15.
[50] Chen C-F-R, Fan Q, Panda R. *Crossvit: cross-attention multi-scale vision transformer for image classification. Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[51] Jin J, et al. *Edge detection guide network for semantic segmentation of remote-sensing images*. IEEE Geosci Remote Sens Lett 2023;20:1–5.
[52] Liu Y, et al. *Richer convolutional features for edge detection. Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[53] Bogunović H, et al. *RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge*. IEEE Trans Med Imaging 2019;38(8):1858–74.
[54] Kermany DS, et al. *Identifying medical diagnoses and treatable diseases by image-based deep learning*. Cell 2018;172(5):1122–31.
[55] Hu Y, et al. *AMD-SD: an optical coherence tomography image dataset for wet AMD lesions segmentation*. Sci Data 2024;11(1):1014.
[56] Gu Z, et al. *Ce-net: context encoder network for 2d medical image segmentation*. IEEE Trans Med Imaging 2019;38(10):2281–92.
[57] Xie E, et al. *SegFormer: simple and efficient design for semantic segmentation with transformers*. Adv Neural Inf Proces Syst 2021;34:12077–90.
[58] Cao H, et al. *Swin-unet: Unet-like pure transformer for medical image segmentation. European conference on computer vision*. Springer; 2022.
[59] Schlemper J, et al. *Attention gated networks: learning to leverage salient regions in medical images*. Med Image Anal 2019;53:197–207.