The Journal of Supercomputing (2025) 81:665
https://doi.org/10.1007/s11227-025-07136-7

WS-SAM: self-prompting SAM with wavelet and spatial domain for OCTA retinal vessel segmentation

Zhaohui Zhang[1] · Fei Ma[1] · Hongjuan Liu[1] · Xiwei Dong[2] · Yanfei Guo[1] · Jing Meng[1]

Accepted: 26 February 2025

Abstract

Optical coherence tomography angiography (OCTA) technology can accurately depict the microvascular architecture of the retina, providing crucial evidence for the diagnosis of retinal disease. Nevertheless, OCTA images are typically accompanied by artifacts and low signal-to-noise ratios, exerting a severe impact on diagnostic efficiency. Deep learning-based segmentation algorithms are widely recognized for improving retinopathy diagnosis accuracy. This research investigates the application of the segment anything model (SAM) in OCTA retinal vessel segmentation and proposes a self-prompting SAM based on wavelets and spatial domains, named WS-SAM. Specifically, we innovatively designed three key modules: (1) A dual-domain encoder aims to extract multi-scale features through a joint encoding method in the spatial domain and frequency domain to effectively suppress noise interference. (2) A wavelet space fusion module aims to effectively suppress artifact interference and enhance detail texture features by adaptively fusing multi-scale frequency-domain features with spatial-domain features. (3) A Meta Self-Prompter is designed to automatically generate prompt information based on prototype learning algorithms and guide the model to focus on vascular structures via the prompt mechanism, thereby preventing segmentation fractures. Furthermore, we have established a novel dataset, namely OCTA-RV, to augment the data in the field of OCTA retinal vessel segmentation. The experimental results indicate that the Dice coefficients of the WS-SAM model on the datasets of OCTA500-3 M, OCTA500-6 M and OCTA-RV are 0.8754, 0.8949 and 0.7412, respectively, manifesting a remarkable competitiveness compared with contrastive models.

Keywords Optical coherence tomography angiography · Vessel segmentation · Wide-field OCTA

Extended author information available on the last page of the article

Published online: 28 March 2025

1 Introduction

Optical coherence tomography angiography (OCTA) is a noninvasive medical imaging technique to observe vascular structure and blood flow in human tissues [1]. OCTA can accurately display the distribution and morphology of the vascular structure, including capillaries, arteries, veins, etc. OCTA can be used to study the development of vascular structures, abnormal changes and vascular mutations associated with disease. It has effective application in the early diagnosis and monitoring of ocular vascular diseases, such as diabetic retinopathy (DR), age-related macular degeneration (AMD) and venous occlusion [2, 3]. The advancement of OCTA technology has significantly propelled clinical ophthalmic research [4, 5]. On the other hand, OCTA allows noninvasive 3D analysis of the retinal and choroidal vascular system and can be segmented to view each vascular plexus individually. OCTA can visualize the vascular structure at different depths [6]. OCTA allows for high-definition visualization of choroidal neovascularization (CNV) structures without the contrast injection and is also useful for central retinal vein occlusion (CRVO), branch retinal vein occlusion (BRVO) and other retinal vasculopathies [7]. Therefore, it is of great importance to study the vascular segmentation of OCTA images.

Compared with traditional time-domain optical coherence tomography (TD-OCT), SS-OCT (swept source optical coherence tomography) has faster scanning speed and deeper tissue fluoroscopy capability. Wide-field OCTA (WF-OCTA) can be realized by SS-OCT, in which the single scan area can reach 15×9 mm^2 or 12×12 mm^2. Figure 1 shows that OCTA-RV contains OCTA image data within

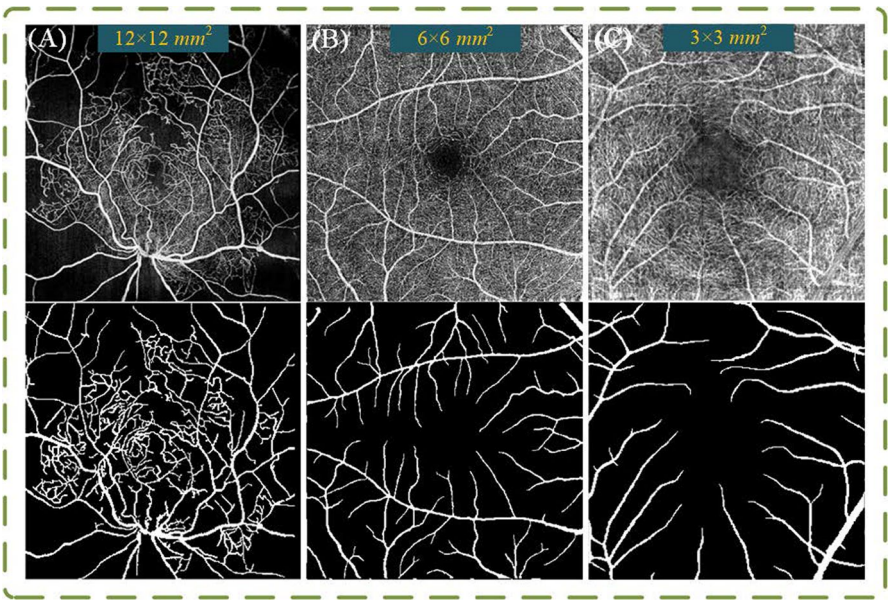

Fig. 1 Samples from different OCTA datasets. From left to right, the original images and ground truth of **A** OCTA-RV, **B** OCTA500-6 M and **C** OCTA500-3 M

a $12 \times 12\text{mm}^2$ fovea-centered field of view (FOV). WF-OCTA can provide a wide range of field of view, thus enabling the capture of a wider range of vascular structures in a single image. This is essential for a comprehensive understanding of the layout and variation of the ocular vascular.

However, there exist many difficulties in the segmentation task of OCTA retinal images. For example, there are usually many artifacts and noises in OCTA retinal images. The blood vessels by OCTA may be irregular, twisted, bifurcated and even broken. Currently, there is a limited availability of OCTA retina datasets, etc. The above difficulties are shown in Fig. 1, which shows the representative OCTA images of the fundus. To overcome these problems, current researches mainly focus on designing deep learning frameworks and modules [8–12], and have achieved better results for retinal vessel segmentation in OCTA images.

Currently, many methods based on encoder–decoder structures have been applied to medical image segmentation tasks [13–15], significantly improving segmentation performance. To better capture long-range contextual information and enhance segmentation accuracy, many researchers have focused on incorporating attention blocks into the models [8–10]. The Transformer model has been shown to be more accurate and robust for segmentation tasks [16–18]. Existing research indicates that integrating frequency-domain and spatial-domain features can provide a more comprehensive way to extract information at multiple scales and frequency ranges. Furthermore, each frequency component in the frequency domain corresponds to a specific aspect of the image. Low-frequency components capture the overall structure, while high-frequency components reveal local details. [19–21] enhance the model's ability to capture low-frequency global structures and high-frequency details by introducing frequency-domain information. Moreover, some unsupervised approaches, such as homomorphic filtering [22], have also exhibited outstanding performance in retinal vessel segmentation and are capable of precisely extracting the contours of retinal vessels. The denoising of OCTA images is of significant importance for the subsequent image analysis and processing tasks [23, 24].

The high-quality OCTA datasets have consistently remained a pressing demand for researchers in the field. The construction of the new OCTA-RV dataset aims to meet these demands and promote the development of retinal vessel segmentation and related research. As shown in Table 1, compared with the existing OCTA datasets, the images in the OCTA-RV dataset have a larger field of view ($12 \times 12\ \text{mm}^2$), which can cover more peripheral retinal areas and capture a wider range of vascular structures. The 2D images in the OCTA-RV dataset are generated from maximum intensity projections of corresponding anatomical layers within the original 3D volumetric data, specifically including the superficial vascular layer (SVL), deep vascular layer (DVL), outer retinal layer (ORL) and pigment epithelium detachment (PED). This multilayered projection strategy enables the images on OCTA-RV dataset to capture depth-resolved vascular architectures, thereby revealing structural details of deep retinal microvasculature that are critical for clinical analysis. Additionally, OCTA-RV places particular emphasis on pixel-level annotation, labeling a substantial number of deep microvessels within the included images in Fig. 2.

Table 1 Comparison of OCTA-RV dataset with public OCTA datasets

Dataset	Image	Field of view	Resolution	Projection maps	Pixel-level label	Year
OCTA500 3 M	200	3×3 mm^2	304×304	ILM&OPL&BM	N	2020
OCTA500 6 M	300	6×6 mm^2	400×400	ILM&OPL&BM	N	2020
Rose-1	117	3×3 mm^2	304×304	SVC&DVC	Y	2021
Rose-2	112	3×3 mm^2	512×512	SVC&DVC	Y	2021
OCTA-RV (our)	372	12×12 mm^2	304×304	SVL&DVL &ORL&PED	Y	2024

Note: "ILM" denotes the internal limiting membrane layer, "OPL" denotes the outer plexiform layer, "BM" denotes the bruch's membrane, "SVC" denotes the superficial vascular complexes, "DVC" denotes the deep vascular complexes, "SVL" denotes the superficial vascular layer, "DVL" denotes the deep vascular layer, "ORL" denotes the outer retinal layer, and "PED" denotes the pigment epithelial detachment layer. The number of OCTA-RV images after data enhancement is 372

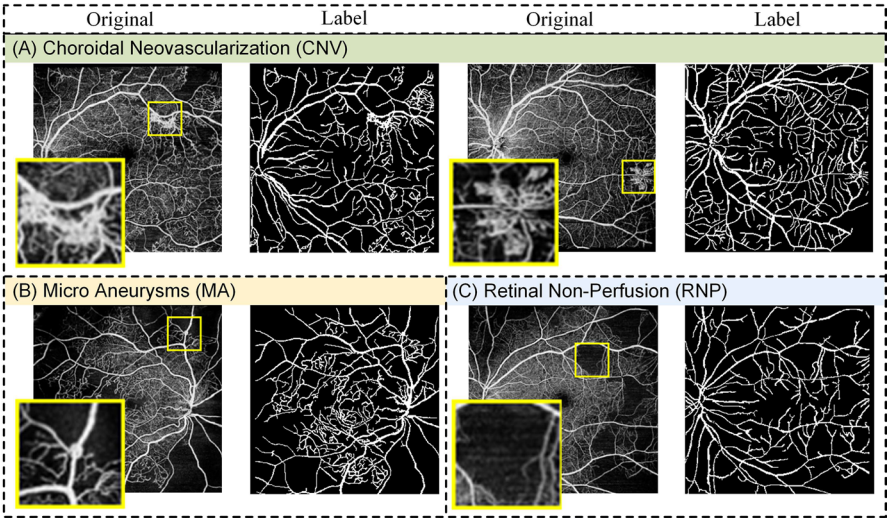

Fig. 2 Example of three different OCTA images with retinal diseases in the OCTA-RV dataset

Recently, segment anything model (SAM) [25] has demonstrated excellent generalization performance and zero-shot capabilities for image segmentation tasks. SAM is trained on over 1 billion masks on 11 million images. SAM is a based prompt segmentation model that supports four kinds of segmentation prompts, which are sparse prompt (point, box and text) and dense prompt (mask). High-quality manual annotation as segmentation prompt is the key to determine the segmentation performance of SAM. However, professional medical prompts are expensive and time-consuming, which leads to the decline of SAM segmentation effect on medical images.

To address the above issues, we propose a new self-prompting SAM, named WS-SAM. Based on the architecture of the segment anything model, WS-SAM

investigates the feasibility of implementing self-prompting. Specifically, we have devised a dual-domain encoder that integrates spatial-domain and frequency-domain modeling. In the frequency domain, the dual-domain encoder employs multi-level wavelet decomposition to segregate high-frequency noise from low-frequency structural information. Thus, it achieves the goal of effectively suppressing noise and preserving the integral structural information of the image. In the spatial domain, the dual-domain encoder performs pixel-wise spatial encoding, which can extract more local details to ensure the accuracy of segmentation. On this basis, we proposed the wavelet space fusion module to effectively fuse frequency and spatial-domain features, further improving the model's performance in terms of global structure and local details. Furthermore, we propose a Meta Self-Prompter, which automatically acquires prompt information via a prototype learning algorithm and utilizes the prompt information to guide the model to focus on the structural morphology of retinal vessels, ensuring the continuity of blood vessel segmentation. Our contributions can be summarized as follows:

- We propose a novel self-prompting SAM based on wavelets and spatial domains for OCTA vessel segmentation, named WS-SAM, which can efficiently accomplish automatic vessel segmentation.
- We propose a dual-domain encoder that effectively suppresses noise by introducing multi-scale frequency-domain information. In particular, the pixel-level image encoder adopts a dual-path pixel-channel window attention mechanism to achieve fine-grained perception of feature information. We also proposed a wavelet space fusion module to achieve adaptive fusion of multi-scale frequency-domain and spatial-domain features.
- We design the Meta Self-Prompter, which automatically generates prompt information through a prototype learning algorithm to guide the model to focus on retinal vessel structures.
- We have constructed a new OCTA retinal vessel segmentation dataset, named OCTA-RV. As illustrated in Fig. 2, OCTA-RV includes various OCTA images with retinal diseases, such as choroidal neovascularization (CNV), microaneurysm (MA) and retinal non-perfusion (RNP).

2 Related work

In recent years, many deep learning-based methods have been proposed for medical image segmentation. Ronneberger et al. [13] designed a U-shaped architecture, named UNet. Oktay et al. [14] proposed a new medical image attention gate model and combined with UNet. For frequency-domain methods, Li et al. [19] proposed GFUNet, a global frequency-domain UNet, achieving efficient and effective medical image segmentation by integrating the Fourier transform with the UNet structure. Bui et al. [20] proposed MEGANet, a multi-scale edge-guided attention network that effectively retains high-frequency information by integrating classic edge detection techniques with attention mechanisms. Wu et al. [21] proposed MFMSNet,

a multi-frequency and multi-scale interactive network integrating CNN and Transformer, for the segmentation of breast ultrasound images.

2.1 OCTA image segmentation

The convolutional neural network (CNN) can effectively extract features for the OCTA vascular segmentation task. Wang et al. [26] proposed DBUNet, which consists of a pure convolutional branch for extracting minutiae features and a UNet branch. Ma et al. [27] proposed a coarse-to-fine vessel segmentation network named OCTA-Net and a new OCTA retinal dataset (ROSE). Hao et al. [28] designed a novel Voting-based Adaptive Feature Fusion multitask network (VAFF-Net) designed for the joint segmentation, detection and classification of retinal vessels (RV), foveal avascular zone (FAZ) and retinal vascular junctions (RVJ) in OCTA images. Ning et al. [29] designed FRNet, a full-resolution convolutional network composed of improved Recurrent ConvNeXt Blocks, and introduced a new OCTA dataset. Abtahi et al. [30] proposed a multimodal fusion deep learning neural network named MF-AV-Net, and designed early fusion and late fusion architectures for OCTA retinal vessel segmentation, respectively. Liu et al. [31] designed a network architecture called ACRROSS to achieve retinal blood vessel and capillary segmentation on limited OCTA images by learning local contrast and vessel structure. Xu et al. [32] introduced a cascade neural network named AV-casNet that integrates convolutional neural networks (CNN) and graph neural networks (GNN) to enhance the connectivity of retinal vessel segmentation in OCTA images.

The Transformer can learn the dependencies between different positions in the input sequence through the self-attention mechanism, which can effectively capture the global features of OCTA images. Shi et al. [16] designed a network architecture integrated with Transformer and UNet, named TCU-Net. The TCU-Net employs Transformer to compensate for the lack of global dependence of pure convolution operation on OCTA images for retinal vessel segmentation. Tan et al. [17] designed an end-to-end transformer network OCT2Former based on an encoder–decoder structure, which is composed of a dynamic transformer encoder and a lightweight decoder. The SAM framework has been introduced for OCTA vessel segmentation tasks. Chen et al. [18] designed a new SAM-based OCTA vessel segmentation model, SAM-OCTA, for local vessel and arteriovenous vessel segmentation.

To overcome the problem of limited OCTA dataset, many works have proposed corresponding solution schemes. Li et al. [11] collected a new OCTA dataset, named OCTA500, which is currently the largest and most comprehensive OCTA dataset, and proposed a 3D-to-2D segmentation network, named IPN-V2. Chinkamol et al. [12] proposed OCTAve, a weakly supervised learning framework for microvascular segmentation in OCTA, achieving excellent performance with minimal labeled training data. Wu et al. [33] designed a novel 3D-to-2D segmentation network, named PAENet. PAENet consists of a 3D feature learning path and a 2D segmentation path, which guides the 2D segmentation with 3D features. Yang et al. [34] designed a layer attention network (LA-Net) for 3D-to-2D retinal vessel segmentation. This network consists of a 3D projection path and a 2D segmentation path. The

multi-scale layer attention module in the 3D path can learn the layer features of OCT and OCTA images and capture 3D multi-scale information. The reverse boundary attention module in the 2D path maintains the boundary and shape features of retinal vessels by focusing on non-salient regions. Yang et al. [35] proposed the ODDF-Net segmentation network for the simultaneous 2D segmentation of RC, RA, RV and FAZ in 3D OCTA. ODDF-Net introduced the concept of optical density to generate additional input images to enhance specificity. Moreover, ODDF-Net designs auxiliary classification heads and cross-dimensional feature fusion modules to model the relationship between diseases and retinal structures.

2.2 Medical variants of SAM

Recently, the segment anything model (SAM) family has demonstrated superior segmentation performance and zero-shot capabilities on medical images. Cheng et al. [36] designed a prompt-free adaptive SAM for medical image segmentation named H-SAM. H-SAM has a two-stage layered decoder, in which the latter stage refines the predictions of the previous stage. Ma et al. [37] proposed MedSAM, a foundation model designed for universal medical image segmentation. Wu et al. [38] designed a new paradigm toward the universal medical image segmentation, termed "One-Prompt Segmentation". OnePrompt Segmentation combines the strengths of one-shot and interactive methods. Gao et al. [39] proposed decoupled SAM (DeSAM), which modified the mask decoder of SAM by introducing the prompt-relevant IoU module (PRIM) and the prompt-decoupled mask module (PDMM), achieving a decoupled design. Wu et al. [40] proposed the Medical SAM Adapter (Med-SA), which integrates medical knowledge into the segmentation model by using lightweight and efficient adaptive techniques and introduces spatial depth transpose (SD-Trans) to adapt 2D SAM to 3D medical images.

The vanilla SAM model [25] is less effective in medical images, which is due to the lack of large-scale medical image data and professional domain knowledge prompt. However, professional medical domain prompts require the participation of medical domain experts, which is expensive and time-consuming, and not friendly to clinical scenarios. To address the above issues, [25, 41, 36, 42, 43–46] proposed free -prompt or self-prompting methods, which tend to generate coarse dense prompt information. However, dense -prompt may cause excessive constraints in OCTA images, thus causing detailed features to be ignored. To address these issues, we propose a Meta Self-Prompter, named MSPrompter, which can transfer prompts while enhancing prototype features. Compared with other self-prompting methods, the advantage of MSPrompter lies in adoption of meta-learning strategies, especially prototypical learning, which achieves effective transfer of prompts by extracting common prototype features. Specifically, MSPrompter improves the representation of target sample features by selectively extracting common prototype features. It also facilitates the deep fusion of prompt information with common prototype features, indirectly enabling prompt transfer and generating new sparse prompts.

3 Methodology

In this section, we introduce the new method WS-SAM, including pixel-wise image encoder, wavelet space fusion (WSF-Former) module, Meta Self-Prompter (MSPrompter) and Simplified Mask Decoder. Finally, we detail the loss function of WS-SAM.

3.1 Preliminary

Segment Anything Model (SAM) SAM is a prompt-based image segmentation model that accommodates two types of prompts: sparse (points, boxes) and dense (masks). SAM was trained with over 1 billion masks on 11 million images. Due to its flexible prompting feature, SAM can perform zero-shot transfer between different image tasks, enabling efficient segmentation even without specific training data. Furthermore, SAM proposes a prompt-based segmentation task, aiming to return valid masks based on arbitrary segmentation prompts.

SAM comprises three components: image encoder, prompt encoder and mask decoder. The image encoder is a pre-trained Vision Transformer (ViT) that is optimized for handling high-resolution inputs with minimal computational overhead. The prompt encoder conducts embedding encoding for each type of prompt. The mask decoder is a dual-path decoder that effectively maps image embeddings, prompt embeddings and output tokens to segmentation masks.

Discrete Wavelet Transform (DWT) DWT is a multi-scale frequency analysis method that can decompose image information into different frequency components using 1D wavelet basis functions, as illustrated in Fig. 3. Comparison to the Fourier transform, DWT can capture local image details via high-frequency components while preserving overall structural information via low-frequency components. To capture high-frequency details in images while preserving overall structure, we uniquely use discrete wavelet transform, rather than other frequency methods, to extract frequency-domain information. The definition of DWT is presented as in Eq. 1:

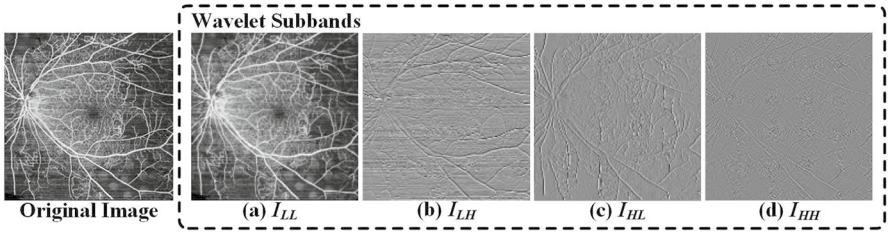

Fig. 3 Visualization of Haar discrete wavelet transform on OCTA-RV image

$$
\begin{aligned}
LL(m,n) &= \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} f(p,q)\cdot h(m-p)\cdot h(n-q),\\
LH(m,n) &= \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} f(p,q)\cdot h(m-p)\cdot g(n-q),\\
HL(m,n) &= \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} f(p,q)\cdot g(m-p)\cdot h(n-q),\\
HH(m,n) &= \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} f(p,q)\cdot g(m-p)\cdot g(n-q).
\end{aligned}
\tag{1}
$$

Here $f(m,n) \in \mathbb{R}^{C\times H\times W}$ is the pixel value of the input image. M and N are the number of rows and columns of the image. h and g are low-pass and high-pass filters, whose coefficients are derived from the corresponding wavelet basis functions. $LL, LH, HL, HH \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$ are the low-low (LL) sub-band, the low–high (LH) sub-band, the high–low (HL) sub-band and the high–high (HH) sub-band, respectively.

3.2 Overall architecture

Figure 4 shows the architecture of our proposed WS-SAM. The architecture of WS-SAM follows the vanilla SAM design and consists of three parts: image encoder, prompt encoder and mask decoder. Different from the vanilla SAM [25], WS-SAM divides the image encoder into two parts: a pixel-wise image encoder and a wavelet

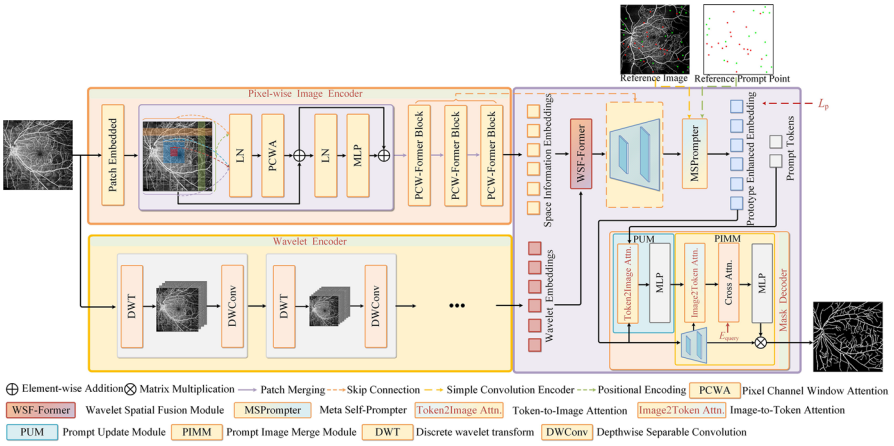

Fig. 4 Overall architecture of the proposed WS-SAM. The proposed WS-SAM takes a query image and a reference image with prompt points as inputs. The features of the queried images are extracted from the spatial domain and the frequency domain by the dual encoder. The sparse prompt of the query image is generated by Meta Self-Prompter based on the enhanced prototype features and the prompt information of the reference image. The Simplified Mask Decoder takes sparse prompt and the prototype enhanced query image as inputs to predict the segmentation result

encoder. The wavelet encoder is composed of six layers of discrete wavelet transform (DWT) with Haar wavelet and depthwise convolution. The inter-domain fusion of wavelet frequency domain and spatial domain is realized by WSF-Former. After inter-domain fusion, image encoding is performed by skip connection and a simple pixel decoder for initial decoding and feature integration. Subsequently, the initial decoded image encoding is embedded as the query image embedding. The query image embedding, reference image embedding and reference prompt embedding are fed into the Meta Self-Prompter (MSPrompter). MSPrompter consists of two critical paths: the prototype enhancement path and the prompt update path, which generate the enhanced prototype embedding and the prompt embedding of the query image, respectively. Finally, the enhanced prototype embedding and the updated prompt embedding are fused and decoded by the Simplified Mask Decoder to predict the segmentation results.

3.3 Pixel-wise image encoder

The current researches demonstrate that many efficient ViT models are unable to form sufficient information mixing through stacking due to depth degradation effects. This leads to the fact that ViT-based models often fail to achieve the expected performance. To solve the above problems and enhance window attention to fine-grained perception, we propose a new pixel-channel window attention module, named PCWA. The PCWA is able to enhance the receptive field and fine-grained perception of window attention while avoiding cross-layer information degradation caused by ViT stacking.

Neighborhood Attention (NA) Neighborhood attention [47] is able to restrict the receptive field of each query token to a fixed-size neighborhood around the corresponding token of the key–value pair. NA can expand local receptive fields without the need for extra operations, maintaining translational equivariance while

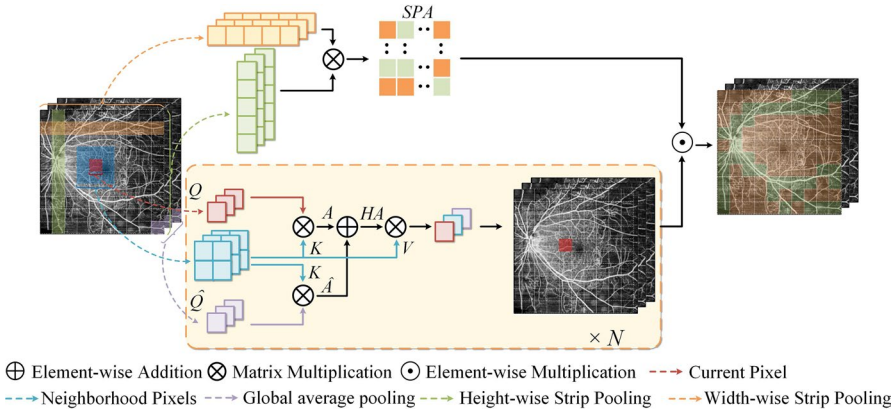

Fig. 5 Overall structure of pixel-channel window attention (PCWA). N is the number of pixels in the query image

introducing local inductive biases. Neighborhood attention on a single pixel can be defined in Eq. 2 as follows:

$$NA(X_{(i,j)}) = \text{softmax}(\frac{Q_{(i,j)}K^T_{(i,j)\sim\rho(i,j)} + B_{(i,j\sim\rho)}}{\sqrt{d}})V_{(i,j)\sim\rho(i,j)}, \tag{2}$$

where $Q_(i,j)$ is $X_{(i,j)} \in \mathbb{R}^{1\times C}$ at the coordinate (i, j) as the query token. $K_{(i,j)\sim\rho(i,j)} \in \mathbb{R}^{k^2\times C}$ is the neighbor token, k^2 is the size of the neighborhood window. $(i,j) \sim \rho(i,j)$ is the neighbor token coordinate within a fixed window size of k×k centered on the query token at (i, j) and ρ is the maximum range. $B_(i,j \sim \rho)$ is the learnable relative positional biases. $\sqrt{d}$ is the scaling parameter.

Pixel-Channel Window Attention To introduce more global information into pixel-level neighborhood attention while enhancing the interaction between neighborhoods, we propose a dual-path pixel-channel window attention in Fig.5. This is a two-path cross-attention mechanism wherein the neighborhood token is employed as the key–value pair, while the center pixel token and the global average pooling token are, respectively, employed as the query pairs. The pixel-channel window attention consists of two core steps: the hybrid neighborhood attention (HNA) and the strip pooling weighting operation. Specifically, the central pixel token and the global average pooling token perform matrix multiplication with the neighborhood tokens, respectively, to obtain the affinity matrices A and $\hat{A}$. This process is illustrated in Eq. 3. The affinity matrices A and $\hat{A}$ are fused to yield the hybrid affinity matrix HA via element-wise addition and convolutional filtering, as defined in Eq. 4. Finally, the hybrid affinity matrix HA performs matrix multiplication with the neighborhood tokens to obtain a new token that aggregates pixel-channel and neighborhood information, as defined in Eq. 5.

$$\begin{aligned} A_{(i,j)\sim\rho(i,j)} &= Q_{(i,j)}K^T_{(i,j)\sim\rho(i,j)}, \\ \hat{A}_{(i,j)\sim\rho(i,j)} &= \hat{Q}K^T_{(i,j)\sim\rho(i,j)}, \end{aligned} \tag{3}$$

$$HA_{(i,j)} = DW(A_{(i,j)\sim\rho(i,j)} + \hat{A}_{(i,j)\sim\rho(i,j)}), \tag{4}$$

$$HNA(X_{(i,j)}) = softmax(\frac{HA_{(i,j)} + B_{(i,j\sim\rho)}}{\sqrt{d}})V_{(i,j)\sim\rho(i,j)}, \tag{5}$$

where $Q_{(i,j)} \in \mathbb{R}^{1\times C}$ is the pixel token located at (i, j). $\hat{Q} \in \mathbb{R}^{1\times C}$ is the global average pooling token. $K_{(i,j)\sim\rho(i,j)} \in \mathbb{R}^{k^2\times C}$ and $V_{(i,j)\sim\rho(i,j)} \in \mathbb{R}^{k^2\times C}$ are neighborhood tokens, and k^2 is the size of the neighborhood window. $A_{(i,j)\sim\rho(i,j)}$ and $\hat{A}_{(i,j)\sim\rho(i,j)} \in \mathbb{R}^{k^2\times 1}$ are the affinity matrices obtained from the central pixel token and the global average pooling token, respectively. DW is depthwise separable convolution. $HA_{(i,j)} \in \mathbb{R}^{k^2\times 1}$ is the hybrid affinity matrix. HNA is hybrid neighborhood attention, and its output is an information aggregation token with a shape of $1 \times C$. Pixel-channel window attention only aggregates the neighborhood receptive field to a single pixel, but such

aggregation capability is limited. Considering that neighborhood external interaction is also necessary for receptive field expansion, we additionally add a neighborhood interaction mechanism consisting of strip pooling attention around pixel-channel window attention, as defined in Eqs. 6 and 7.

$$\mathrm{SPA}(X_{(i,j)}) = \left(\frac{1}{H}\sum_{0\le i<H} X_{(i,j)} \otimes \frac{1}{W}\sum_{0\le j<W} X_{(i,j)}\right), \tag{6}$$

$$\mathrm{PCWA}(X_{(i,j)}) = \mathrm{HNA}(X_{(i,j)}) \odot \mathrm{SPA}(X_{(i,j)}). \tag{7}$$

Here H and W are the height and width of the original image, respectively. *PCWA* is pixel-channel window attention, and its output is a feature map with a shape of $C \times H \times W$. *SPA* is strip pooling attention, and its output is a spatial weight matrix of shape H×W. $\odot$ is the Element-wise Product (Fig. 6).

3.4 Wavelet spatial fusion module

There are inter-domain differences between frequency domain and spatial domain, which makes it impossible to directly fuse frequency-domain features and spatial-domain features. In order to handle the above problems, we propose a novel wavelet spatial fusion module, named WSF-Former in Fig.6. WSF-Former can learn inter-domain interaction information while promoting inter-domain fusion and transformation.

After discrete wavelet transform processing, the spatial-domain and frequency-domain features are decomposed into different low-frequency sub-bands and high-frequency sub-bands. The low-frequency sub-band and the high-frequency sub-band represent distinct feature components of the image, respectively. Therefore, in light of the characteristics of the discrete wavelet transform, we separately fuse different types of wavelet sub-bands. Specifically, we apply single-head and multi-head cross-attention mechanisms to the low-frequency sub-bands and high-frequency sub-bands, respectively, which are transformed from different domains. The process is shown in Eqs. 8 and 9. After the frequency-domain fusion, we concatenate the fused low-frequency and high-frequency features, and convert them to the spatial domain via the inverse wavelet transform to obtain the restored feature X_h. This process is presented in Eq. 10. Subsequently, we fuse the restored feature X_h with the original spatial-domain feature through the cross-attention mechanism. Finally, the spatial voting mechanism is employed to weight the spatial features to obtain the feature X_{Es}, thereby achieving adaptive cross-domain fusion of the frequency-domain and spatial-domain features. This process is presented in Eq. 11.

$$CA(X_s, X_w) = \mathrm{softmax}\left(\frac{D_i(Q_s)D_i(K_w^T)}{\sqrt{d_k}}\right)D_i(V_w), \quad i = 1,2,3 \ \ or \ \ i = 4, \tag{8}$$

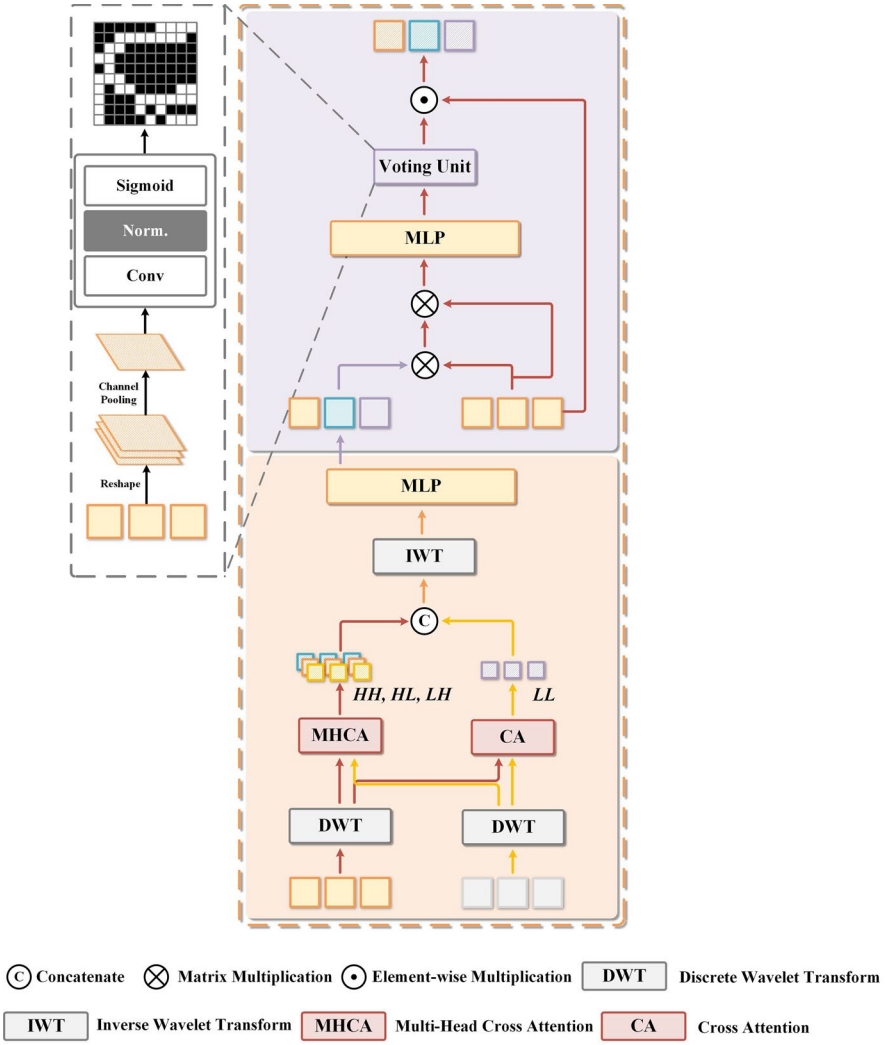

Fig. 6 Overall structure of wavelet space fusion (WSF-Former) module

$$\begin{aligned} \mathrm{MHCA}(X_s, X_w) &= \mathrm{Concat}((\mathrm{head}_1, \mathrm{head}_2, \mathrm{head}_3)W^0), \\ \mathrm{head}_i &= CA(X_s, X_w), \end{aligned} \tag{9}$$

$$X_h = IWT(\mathrm{Concat}(\mathrm{MHCA}(D_i(X_s), D_i(X_w)), CA(D_i(X_s), D_i(X_w))), \tag{10}$$

$$X_{Es} = \mathrm{Vote}(CA(X_s, X_h)) \odot X_s. \tag{11}$$

Here $X_w \in \mathbb{R}^{L \times C}$ and $X_s \in \mathbb{R}^{L \times C}$ are frequency-domain and spatial-domain features. $\{Q, K, V\}_s$ and $\{Q, K, V\}_w$ are the linear projection results of X_w and X_s, respectively.

D is discrete wavelet transform and $D_{\{i=1,2,3,4\}}$ is four different wavelet sub-bands. *IWT* is inverse discrete wavelet transform. *CA* is cross-attention. *MHCA* is multi-head cross-attention. *Vote* is a voting unit consisting of pooling operation and several convolutions. $X_h \in \mathbb{R}^{L \times C}$ is the reconstructed feature through the inverse wavelet transform (IWT). $X_{Es} \in \mathbb{R}^{L \times C}$ is the enhanced spatial-domain feature via dual-domain feature fusion and the voting mechanism. It is notable that since WSFF-Former is located at the lowest layer of the model, its computational consumption is relatively low (Fig. 7).

3.5 Meta Self-Prompter

Medical image prompts are dependent on expert knowledge, which may cause a decline in the performance of the SAM model on medical images, especially when it lacks sufficient medical prompt annotation. Meanwhile, manual prompts lack

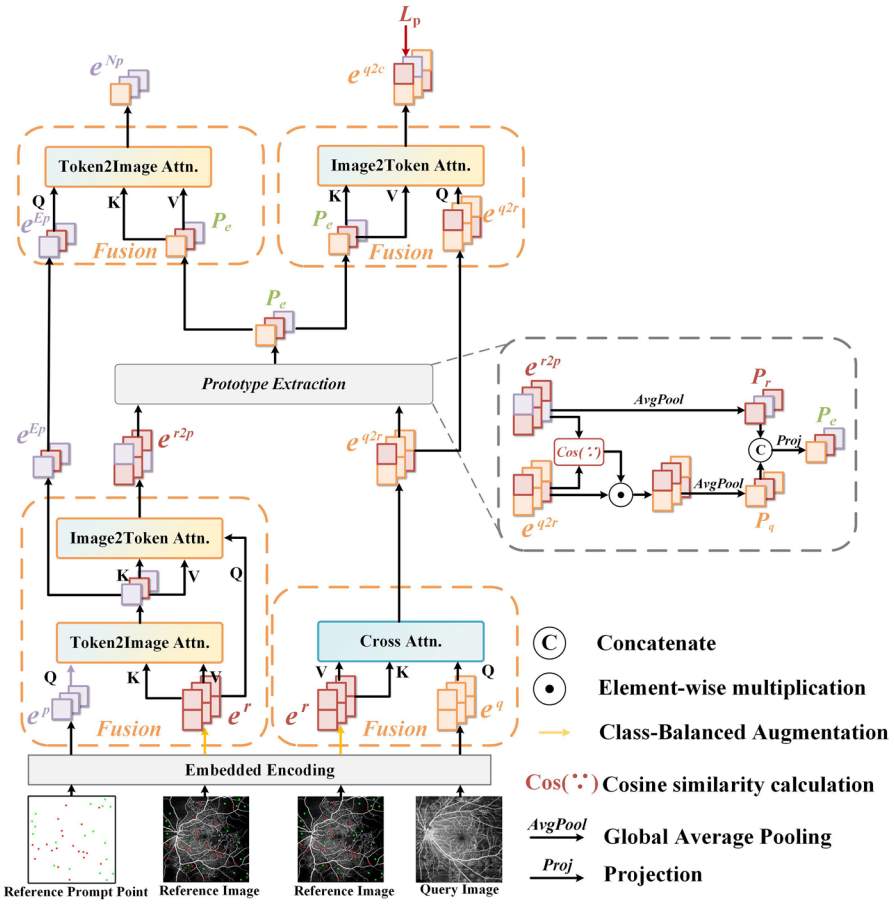

Fig. 7 Overall structure of Meta Self-Prompter (MSPrompter)

clinical feasibility. To reduce the dependency of SAM models on expert knowledge, self-prompting is considered a potentially effective solution. Therefore, we design a new Meta Self-Prompter (MSPrompter) module in Fig.7 inspired by the methods under few-shot scene, especially prototypical network in the meta-learning strategy. Specifically, the Meta Self-Prompter module is mainly divided into three stages: common prototype extraction, query image enhancement and prompt transfer. In the common prototype extraction stage, prototype feature extraction is performed via global pooling. Global pooling can achieve the aggregation and compression of global information in feature maps, representing the feature prototype of the entire image. The common prototype features are extracted by concatenating and projecting the global pooling features of the Query image and Reference image. In addition, prior to the common prototype extraction, we designed a dual-branch fusion mechanism to ensure that the sampled common prototype features are biased toward the common foreground features of the Query image and Reference image. The prompt transfer stage is designed to generate a query image-biased prompt embedding by fusing the prompt embeddings with the common prototype features. In the query image enhancement stage, the query image features are further enhanced through fusion with the common prototype features.

The MSPrompter module's input primarily consists of three components: query image embedding, reference image embedding and reference prompt embedding. The selection of reference images employs a methodology combining principal component analysis (PCA) and the K-means clustering algorithm. First, feature dimensionality reduction is performed on the dataset using principal component analysis (PCA). Subsequently, clustering analysis is conducted on the reduced-dimensional feature space through the K-means algorithm. Finally, select the image with the highest similarity to the features of multiple cluster centers as the reference image. The purpose of choosing the reference image is to manually select the prompt points on the reference image as templates, providing the model with reference prompt information to guide the segmentation task, thereby achieving the automatic prompt function. The reference image embedding is obtained through simple convolution and downsampling. The reference prompt is the positional representation information generated by encoding the coordinates of the prompt points in the reference image through positional encoding techniques, such as coordinate normalization and trigonometric function encoding.

Class Balance Augmentation Due to the possibility of category imbalance or feature distribution bias in the reference image, which may affect the effect of prompt transfer, we performed class balance augmentation to reduce the negative impact. We adopt class weight [36], according to the principle that the variance of Gaussian noise is inversely proportional to the sample frequency of the class, to weight the samples of the unbalanced class, as defined in Eq. 12:

$$P(gt = i) += N(0, \mathrm{var}(i)), \tag{12}$$

where gt is the ground truth. N is the Gaussian noise function. var is a list of variances.

Selective Common prototype extraction To achieve selective extraction of common prototype features, MSPrompter has designed a two-stage feature extraction process: fusion and prototype extraction. In the first stage, the effective fusion between the reference image and the reference prompt is realized by the Token2Image attention and the Image2Token attention, obtaining the features e^{E_p} and e^{r2p}, respectively. This process is illustrated in Eq. 13 and 14. Through cross-attention, the query image and reference image are effectively fused to obtain the feature e^{q2r}, as defined in Eq. 15.

$$e^{E_p} = T2I\ \text{Att}(e^p, e^r) + e^p, \tag{13}$$

$$e^{r2p} = I2T\ \text{Attn}(e^r, e^{E_p}) + e^r, \tag{14}$$

$$e^{q2r} = \text{Cross Attn}(e^q, e^r) + e^q, \tag{15}$$

where T2I Att is a cross-attention that employs a minimal sequence as a query. *I2TAttn* a cross-attention that employs a minimal sequence as a key and value. $e^q, e^r \in \mathbb{R}^{L\times C}$ are the query and reference image sequences, respectively. $e^p \in \mathbb{R}^{K\times C}$ is the reference prompt sequence, and K is the number of prompt points. $e^{E_p} \in \mathbb{R}^{K\times C}$ is the enhanced prompt sequence after fusion. $e^{r2p}, e^{q2r} \in \mathbb{R}^{L\times C}$ are the enhanced query and reference image sequence after fused, respectively.

In the second stage, the enhanced reference image feature e^{r2p} is extracted through average pooling to obtain the prototype feature P_r, as defined in Eq. 16. Subsequently, the common features between e^{r2p} and e^{q2r} are selected in accordance with the cosine similarity metric, as defined in Eq. 17. Then, the prototype of the selected e^{q2r} is extracted by average pooling, obtaining the query image prototype P_q that is most similar to the reference image. This process is illustrated in Eq. 18. Finally, the P_r and P_q prototype features are concatenated and projected to obtain the common prototype P_c, as defined in Eq. 19.

$$P_r = \text{AvgPool}(e^{r2q}), \tag{16}$$

$$\text{Cos}(\because) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}, \tag{17}$$

$$P_q = \text{AvgPool}(Cos(e^{q2r}, e^{r2p}) \odot e^{q2r}), \tag{18}$$

$$P_c = \text{Proj}(\text{Concat}(P_r, P_q)), \tag{19}$$

where Avgpool is the average pooling. Cos(∵) is the cosine similarity selector, and x,y are the image feature vectors to compute the similarity. Concat is the concatenation operation. *Proj* is a 1×1 projected convolution. $\odot$ is Hadamard product.

$P_r, P_q, P_c \in \mathbb{R}^{1\times C}$ are the query prototype, reference prototype and common prototype, respectively.

Query Image Enhancement The common prototype P_c and query image feature e^{q2r} are fused by Image2Token attention to enhance feature representation learning for the target sample and generate a coarse mask e^{q2c}. This process is illustrated in Eq. 20. e^{q2c} is the output of the initial decoding phase of the overall model.

$$e^{q2c} = I2T\text{Attn}(e^{q2r}, P_c) + e^{q2r}, \tag{20}$$

where $e^{q2c} \in \mathbb{R}^{L\times C}$ is the enhanced representation of the common prototype on the query image sequence.

Prompt Transfer Through Token2Image attentions, the deep fusion of the common prototype P_c and the enhancement prompt features e^{E_p} is achieved. This fusion enables the reference prompt to be transferred indirectly to the query image via the common prototype, as defined in Eq. 21.

$$e^{N_p} = T2I\text{Attn}(e^{E_p}, P_c) + e^{E_p}. \tag{21}$$

Here $e^{N_p} \in \mathbb{R}^{K\times C}$ is the new prompt sequence generated after deep fusion with common prototype.

It is notable that although the MSPrompter module is structurally complex, it still retains a lightweight computational expense. This is mainly due to the fact that in Image2Token and Token2Image computations, at least one of the query (Q), key (K) or value (V) is from a very short prompt sequence or prototype sequence, which effectively reduces the computational complexity and memory consumption. On the other hand, we significantly compressed the computational dimension of the module to achieve the same objective. Figure 8 illustrates the overall workflow of MSPrompter. The fusion of the query image and the common prototype significantly enhances the model's attention to the segmented target area. During the process of deeply fusing the prompt information with the common prototype to realize prompt transfer, the prompt information that was focused on a specific target gradually shifts to a global foreground feature centered on the common prototype.

3.6 Simplified Mask Decoder

As illustrated in Fig. 4, we design a new mask decoder module for the vessel segmentation in OCTA images. The mask decoder contains two parts, namely PUM module which is based on update prompt embedding and PMIM module which is based on merging prompt embedding and image embedding. Since the OCTA vessel segmentation task only performs binary prediction, IoU tokens and the IoU prediction branch are omitted. To be able to capture pixel features, transposed convolution and depthwise separable convolution are inserted before Image2Token attention in MPIM module. In the MPIM module, we additionally insert a learnable query embedding E_{query} for supervised prediction results.

In the Simplified Mask Decoder, the encoded image embeddings are restored by PMIM modules, and the prompt embeddings required for the PMIM module

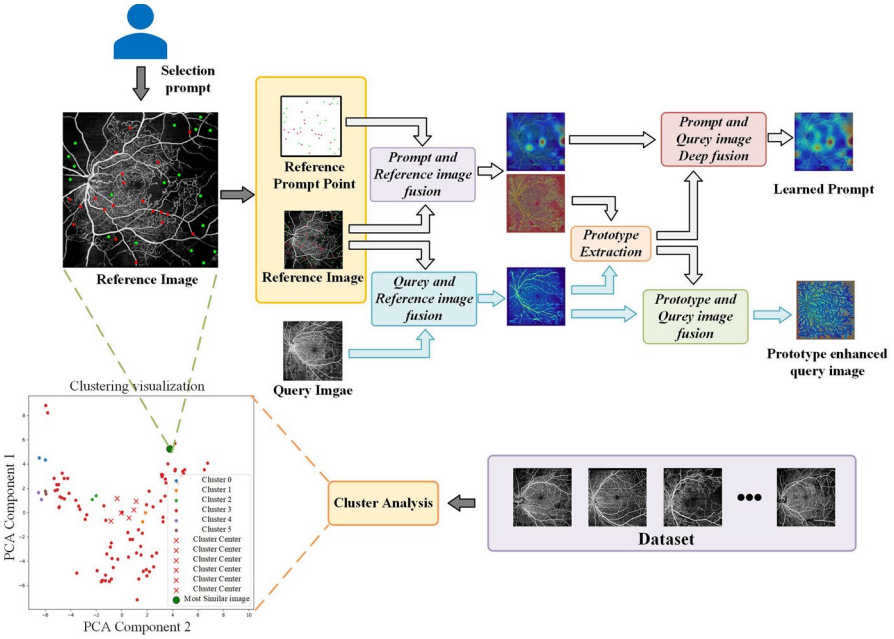

Fig. 8 Overall workflow of MSPrompter. The input reference image is selected from the dataset by the K-means algorithm, which is the image closest to other image features. The reference prompt is a positional encoding embedding generated by applying coordinate normalization and trigonometric encoding algorithms to the coordinates of the prompt points.

coming from a unique PUM module. The final single-channel output is derived by performing a matrix multiplication between the learnable query embedding E_{query} and the decoding features.

3.7 Loss function

This paper utilizes a joint loss function, consisting of the global loss $\mathcal{L}_S$ for final segmentation prediction and the prompt prediction loss $\mathcal{L}_P$ for guiding the decoding at the primary stage. The joint loss function is calculated as follows:

$$\mathcal{L}_{\text{Joint}} = \lambda \mathcal{L}_S + (1 - \lambda)\mathcal{L}_P, \tag{22}$$

where λ is the confidence coefficient that balances these two types of losses. The parameter λ is initially set to 0.8 and decreases exponentially with a decay factor of 0.005. $\mathcal{L}_S$ is the sum of the binary cross-entropy loss and the Dice loss, computed between the predicted category and the ground truth. $\mathcal{L}_P$ is the sum of the binary cross-entropy loss and the Dice loss computed between the prompt-guided coarse mask and the downsampled ground-truth labels. Binary cross-entropy loss and Bice loss, defined as follows:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(p_i \cdot y_i)}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} y_i^2}, \tag{23}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log(p_i) + (1 - y_i) log(1 - p_i)\right]. \tag{24}$$

Here N is the total number of pixels, y_i is the ground-truth label for the i^{th} pixel, and p_i is the predicted probability for the i^{th} pixel. The Dice loss function is mainly used for image-level optimization, while the binary cross-entropy loss function is used for pixel-level optimization.

4 Experiment

4.1 Datasets and implementation details

1) Datasets: All methods are conducted on the OCTA500-3 M and OCTA500-6 M datasets and the newly collected OCTA-RetinaVessel dataset (OCTA-RV).

OCTA500 The OCTA500 [11] dataset contains two subsets with different fields of view (FOVs), OCTA500-3 M and OCTA500-6 M. The OCTA500 dataset was collected using a commercial 70 kHz frequency-domain OCT system (RTVue-XR, Optovue, CA) with a center wavelength of 840 nm. OCTA500-3 M contains OCTA images from 200 different subjects, mainly from the normal population, each with a field of view of 3×3 mm and a resolution of 304×304 pixels. OCTA500-6 M contains OCTA images from 300 different subjects, mainly from the population with retinal common diseases. Each image has a field of view of 6×6 mm and a resolution of 400×400 pixels. The pixel-level ground truth was hand-mapped by five trained researchers and reviewed by three ophthalmologists.

OCTA-RV This dataset was captured and approved by the Institutional Review Committee of the Sixth People's Hospital of Shanghai Jiaotong University, and this study strictly followed the Declaration of Helsinki. All subjects have signed an informed consent form. The OCTA-RV dataset was captured by SS-OCT system (VG200D, SVision Imaging, Ltd, China) from 62 people (including diabetic retinopathy patients and healthy people) with a scanning area of 12×12 mm^2. The OCTA-RV dataset contains images with a resolution of 304×304 pixels. Benefiting from WF-OCTA (wide-field optical coherence tomography angiography), 12×12 mm^2 FOV provides a wider field of view compared to 6×6 mm^2 FOV and 3×3 mm^2 FOV, enabling more ocular structures and lesions to be captured in a single scan. All the subjects have been certified by professional physicians, and the work process strictly follows the Privacy Protection protocol. The annotation of samples were completed by researchers under the guidance of ophthalmologist, and the accuracy of each image annotation was based on pixel level. As illustrated in Fig. 9, OCTA-RV contains choroidal neovascularisation (CNV), microaneurysm (MA), retinal non-perfusion (RNP), intraretinal microvascular abnormality (IMA) and other

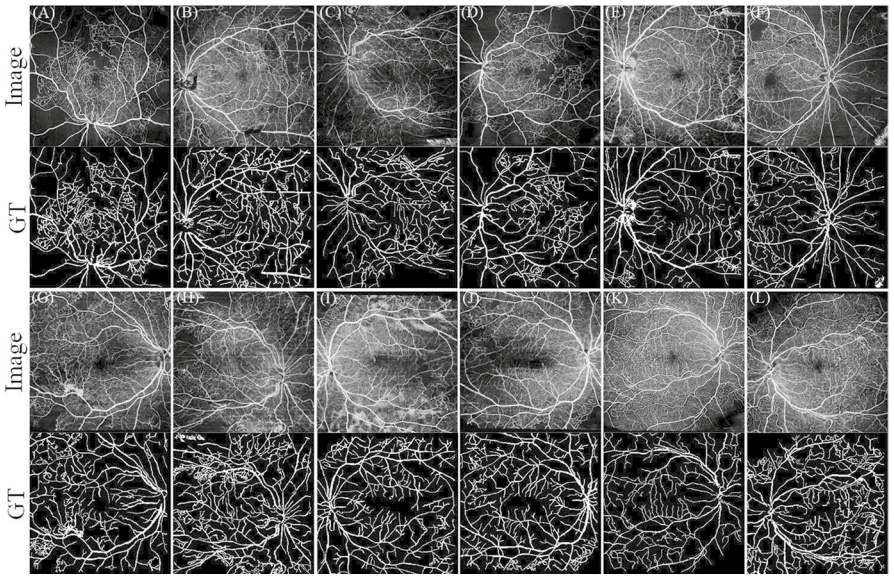

Fig. 9 OCTA images and ground truths in the OCTA-RV dataset. Top Row: the original OCTA retinal images. Bottom Row: the ground truths of retinal images. To enrich the image modalities of the OCTA-RV dataset, the images and labels within the OCTA-RV dataset were subjected to image enhancement processing, including random horizontal flipping, vertical flipping and random rotation.

retinal vascular lesions samples. The new dataset will be of great significance for vessel segmentation of retinal lesions. Due to the difficulty of data collection and the high cost of labeling time, only 62 OCTA images are included in the OCTA-RV dataset, and a total of 372 images are used for experiments after data augmentation.

2) Implementation Details: We implemented all the experiments using the PyTorch platform and performed them on a single RTXA4000 GPU (16GB). Following the fairness principle, all methods uniformly use Adam with an initial learning rate of 0.0005 and weight decay of 0.001 as the optimizer for training. All the images of the datasets are resized to a multiple of 32 through zero-pixel padding to avoid size mismatch problems during the model inference process. On all datasets, the training is performed for 300 epochs with a batch size of 4. All the datasets were data augmented by random horizontal/vertical flips and random rotation and divided into training/validation/test according to the ratio of 8:1:1. It is worth noting that to avoid overfitting of the model during training, we employed K-fold cross-validation with $K = 10$ for each experiment. The method is executed ten times, and the average score of the best result obtained on the test set for each fold is used as the performance evaluation score for the model. For the comparison model, the remaining parameters are set strictly according to the reference paper. For all prompt-based models, we only experimented with the point prompt setting due to the OCTA image characteristics. For the point prompt setting, we randomly select 20 positive/negative points in the foreground/background in the ground truth as the point prompt. The whole training process of WS-SAM only needs to input a reference image and

the corresponding reference prompt. The source code will be available at https://zhaohuizhang0809.github.io/OCTA-RV.

3) Evaluation Metrics: To objectively evaluate the segmentation performance of the proposed WS-SAM and the competitive methods, we utilize the following metrics, including sensitivity (SEN), specificity (SPE), Dice coefficient (DICE) and Jaccard Index (JAC). The related definitions are as follows:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \tag{25}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}), \tag{26}$$

$$\text{Dice Coefficient} = 2 \times \text{TP}/(2 \times \text{TP} + \text{FP} + \text{FN}), \tag{27}$$

$$\text{Jaccard} = \text{TP}/(\text{TP} + \text{FP} + \text{FN}). \tag{28}$$

Here *TP* (true positives) denotes the number of correctly predicted blood vessel pixels, *TN* (true negatives) denotes the number of correctly predicted non-vessel pixels. *FP* (false positives) denotes the number of non-vessel pixels that are incorrectly predicted as vessel pixels, and *FN* (false negatives) denotes the number of vessel pixels that are incorrectly predicted as non-vessel pixels.

4.2 Ablation study

1) Effectiveness of the Proposed Modules: To investigate the effectiveness of the wavelet encoder, pixel-wise image encoder, wavelet space fusion module, Meta Self-Prompter and Simplified Mask Decoder, we further performed ablation experiments of modules on the OCTA500 and OCTA-RV datasets, as illustrated in Table 2. The results of all ablation experiments were obtained on the test sets of the three datasets, respectively. Firstly, the effectiveness of the dual-branch encoder is verified by removing the wavelet encoder and pixel-wise image encoder. By removing the wavelet encoder from WS-SAM (w/o wavelet encoder), the performance declines with the average Dice of 2.8%, 1.46% and 3.49% on the three datasets. Similarly, removing the pixel-wise image encoder (w/o pixel-wise image encoder) causes WS-SAM to decrease the average Dice by 1.79%, 1.10% and 1.50% on the three datasets, respectively. The experimental results show that the design of dual-branch encoder is effective for improving the performance of WS-SAM in vessel segmentation on OCTA images. To verify the effectiveness of integrating wavelet space domain features, we removed the wavelet space fusion module (w/o wavelet space fusion module). By removing the wavelet space fusion module from WS-SAM, the performance declines with the average Dice of 1.68%, 0.95% and 1.47% on the three datasets. Finally, to verify the effective generation and integration of prompt information by Meta Self-Prompter and Simplified Mask Decoder, we removed each component individually (w/o Meta Self-Prompter; w/o Simplified Mask Decoder). By removing the Meta Self-Prompter from WS-SAM, the performance declines with the average Dice of 1.88%, 1.06% and 1.26% on the three datasets. Similarly, removing

Table 2 Ablation experiments on OCTA500 and OCTA-RV datasets. Best results are in bold

Datasets	Methods	SEN	SPE	DICE	JAC
OCTA500_3M	w/o Wavelet Encoder	0.8623	0.9905	0.8705	0.7689
	w/o Pixel-wise Image Encoder	0.8587	**0.9927**	0.8682	0.7653
	w/o WSF-Former	0.8634	0.9904	0.8712	0.7692
	w/o MSPrompter	0.8589	0.9900	0.8695	0.7634
	w/o Simplified Mask Decoder	0.8651	0.9908	0.8724	0.7705
	Our	**0.8716**	0.9917	**0.8754**	**0.7746**
OCTA500_6M	w/o Wavelet Encoder	0.8857	0.9883	0.8803	0.7871
	w/o Pixel-wise Image Encoder	0.8653	**0.9921**	0.8838	0.7993
	w/o WSF-Former	0.8776	0.9912	0.8854	0.8017
	w/o MSPrompter	0.8741	0.9912	0.8843	0.8002
	w/o Simplified Mask Decoder	0.8826	0.9904	0.8844	0.8002
	Our	**0.8914**	0.9903	**0.8949**	**0.8150**
OCTA-RV	w/o Wavelet Encoder	0.7417	0.9417	0.7063	0.5483
	w/o Pixel-wise Image Encoder	0.7733	0.9417	0.7262	0.5730
	w/o WSF-Former	0.7741	0.9418	0.7265	0.5733
	w/o MSPrompter	0.7665	0.9451	0.7286	0.5762
	w/o Simplified Mask Decoder	0.7769	0.9417	0.7282	0.5752
	Our	**0.7913**	**0.9465**	**0.7412**	**0.5902**

Simplified Mask Decoder caused WS-SAM to decrease the average Dice by 1.63%, 1.05% and 1.30% on the three datasets, respectively. The ablation experiments show that the model framework of our proposed WS-SAM is effective, and the design of each module plays an important role in promoting the segmentation effect of WS-SAM on OCTA images.

2) Analysis of PCW-Attention: To verify the performance and efficiency of the proposed PCW-Attention, an attention comparison experiment is conducted. Specifically, we replaced the spatial encoding backbone of WS-SAM with PCWA, Swin Transformer (SwinT) and Vision Transformer (ViT), respectively, and conducted a performance comparison. To ensure experimental fairness, all backbone networks are set to 3 layers, containing 6, 2 and 2 sub-layers, respectively. As shown in Fig. 12, heatmaps were visualized at the end of the spatial encoding and each layer of the decoder output. The heatmap results indicate that, compared to SwinT and ViT, PCW-Former establishes more effective cross-regional non-local dependencies during encoding, thereby facilitating the inference and recovery of fine-grained features in decoding. Compared with traditional attention mechanisms, the heat maps generated by PCWA-Former are more focused and distinct. As evidenced by the heatmaps, PCWA-Former demonstrates superior efficiency in capturing critical information when processing long sequences or complex features, thereby leading to significant improvements in prediction accuracy. As shown in Table 5, PCW-Former not only has a faster computational speed than traditional self-attention backbones, but also achieves higher quantitative indicators in task execution. As shown in Fig. 14, compared with ViT and SwinT, our proposed PCWA-Former can

augment the global receptive field of the model to capture a wider range of contextual information.

3) Analysis of Hyper-parameter: It is worth noting that the number of prompt points is crucial to WS-SAM, as it affects the effectiveness of prompt transfer. For the number range of prompt point numbers in $10 \leq n \leq 50$, we perform a grid search using Dice metrics on three datasets. As illustrated in Fig. 10, when n equals 20, our WS-SAM attains the best performance. We hold the view that too few prompts may not provide effective reference information, while excessive prompts may interfere with the transfer of prompts. Overall, a moderate number of prompts is beneficial to WS-SAM and is capable of achieving favorable prompt transfer effects.

4.3 Comparison with the state-of-the-art methods

1) Qualitative Comparison: Figure 11 shows the predicted segmentation results of the proposed WS-SAM and the other competing methods on OCTA500-3 M, OCTA500-6 M and OCTA-RV, respectively. By comparing the prediction results on different datasets with the same noise interference, vessel distribution and morphological size, it can be found that WS-SAM can make more accurate prediction of tiny vessels better than the compared methods. As illustrated in Fig. 11, most of the compared networks cannot make accurate predictions of the blood vessels in the central concave region of the retina on the OCTA-RV dataset. Most of compared methods cannot accurately segment the microvessels on the OCTA500-3 M dataset. As the above analysis on three OCTA images with different resolutions, noise interference and vessel distribution, it has been proved the effectiveness and robustness of our proposed method.

2) Quantitative Comparison: As illustrated in Table 4, to evaluate the vessel segmentation effect of WS-SAM on OCTA images, we perform the comparisons with state-of-the-art medical image segmentation methods and medical SAM variants on OCTA500-3 M, OCTA500-6 M and OCTA-RV datasets. To ensure a fair comparison, we fine-tune all models to adapt to OCTA images on the same training set. For non-SAM methods, WS-SAM overall outperforms the second-best methods on OCTA500-3 M, OCTA500-6 M and OCTA-RV datasets. WS-SAM achieves a 1.87% and 1.02% Dice increase over the second-best DBUNet [26] on OCTA500-3 M and OCTA500-6 M datasets, and a 1.07% Dice increase over the second-best OCT2Former [17] on OCTA-RV dataset. Compared with MEGANet [20] that also integrates frequency-domain and spatial-domain information, the Dice values of WS-SAM on the three datasets were, respectively, increased by 4.11%, 1.21% and 6.83%. Compared to the self-prompting SAM variants, WS-SAM achieves a Dice increase of 11.89%, 4.80% and 6.0% over H-SAM [36] on the OCTA500-3 M, OCTA500-6 M and OCTA-RV datasets, respectively. Compared to Vanilla SAM [25] with point prompts, WS-SAM also achieves better results, with a Dice increase of 11.49%, 3.94% and 6.78% on the three datasets. Compared with MedSAM, the dice of WS-SAM has increased by 5.86%, 2.83% and 4.94%, respectively, on the three datasets. We believe that the suboptimal performance of traditional SAM model and their variants in OCTA image can be primarily attributed to

Table 3 Parameters and FLOPs comparison results (the input resolution size is 304 × 304 pixels)

Methods	FLOPs(G)	Params(M)
UNet	10.62	8.29
Attention UNet	103.99	34.88
MsTGANet	27.24	11.60
RetiFluidNet	25.22	8.56
DBUNet	102.86	24.79
OCT2Former	102.20	7.35
GFUNet	9.19	3.94
MsEGANet	21.39	29.27
MFMSNet	34.99	54.54
Vanilla SAM	45.78	40.88
H-SAM	43.03	97.47
MedSAM	41.75	67.21
DeSAM	41.64	85.89
Med-SA	45.39	93.13
WS-SAM	21.32	8.58

Note: "Params" denotes the number of parameters in the model, and "FLOPs" denotes the computational overhead

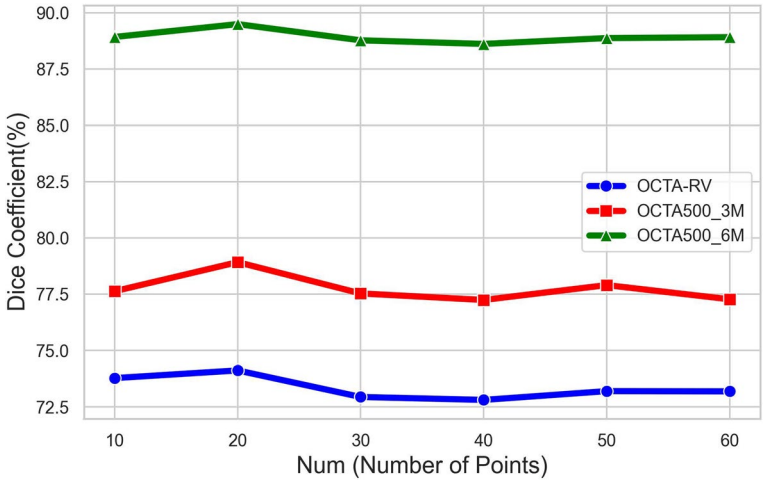

Fig. 10 Hyper-parameter analysis for the number of prompts on OCTA retinal vessel segmentation

two main aspects: Firstly, the current OCTA images have relatively low resolution, while SAM and its variants typically use high-resolution images as input. Secondly, the insufficiency of a medical image dataset of adequate scale for model training is likely to impose constraints on model performance. In contrast, the training data volumes of both the SAM and MedSAM models have reached the magnitude of several million. The experimental results show that WS-SAM is an effective OCTA vessel

Table 4 Comparison with state of the art on OCTA500 and OCTA-RV datasets. Best results are in bold

Methods	OCTA500-3 M					OCTA500-6 M						OCTA-RV			
	SEN	SPE	DICE	JAC	*p*-value	SEN	SPE	DICE	JAC	*p*-value	SEN SPE		DICE	JAC	*p*-value
UNet [13]	0.8412	0.9922	0.8568	0.7509	<0.001	0.8809	0.9903	0.8766	0.7986	<0.001	0.7508	0.9420	0.7132	0.5565	<0.001
Attention UNet [14]	0.8590	0.9909	0.8585	0.7537	<0.001	0.8755	0.9900	0.8818	0.8030	<0.001	0.7886	0.9362	0.7241	0.5702	<0.001
MsTGANet [9]	0.8336	0.9936	0.8617	0.7585	<0.001	0.8678	0.9868	0.8834	0.7923	<0.001	0.7210	0.9553	0.7216	0.5598	<0.001
RetiFluid-Net [10]	0.8644	0.9905	0.8589	0.7540	<0.001	0.8563	0.9924	0.8830	0.7915	<0.001	0.7907	0.9339	0.7204	0.5654	<0.001
DBUNet [26]	0.8657	0.9906	0.8606	0.7569	<0.001	0.8683	0.9913	0.8847	0.7943	<0.001	0.7680	0.9425	0.7281	0.5814	<0.001
OCT-2Former [17]	0.8316	0.9949	0.8610	0.7575	<0.001	0.8755	0.9903	0.8845	0.7942	<0.001	0.7369	0.9546	0.7305	0.5788	<0.001
GFUNet [19]	0.8089	0.9896	0.8189	0.6952	<0.001	0.8379	0.9871	0.8454	0.7333	<0.001	0.6623	0.9493	0.6703	0.5071	<0.001
MEGANet [20]	0.8388	0.9925	0.8575	0.7518	<0.001	0.8762	0.9898	0.8821	0.7901	<0.001	0.7820	0.9130	0.6785	0.5147	<0.001
MFMSNet [21]	0.8446	0.9909	0.8497	0.7402	<0.001	0.8646	0.9887	0.8695	0.7701	<0.001	0.7344	0.9349	0.6885	0.5271	<0.001
Vanilla SAM [25]	0.7743	0.9889	0.7942	0.6605	<0.001	0.8448	0.9880	0.8555	0.7486	<0.001	0.7029	0.9376	0.6734	0.5095	<0.001
H-SAM [36]	0.7271	0.9900	0.7653	0.6304	<0.001	0.8459	0.9858	0.8469	0.7355	<0.001	0.7636	0.9218	0.6812	0.5240	<0.001
MedSAM [37]	0.8564	0.9848	0.8168	0.6919	<0.001	0.8509	0.9883	0.8666	0.7686	<0.001	0.7056	0.9452	0.6918	0.5411	<0.001

Table 4 (continued)

Methods	OCTA500-3 M					OCTA500-6 M					OCTA-RV				
	SEN	SPE	DICE	JAC	*p*-value	SEN	SPE	DICE	JAC	*p*-value	SEN SPE		DICE	JAC	*p*-value
DeSAM [39]	0.8201	0.9913	0.8370	0.7212	<0.001	0.8564	0.9884	0.8623	0.7589	<0.001	0.6985	0.9437	0.6862	0.5231	<0.001
Med-SA [40]	0.8020	0.9905	0.8215	0.6986	<0.001	0.8471	0.9901	0.8662	0.7650	<0.001	0.7075	0.9515	0.6926	0.5424	<0.001
WS-SAM (our)	**0.8716**	0.9917	**0.8754**	**0.7746**	–	**0.8914**	0.9903	**0.8949**	**0.8150**	–	**0.7913**	0.9465	**0.7412**	**0.5902**	–

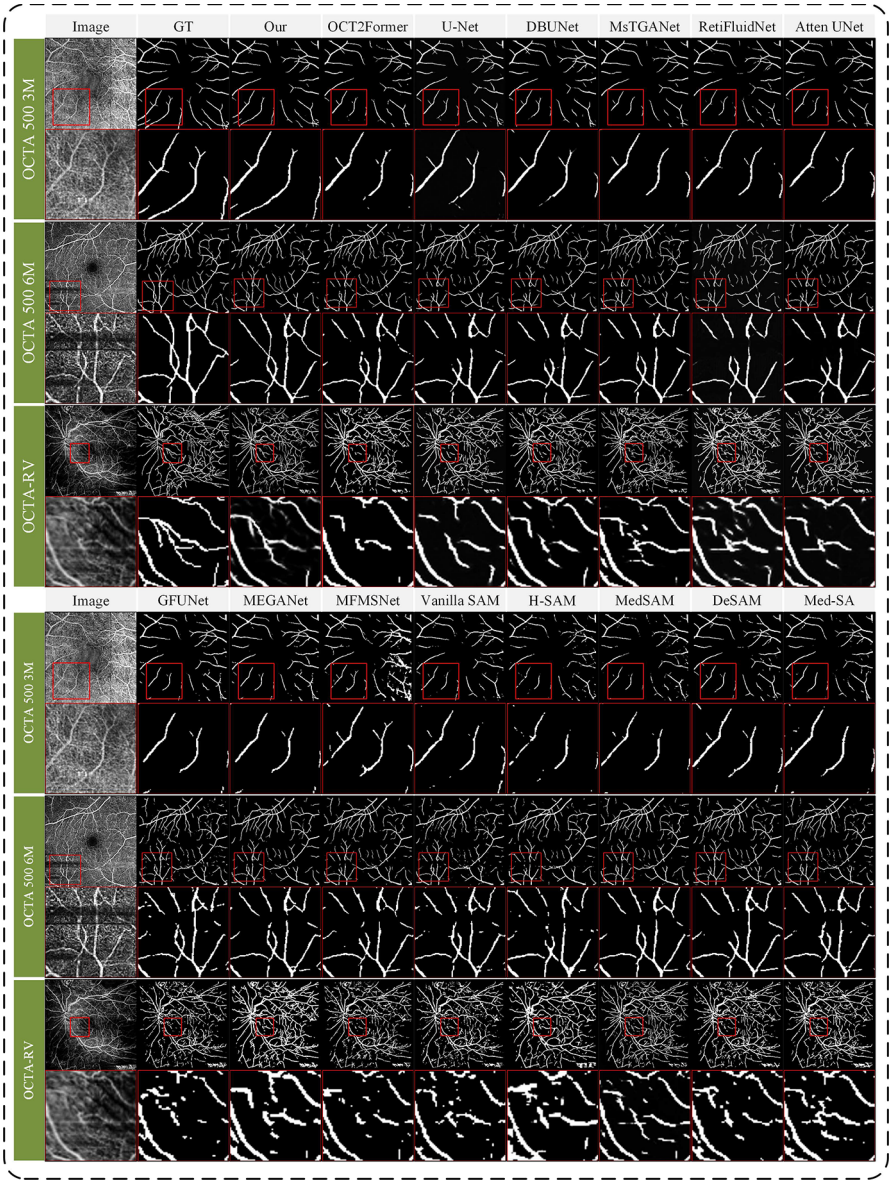

Fig. 11 Retinal vessel segmentation results of the proposed WS-SAM and compared methods. From top to bottom are the predicted segmentation results on OCTA500-3 M, OCTA500-6 M and OCTA-RV

segmentation framework. To demonstrate the significant superiority of the proposed method, the t-test of Dice Coefficient was conducted in all the comparison experiments. Specifically, each experiment was independently repeated ten times to ensure the reliability and reproducibility of the results. As illustrated in Table 4, all the p-values were less than 0.05, indicating that the method in this paper is significantly

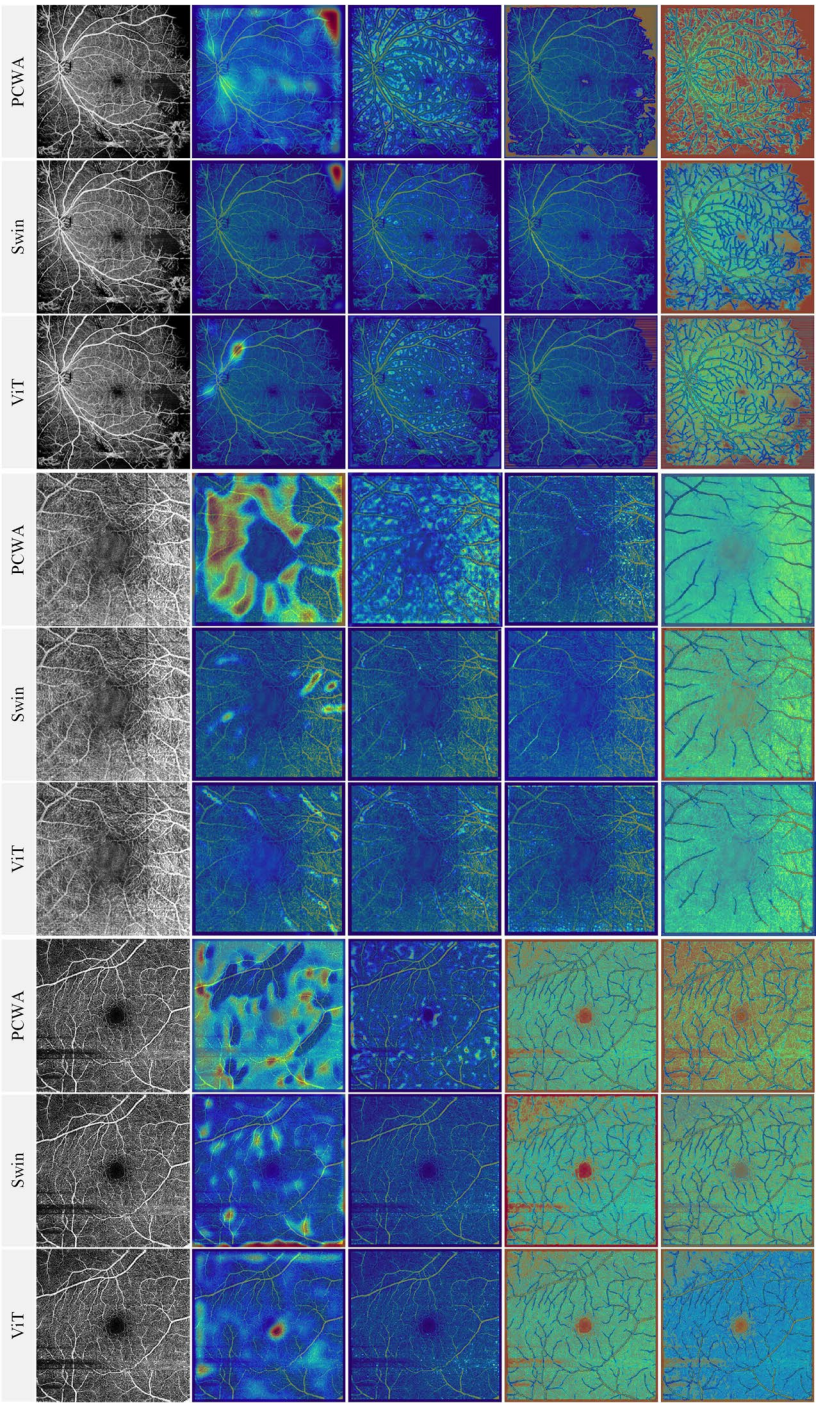

Fig. 12 Comparison of heatmap visualizations of PCWA-Former, Swin Transformer and Vision Transformer as the spatial encoder backbone for segmentation tasks on three datasets

superior to the comparison methods. To conduct a more comprehensive assessment of the segmentation effect, the precision–recall curve (PR), as illustrated in Fig. 13. Compared with other models, WS-SAM can better balance high recall rate and high precision on three datasets. Meanwhile, WS-SAM exhibited steady loss convergence during the training processes of the three datasets, as demonstrated in Fig. 13. Compared with the OCTA500-6 M dataset, the overall accuracy predicted by the OCTA500-3 M dataset is lower. This is mainly due to the fact that the images contained in the OCTA500-3 M dataset have lower-resolution and low-contrast areas are widespread.

3) Efficiency Analysis: Table 3 shows the computational consumption (Flops) and parameter scale (Params) of WS-SAM and all the comparative networks. Following the fairness principle, we make adjustments to the maximum depth and dimension to ensure uniformity for all methods. As illustrated in Table 3, compared with OCT-2Former [17], WS-SAM reduces the computational consumption by approximately 5 times and has similar parameter numbers. In comparison with DBUNet [26], WS-SAM also reduces the computational consumption by about 5 times, and the number of model parameters is only one-third of DBUNet [26]. Compared with other models, the computational consumption of WS-SAM is only greater than that of UNet [13] and GFUNet [19], while the number of its parameters is relatively small. Moreover, while remaining at a relatively small Params and Flops, WS-SAM surpasses all other comparative models in terms of segmentation performance. To facilitate the quantitative analysis of the computational overhead of WS-SAM, we have meticulously analyzed the parameters and computational overheads of each module and conducted a comparative analysis of common model backbones, as shown in Tables 5 and 6. As indicated in Table 6, the WS-SAM model has a relatively small number of parameters, which is attributed to the effective compression of parameters

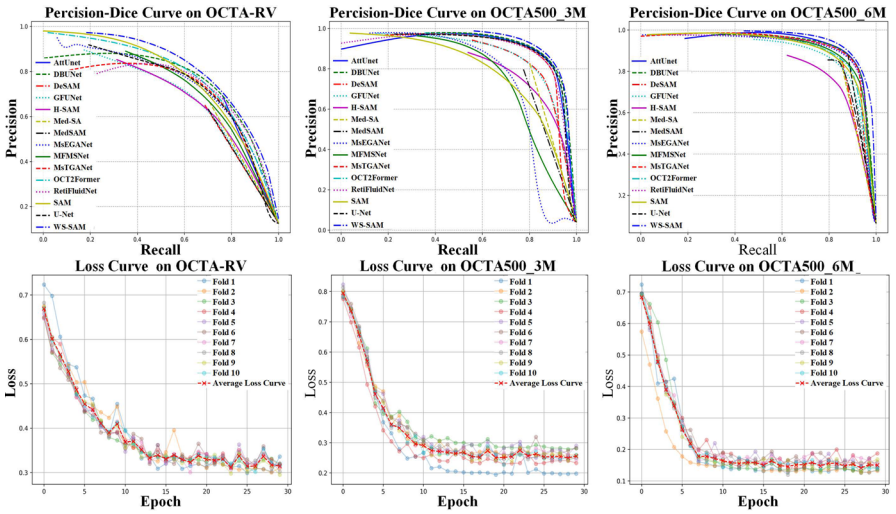

Fig. 13 Precision–recall curve of different methods on test samples and the loss convergence curve of WS-SAM on train sample for OCTA-RV, OCTA500-3 M and OCTA500-6 M

Table 5 Comparison of PCW-Former, Swin Transformer (SwinT) and Vision Transformer (ViT) as spatial encoder backbone on OCTA-RV and OCTA500 dataset. Best results are in bold

dataset	Methods	Flops (G)	Params (M)	Dice (%)
OCTA-RV	ViT	9.68	5.47	0.7302
	SwinT	**7.37**	**5.48**	0.7271
	PCW-Former (Our)	7.39	5.69	**0.7412**
OCTA500-3 M	ViT	9.68	5.47	0.7756
	SwinT	**7.37**	**5.48**	0.7782
	PCW-Former (Our)	7.39	5.69	**0.7892**
OCTA500-6 M	ViT	19.06	5.47	0.8880
	SwinT	**12.45**	**5.48**	0.8897
	PCW-Former (Our)	12.48	5.69	**0.8949**

Table 6 Number of parameters (Params) and computational overhead (Flops) of each module e on OCTA-RV dataset

Methods	Params (M)	Flops (G)
Pixel-wise encoder	5.69	7.39
Wavelet encoder	0.41	0.55
WSF former	0.47	3.55
MSPrompter	0.46	4.53
Mask decoder	0.21	4.66

in each module. For example, the image encoder of the traditional SAM is directly composed of several stacked transformers. On the contrary, the dual-domain encoder we designed performs downsampling operations between each layer, effectively suppressing the number of parameters. Furthermore, we have endeavored to minimize the size of the channel dimension to the greatest extent possible. As shown in Table 5, compared with the common encoder backbones, the PCW-Former we proposed not only maintains a smaller number of parameters and a faster inference speed, but also demonstrates a significant advantage in segmentation accuracy. This is predominantly attributed to the fact that we utilized a window partitioning strategy for designing the attention mechanism, significantly reducing the computational consumption while guaranteeing the segmentation effect. In Tables 3, 5 and 6, the proposed model demonstrates significant lightweight characteristics in terms of the overall parameter, the parameter distribution of each module and the design of the attention mechanism. These results indicate that WS-SAM has achieved an effective balance in aspects such as parameter optimization and computational efficiency, demonstrating its application potential in low-resource allocation environments.

The advantages of our proposed method are twofold: (1) Firstly, we design a pixel-wise image encoder with a small amount of parameters and computational consumption, which can improve fine-grained perception and ensure high-quality segmentation results. (2) Secondly, the meta-prompeter can guide the inference process and improve the segmentation effect of the original encoder–decoder.

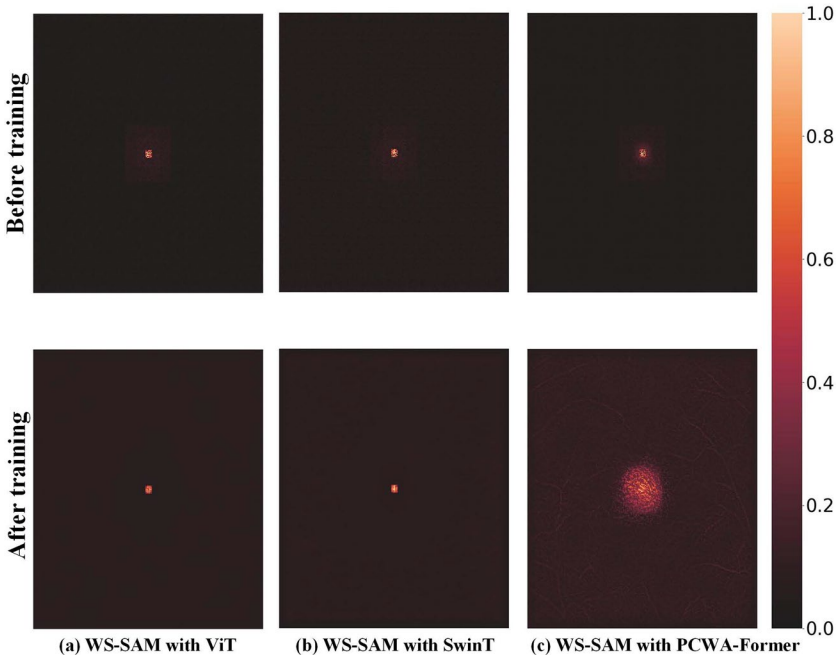

Fig. 14 Receptive field size comparison before and after training when performing ViT, SwinT and PCWA-former as backbones for WS-SAM, respectively

Effect analysis: As shown in Fig. 15, the confusion matrix analysis indicates that the WS-SAM model demonstrates a high true positive rate (TPR) and a low false negative rate (FNR) in the retinal vessels segmentation task, suggesting that the model can accurately identify most of the true vascular pixels. Additionally, the low false positive rate of the WS-SAM model further validates its superior performance in identifying non-vascular regions.

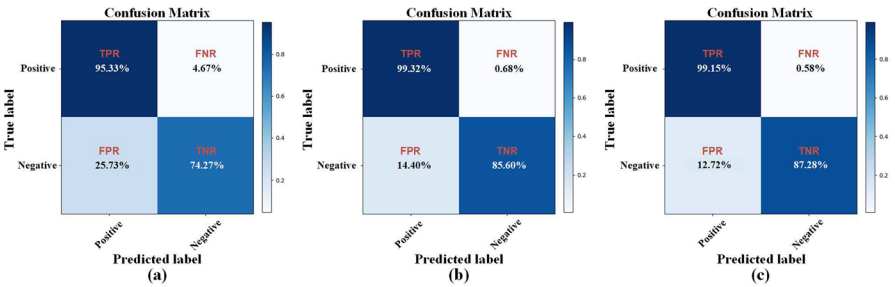

Fig. 15 Confusion matrix results of WS-SAM on three datasets. **a** OCTA-RV, **b** OCTA500_3M, **c** OCTA500_6M. In the confusion matrix, the rows denote the actual foreground and background labels, whereas the columns represent the predicted labels. TPR denotes true positive rate, FNR denotes false negative rate, FPR denotes false positive rate, and TNR denotes true negative rate

5 Discussion

In OCTA images, noise and artifacts can have an adverse effect on the diagnosis of retinal diseases. Wide-field OCTA (WF-OCTA) is capable of offering a broader field of view, but the minute blood vessels contained therein are hard to discriminate by the human eye. To address this issue, we propose a self-prompting model named WS-SAM. Based on self-prompting for segmentation, WS-SAM realizes an enhancement in segmentation accuracy and strictly controls its parameters quantity and computational consumption. Although the method we proposed has made progress in OCTA retinal vessel segmentation, it is not without limitations. One of the limitations of this approach is that we choose to employ class weight [36] to tackle the issue of class imbalance, and this method is not learnable. Although class weight can alleviate to some extent the problem of class imbalance, it is difficult to overcome the situation of extreme class bias, which requires human intervention. Additionally, while we have greatly reduced the number of parameters and computational overhead of the model, we still hope it will be more lightweight.

Our future work will address these limitations by developing more comprehensive and adaptable methodologies. We hope to solve the problem of class imbalance by introducing learnable similarity metric parameters to establish effective distance measures between samples, thus distinguishing minority class samples from similar samples. Furthermore, we hope to introduce lightweight attention or state spaces [48], thereby achieving further model lightweighting.

6 Conclusion

To improve the effectiveness of vessel segmentation in OCTA images, we propose a self-prompting SAM based on wavelet and spatial domain, named WS-SAM. We employed the modality of joint frequency-domain and spatial-domain coding, with the purpose of capturing multi-scale feature information from distinct domains. WS-SAM is composed of wavelet encoder, pixel-wise image encoder, wavelet space fusion (WSF-Former) module, Meta Self-Prompter (MSPrompter) and Simplified Mask Decoder. In particular, the wavelet encoder and pixel-wise image encoder have been proposed to extract wavelet and spatial-domain features, which provide cross-domain information of different viewpoints instead of separate spatial-domain information. Then, the WSF-Former module is designed to adaptively fuse the cross-domain information from the output of wavelet encoder and pixel-wise image encoder. To implement self-prompting, we further propose MSPrompter, which implements prototype enhancement and prompt transfer in the way of meta-learning. Finally, we simplify the mask decoder proposed by vanilla SAM. To handle the data shortage problem, we constructed a new OCTA vessel segmentation dataset named OCTA-RV, which focuses on the labeling deep microvessels. The extensive experiments with state-of-the-art methods on

both OCTA500 and OCTA-RV datasets show that the proposed WS-SAM can achieve excellent segmentation performance on OCTA retinal vessel segmentation task.

Acknowledgements The authors gratefully acknowledge the financial supports by the Natural Science Foundation of Shandong Province (No. ZR2020MF105), the Guangdong Provincial Key Laboratory of Biomedical Optical Imaging Technology (No. 2020B121201010), the Natural National Science Foundation of China (62175156,61675134), the Science and technology innovation project of Shanghai Science and Technology Commission (19441905800, 22S31903000), the Qufu Normal University Foundation for High Level Research (116-607001) and the Natural Science Foundation of Jiangxi Province (grant number 20232BAB202053).

Author contribution a . Zhaohui Zhang is responsible for the writing of this article. Zhaohui Zhang, Yanfei Guo, Hongjuan Liu are responsible for data statistics and chart design. Xiwei Dong provides experimental ideas. Fei Ma and Jing Meng are responsible for supervising and guiding the experimental process.

Funding This work was supported by the Natural Science Foundation of Shandong Province (No. ZR2020MF105), the Guangdong Provincial Key Laboratory of Biomedical Optical Imaging Technology (No. 2020B121201010), the Natural National Science Foundation of China (62175156,61675134), the Science and technology innovation project of Shanghai Science and Technology Commission (19441905800, 22S31903000), the Qufu Normal University Foundation for High Level Research (116-607001) and the Natural Science Foundation of Jiangxi Province (grant number 20232BAB202053).

Data availability The data related to the current research can be acquired from the corresponding author upon reasonable requests.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval and consent to participate This dataset collection was reviewed and approved by the Institutional Review Committee of the Sixth People's Hospital of Shanghai Jiaotong University, and this study strictly followed the Declaration of Helsinki. The authors take full responsibility for all aspects of the work, ensuring that any issues concerning its accuracy or completeness are thoroughly investigated and addressed.

References

1. Aschauer J, Klimek M, Donner R, Lammer J, Roberts P, Schranz M, Schmidinger G (2024) Non-invasive quantification of corneal vascularization using anterior segment optical coherence tomography angiography. Sci Rep 14:2124
2. Cavichini M, Dans KC, Jhingan M, Amador-Patarroyo MJ, Borooah S, Bartsch D-U, Nudleman E, Freeman WR (2021) Evaluation of the clinical utility of optical coherence tomography angiography in age-related macular degeneration. Br J Ophthalmol 105:983–988
3. Yuan M, Wang W, Kang S, Li Y, Li W, Gong X, Xiong K, Meng J, Zhong P, Guo X, Wang L, Liang X, Lin H, Huang W (2022) Peripapillary microvasculature predicts the incidence and development of diabetic retinopathy: an SS-OCTA study. Am J Ophthalmol 243:19–27
4. Meleppat RK, Fortenbach CR, Jian Y, Martinez ES, Wagner K, Modjtahedi BS, Motta MJ, Ramamurthy DL, Schwab IR, Zawadzki RJ (2022) In vivo imaging of retinal and choroidal morphology and vascular plexuses of vertebrates using swept-source optical coherence tomography. Trans Vision Sci Technol 11:11–11. https://doi.org/10.1167/tvst.11.8.11

5. Meleppat RK, Miller EB, Manna SK, Zhang P, Jr. ENP, Zawadzki RJ (2019) Multiscale Hessian filtering for enhancement of OCT angiography images. In: Manns F, Söderberg PG, Ho A (eds) Ophthalmic Technologies XXIX. SPIE, p 108581K
6. Chung SH, Sin T-N, Dang B, Ngo T, Lo T, Lent-Schochet D, Meleppat RK, Zawadzki RJ, Yiu G (2022) CRISPR-based VEGF suppression using paired guide RNAs for treatment of choroidal neovascularization. Mol Ther - Nucleic Acids 28:613–622. https://doi.org/10.1016/j.omtn.2022.04.015
7. Chen L, Yuan M, Sun L, Chen Y (2022) Three-dimensional analysis of choroidal vessels in the eyes of patients with unilateral BRVO. Front. Med. (Lausanne) 9:854184
8. Mishra S, Wang YX, Wei CC, Chen DZ, Hu XS (2021) VTG-Net: a CNN based vessel topology graph network for retinal artery/vein classification. Front Med (Lausanne) 8:750396
9. Wang M, Zhu W, Shi F, Su J, Chen H, Yu K, Zhou Y, Peng Y, Chen Z, Chen X (2022) MsTGANet: automatic drusen segmentation from retinal OCT images. IEEE Trans Med Imag 41:394–406
10. Rasti R, Biglari A, Rezapourian M, Yang Z, Farsiu S (2023) RetiFluidNet: a self-adaptive and multi-attention deep convolutional network for retinal OCT fluid segmentation. IEEE Trans Med Imag 42:1413–1423
11. Li M, Huang K, Xu Q, Yang J, Zhang Y, Ji Z, Xie K, Yuan S, Liu Q, Chen Q (2024) OCTA-500: a retinal dataset for optical coherence tomography angiography study. Med Image Anal 93:103092
12. Chinkamol A, Kanjaras V, Sawangjai P, Zhao Y, Sudhawiyangkul T, Chantrapornchai C, Guan C, Wilaiprasitporn T (2023) OCTAVE: 2D en face optical coherence tomography angiography vessel segmentation in weakly-supervised learning with locality augmentation. IEEE Trans Bio-med Eng/IEEE Trans Biomed Eng 70(6):1931–1942
13. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention - MICCAI 2015. Springer International Publishing, Cham, pp 234–241
14. Oktay O, Schlemper J, Folgoc LL, Lee MCH, Heinrich MP, Misawa K, Mori K, McDonagh SG, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention U-Net: learning where to look for the pancreas. CoRR abs/1804.03999
15. Mubashar M, Ali H, Grönlund C, Azmat S (2022) R2U++: a multiscale recurrent residual U-Net with dense skip connections for medical image segmentation. Neural Comput Appl 34:17723–17739
16. Shi Z, Li Y, Zou H, Zhang X (2023) TCU-Net: transformer embedded in convolutional U-Shaped network for retinal vessel segmentation. Sensors 23(10):4897
17. Tan X, Chen X, Meng Q, Shi F, Xiang D, Chen Z, Pan L, Zhu W (2023) OCT2Former: a retinal OCT-angiography vessel segmentation transformer. Computer Methods Programs Biomed 233:107454
18. Wang C, Chen X, Ning H, Li S (2024). SAM-OCTA: a fine-tuning strategy for applying foundation model OCTA image segmentation tasks, In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp 1771–1775
19. Li P, Zhou R, He J, Zhao S, Tian Y (2023) A global-frequency-domain network for medical image segmentation. Comput Biol Med 164:107290
20. Bui N-T, Hoang D-H, Nguyen Q-T, Tran M-T, Le N (2024) MEGANet: multi-scale edge-guided attention network for weak boundary polyp segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024. IEEE, pp 7970–7979
21. Wu R, Lu X, Yao Z, Ma Y (2024) MFMSNet: a multi-frequency and multi-scale interactive cnn-transformer hybrid network for breast ultrasound image segmentation. Comput Biol Med 177:108616
22. Ramos-Soto O, Rodríguez-Esparza E, Balderas-Mata SE, Oliva D, Hassanien AE, Meleppat RK, Zawadzki RJ (2021) An efficient retinal blood vessel segmentation in eye fundus images by using optimized top-hat and homomorphic filtering. Computer Methods Programs Biomed 201:105949
23. Chen H, Wei W, Zhang Y (2024) Optical coherence tomography image despeckling based on saliency enhancement and high-order singular value Marchenko-Pastur truncation. Physica Scripta 100:015003
24. Chen H, Qiao H, Wei W, Li J (2024) Time fractional diffusion equation based on caputo fractional derivative for image denoising. Optics Laser Technol 168:109855
25. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick RB (2023) Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, pp 3992–4003

26. Wang C, Ning H, Chen X, Li S (2023) DB-UNet: MLP based dual branch UNet for accurate vessel segmentation in OCTA images. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Rhodes Island, Greece
27. Ma Y, Hao H, Xie J, Fu H, Zhang J, Yang J, Wang Z, Liu J, Zheng Y, Zhao Y (2021) ROSE: a retinal OCT-angiography vessel segmentation dataset and new model. IEEE Trans Med Imag 40:928–939
28. Hao J, Shen T, Zhu X, Liu Y, Behera A, Zhang D, Chen B, Liu J, Zhang J, Zhao Y (2022) Retinal structure detection in OCTA image via voting-based multitask learning. IEEE Trans Med Imag 41:3969–3980
29. Ning H, Wang C, Chen X, Li S (2024) an accurate and efficient neural network for OCTA vessel segmentation and a new dataset. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024. IEEE, pp 1966–1970
30. Abtahi M, Le D, Lim JI, Yao X (2022) MF-AV-Net: an open-source deep learning network with multimodal fusion options for artery-vein segmentation in OCT angiography. Biomed Opt Express 13:4870–4888
31. Liu Y, Carass A, Zuo L, He Y, Han S, Gregori L, Murray S, Mishra R, Lei J, Calabresi PA, Saidha S, Prince JL (2022) Disentangled representation learning for OCTA vessel segmentation with limited training data. IEEE Trans Med Imag 41:3686–3698
32. Xu X, Yang P, Wang H, Xiao Z, Xing G, Zhang X, Wang W, Xu F, Zhang J, Lei J (2023) AV-casNet: fully automatic arteriole-venule segmentation and differentiation in OCT angiography. IEEE Trans Med Imag 42:481–492
33. Wu Z, Wang Z, Zou W, Ji F, Dang H, Zhou W, Sun M (2021) PAENet: a progressive attention-enhanced network for 3D to 2D Retinal vessel segmentation. In: Huang Y, Kurgan LA, Luo F, Hu X, Chen Y, Dougherty ER, Kloczkowski A, Li Y (eds) IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021. IEEE, pp 1579–1584
34. Yang C, Li B, Xiao Q, Bai Y, Li Y, Li Z, Li H, Li H (2024) LA-Net: layer attention network for 3D-to-2D retinal vessel segmentation in OCTA images. Phys Med Biol 69:045019
35. Yang C, Fan J, Bai Y, Li Y, Xiao Q, Li Z, Li H, Li H (2024) ODDF-Net: multi-object segmentation in 3D retinal OCTA using optical density and disease features. Knowl-Based Syst 306:112704
36. Cheng Z, Wei Q, Zhu H, Wang Y, Qu L, Shao W, Zhou Y (2024) Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. IEEE, pp 3511–3522
37. Ma J, He Y, Li F, Han L, You C, Wang B (2024) Segment anything in medical images. Nature Commun 15:654
38. Wu J, Xu M (2024) One-prompt to segment all medical images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. IEEE, pp 11302–11312
39. Yue W, Zhang J, Hu K, Xia Y, Luo J, Wang Z (2024) Surgicalsam: efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp 6890–6898
40. Wu J, Fu R, Fang H, Liu Y, Wang Z, Xu Y, Jin Y, Arbel T (2023) Medical SAM adapter: adapting segment anything model for medical image segmentation. CoRR abs/2304.12620
41. Chen K, Liu C, Chen H, Zhang H, Li W, Zou Z, Shi Z (2024) RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Trans Geosci Remote Sens 62:1–17
42. Hui W, Zhu Z, Zheng S, Zhao Y (2024) Endow SAM with Keen eyes: temporal-spatial prompt learning for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 19058–19067
43. Leng T, Zhang Y, Han K, Xie X (2024) Self-sampling meta SAM: enhancing few-shot medical image segmentation with meta-learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024. IEEE, pp 7910–7920
44. He W, Zhang Y, Zhuo W, Shen L, Yang J, Deng S, Sun L (2024) APSeg: auto-prompt network for cross-domain few-shot semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. IEEE, pp 23762–23772
45. Zhang R, Jiang Z, Guo Z, Yan S, Pan J, Dong H, Qiao Y, Gao P, Li H (2024) Personalize segment anything model with one shot. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net

46. Liu Y, Zhu M, Li H, Chen H, Wang X, Shen C (2024) Matcher: segment anything with one shot using all-purpose feature matching. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net
47. Hassani A, Walton S, Li J, Li S, Shi H (2023) Neighborhood Attention Transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Vancouver, BC, Canada
48. Dao T, Gu A (2024) Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net

Authors and Affiliations

Zhaohui Zhang[1] · Fei Ma[1] · Hongjuan Liu[1] · Xiwei Dong[2] · Yanfei Guo[1] · Jing Meng[1]

✉ Fei Ma
mafei@qfnu.edu.cn

Zhaohui Zhang
zhaohuizhang@qfnu.edu.cn

Hongjuan Liu
liuhongjuan66@qfnu.edu.cn

Xiwei Dong
dxwdxw2005@126.com

Yanfei Guo
guoyanfei2022@qfnu.edu.cn

Jing Meng
jingmeng@qfnu.edu.cn

[1] School of Computer Science, Qufu Normal University, Shandong, China

[2] School of Computer Science, Jiujiang University, Jiangxi, China