The Journal of Supercomputing (2025) 81:1333
https://doi.org/10.1007/s11227-025-07794-7

WHANet: wavelet and hybrid attention network for vessel segmentation in OCTA fundus images

Shuxin Xue[1] · Zhaohui Zhang[1] · Fen Yan[2] · Fei Ma[1] · Guangmei Jia[1] · Yanfei Guo[1] · Yuefeng Ma[1] · Xiaofei Ai[1] · Jing Meng[1]

Received: 6 May 2025 / Accepted: 21 August 2025

Abstract

Optical coherence tomography angiography (OCTA) is a state-of-the-art, non-invasive imaging modality that enables high-resolution visualization of the retinal vasculature, playing a vital role in the early diagnosis of various retinal diseases, including diabetic retinopathy, glaucoma, and choroidal neovascularization. However, traditional OCTA image analysis methods face significant challenges due to the difficulty in detecting retinal microvasculature, the presence of blurred vessel edges, and the complex topological structure of vascular networks. These difficulties are further compounded by limited training data, the inadequate representational capacity of spatial-domain features alone, and the challenges associated with effectively integrating frequency-domain information. To address these limitations, this study introduces WHANet, a novel dual-branch segmentation framework that integrates spatial and frequency-domain feature representations. WHANet comprises two primary modules: the hybrid attention deep convolutional branch (HADCB) and the multi-scale wavelet feature fusion branch (MWFFB). The HADCB enhances spatial feature representation through multi-scale convolutional operations and attention mechanisms, improving the detection of fine vessels, reducing edge blurring and misclassification, and reinforcing local detail perception. In parallel, MWFFB leverages the discrete wavelet transform (DWT) to extract refined vascular structures while preserving global vessel morphology. By complementing spatial-domain features with frequency-domain information, MWFFB strengthens edge representation, enhances the separability of microvessels, and mitigates the effects of image noise. The synergistic integration of these two branches enables comprehensive feature extraction and fusion, significantly boosting the model's ability to handle complex vascular networks. Extensive experiments conducted on three widely used public datasets demonstrate that WHANet consistently outperforms state-of-the-art methods across multiple evaluation metrics, exhibiting superior robustness and segmentation accuracy.

Extended author information available on the last page of the article

Published online: 13 September 2025

Keywords OCTA · WHANet · Retinal vessel segmentation · Hybrid Attention

1 Introduction

The retinal vascular system plays a vital role in maintaining the homeostasis of ocular microcirculation, with its branching density, uniformity of vessel diameter, and hemodynamic characteristics directly influencing oxygen delivery and metabolic efficiency. Systemic diseases such as diabetes mellitus and hypertension can lead to pathological changes in the retinal vasculature, including basement membrane thickening, vascular leakage, and abnormal neovascularization. Numerous studies have shown that conditions like diabetic retinopathy and cardiovascular disease are closely associated with structural and morphological alterations in the retinal vasculature, such as capillary occlusion, venous dilation, and arteriovenous crossing signs [1–3].

In recent years, optical coherence tomography angiography (OCTA) [4, 5] is a new and non-invasive imaging technique of visualization and quantitative analysis of the retinal microvasculature. By detecting the blood flow, OCTA facilitates layered reconstruction of microvascular networks and precise assessment of hemodynamic parameters. As an emerging imaging technology, OCTA provides real-time visualization of the microvascular structures of the retina and choroid, offering detailed blood flow information without fluorescein angiography. OCTA not only eliminates the risks and discomfort associated with contrast dye injection but also delivers clearer vascular detail and faster imaging speed. It can overcome the limitations of two-dimensional fundus photography by providing enriched vascular segmentation, including topological connectivity and spatial distribution.

These advantages have established OCTA as a powerful diagnostic tool for a range of ophthalmic conditions, such as age-related macular degeneration, diabetic retinopathy, and glaucoma [6–8]. However, early approaches for vascular segmentation primarily relied on traditional image processing techniques, including threshold-based Frangi filtering for vessel enhancement [9–14], morphological operations for noise suppression, and Hessian-based methods for enhancing tubular structures [15]. The traditional methods such as k-nearest neighbors (KNN) [16], Bayesian classifiers [17], and AdaBoost [18] have also been explored for retinal vessel segmentation. However, these methods typically suffer from limited robustness and scalability for complex OCTA fundus images [19, 20].

Recently, deep learning-based segmentation frameworks, particularly U-Net and Transformer-based architectures, have garnered increasing attention. These methods can effectively handle the issues of complex microvascular networks from multimodal OCTA data [21–25]. Furthermore, the frequency-domain method [26] can mine the effective features from complex structures. Frequency domain analysis techniques decompose images into various frequency components using Fourier or wavelet transforms, making them well suited for multi-scale characterization of retinal vasculature. In these representations, large vessels typically manifest in low-frequency components, while fine capillary networks are captured in high-frequency bands. The discrete wavelet transform (DWT) [27] enhances vessel boundaries

(high-frequency features) while simultaneously suppressing noise (high-frequency artifacts) through multi-resolution analysis. DWT can effectively mitigate spectral overlap between vascular signals and noise. Spatial domain methods directly process raw pixel intensities, which can preserve the spatial structure and topological continuity of vessels. Compared to the traditional Fourier Transform (FT), the Discrete Wavelet Transform (DWT) offers time-frequency localization capabilities, providing good resolution in both spatial and frequency domains. While the Fourier Transform focuses on the global frequency components, DWT can do multi-scale decomposition, making it capable of detecting high-frequency discontinuities (such as edges and noise) while preserving low-frequency structural information. This makes DWT adept at handling image processing tasks such as denoising, compression, and feature extraction. To accurately extract high-frequency details, such as fine textures and edge features, while preserving the overall structural integrity of the image, we adopt DWT instead of other frequency-domain methods (e.g., Short-Time Fourier Transform or Discrete Cosine Transform). The unique strength of DWT lies in its adaptive decomposition capability, which allows the analysis granularity to be adjusted according to the scale of different features, thereby enabling more effective separation and enhancement of critical frequency-domain information [28]. As shown in Fig. 1 , the Discrete Wavelet Transform (DWT) decomposes the original OCTA image into four frequency sub-bands: a low-frequency approximation component (LL) and three high-frequency detail components–LH (vertical details, marked in green), HL (horizontal details, marked in blue), and HH (diagonal details, marked in red). In the figure, red, green, and blue primary colors are used to visualize high-frequency components in different orientations. This not only enhances the visual contrast of the image but also facilitates the identification and analysis of structural features in various directions, such as vessel orientation and texture details. The LL sub-band preserves the overall structural information of the image, while the LH, HL, and HH sub-bands emphasize directional edge and detail features, providing a richer frequency-domain representation for subsequent segmentation and feature modeling tasks.

OCTA imaging is inherently prone to substantial speckle noise and intensity inhomogeneity, which can obscure vascular morphology and compromise segmentation

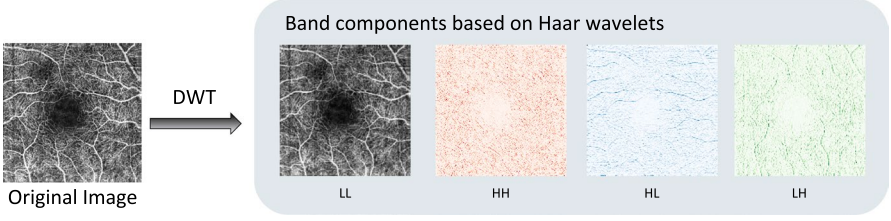

Fig. 1 Visualization of the frequency band decomposition results of the OCTA500-3 M image

fidelity. The characteristically low signal-to-noise ratio (SNR) of OCTA data manifests as ambiguous contrast differentiation between fine vascular structures (particularly capillaries) and background tissue, frequently inducing topological errors including vessel fragmentation, boundary leakage, or spurious detection. Retinal vasculature exhibits multi-scale structural complexity, with vessels demonstrating: (1) fractal-like branching patterns spanning 6-8 orders of magnitude in diameter, (2) anisotropic orientation distribution, and (3) dense spatial packing in the capillary plexus where vessels approach the imaging resolution limit (typically 1-2 pixels wide). These characteristics impose fundamental challenges for precise boundary delineation, as conventional segmentation approaches struggle to concurrently resolve micron-scale capillaries while maintaining structural continuity.

Current methodologies exhibit three principal limitations: First, insufficient cross-domain feature fusion fails to synergistically integrate spatial-textural characteristics with frequency-domain representations (e.g., wavelet coefficients or Fourier harmonics), thereby neglecting complementary vascular signatures. Second, the prevalent use of monolithic feature extractors inadequately addresses the scale discrepancy between major vessels and capillaries. Third, most frameworks lack explicit mechanisms to preserve vascular connectivity in low-SNR regions, leading to compromised microvascular network integrity.

To address these challenges, we attempt to design a dual-branch network model operating synergistically in both spatial and frequency domains, named WHANet. In summary, our contributions are three folds:

(1) We design a wavelet and hybrid attention network for vessel segmentation in OCTA fundus images (WHANet), which comprises a hybrid attention deep convolutional branch (HADCB) and a multi-scale wavelet feature fusion branch (MWFFB). With the fusion of spatial-domain attention and frequency-domain wavelet decomposition, WHANet significantly captures the effective features of complex vessel.

(2) We propose a hybrid attention deep convolutional branch (HADCB), which integrates multi-scale depthwise separable convolution with a channel-spatial attention mechanism. It extracts the features of vessel edge through dynamic weight allocation, meanwhile effectively suppressing noise and artifact interference.

(3) We design a multi-scale wavelet feature fusion branch (MWFFB), which utilizes Discrete Wavelet Transform (DWT). This module decomposes the samples into high- and low-frequency components. A cross-frequency fusion strategy is employed by combining high-frequency feature enhancement with low-frequency guidance to denoise the image and accurately capture the frequency-specific characteristics of deep microvessels.

2 Related work

Retinal vessel segmentation plays a crucial role in ophthalmic diagnosis, particularly in optical coherence tomography angiography (OCTA), where the task is especially challenging due to image noise, low contrast, and the intricate structure of fine vessels. Existing approaches can be divided into two categories: statistics-based methods and deep learning-based methods.

Early studies primarily relied on hand-crafted features and morphological operations. For instance, the Frangi filter [29] enhances tubular structures using Hessian matrices, but it is highly sensitive to noise and struggles with accurately detecting vessel bifurcations. Region-growing methods [30] iteratively expand vascular regions based on pixel similarity; however, they are vulnerable to uneven illumination. Threshold-based segmentation techniques, such as OTSU algorithm [31], are simple and computationally efficient, but often fail to distinguish microvessels from background noise. In summary, statistics-based methods depend heavily on manual features, and are inadequate for handling the complex and fine scale vascular structures in OCTA images.

Recently, U-Net [32] has emerged as a benchmark model for medical image segmentation. Its encoder–decoder architecture effectively fuses multi-scale features via skip connections. To extract effective features from significant regions, some U-Net-based methods are presented with attention mechanism. Attention U-Net [21] incorporates a channel-spatial attention mechanism to dynamically focus on vascular regions. DBU-Net [33] integrates DenseNet dense connectivity mechanism into U-Net to alleviate the vanishing gradient problem, strengthen feature propagation. In this model, feature maps within each dense block are interconnected through cross-layer connections, effectively mitigating the vanishing gradient problem and improving the representation. DBU-Net performs well in segmenting complex vascular bifurcations, however, its dense connections significantly increase memory consumption, and impose high hardware demands. Gu [34] introduced CENet with a Context-Aware Module that combines global contextual information with local edge details. It also employs an edge-guided loss function to extract accurate boundary of vessels. While CENet achieves better vessel continuity and performs well in low-contrast regions. The edge-guided module requires more annotations, which increases labeling costs. GAO et al. [35] recently proposed a spatio-temporal correspondence attention network for solving the discontinuity and loss problem in vessel segmentation. The network employs an innovative encoder–decoder structure that contains a spatio-temporal correspondence block and two attention blocks. The spatio-temporal correspondence block extracts spatio-temporal features from the previous frame to enhance the feature representation of the current frame, while the spatial and channel attention blocks enhance the segmentation of foreground vessels.

Multi-scale fusion strategies have been extensively designed for image segmentation. For example, the Atrous Spatial Pyramid Pooling (ASPP) module proposed in DeepLabv3+ [36], which significantly improves spatial feature representation by leveraging multi-rate convolutions to capture multi-scale objects and their contextual dependencies. There exists plenty of effective information in frequency domains. The recent researches based on deep learning methods have further investigated cross-domain mechanisms that jointly optimal representations in both the spatial and frequency domains. For example, FreqMamba [37]introduces bandwidth analysis to exploit frequency correlations, and integrates the Fourier Transform to model global degradation to learn effective features from images.

Furthermore, to balance global context modeling with the preservation of local details, TransUNet [25]was the first to incorporate Transformers into medical image segmentation. By combining the strengths of Transformer architectures with the

encoder–decoder structure of U-Net, TransUNet provides a powerful framework for medical segmentation tasks. Specifically, it takes the CNN-derived feature maps as sequences of image patches and encodes them by using a Transformer to capture long-range dependencies. Simultaneously, the decoder upsamples the encoded features and fuses them with high-resolution CNN features to achieve precise localization.

Transformer architectures, driven by self-attention mechanisms, have demonstrated remarkable advantages in ocular image segmentation tasks. MsTGANet [38], a novel multi-scale Transformer-based global attention network with a U-shaped architecture, introduces a multi-scale Transformer non-local (MsTNL) module, strategically placed at the top of the encoder path. This module is specifically designed to extract multi-scale non-local features with long-range dependencies across different encoder layers, thereby enhancing global feature perception. In addition, Tan et al. [39] designed OCT2Former, which is a dynamic token aggregation Transformer and leverages a multi-head dynamic token aggregation attention mechanism to effectively capture global retinal vascular structures while significantly reducing computational costs in both time and space. To address the inherent lack of inductive bias in Transformer architectures, OCT2Former incorporates an auxiliary convolutional branch, which not only compensates for this limitation but also facilitates faster convergence with negligible increase in model parameters. Luo et al. [40]proposed a novel component, Lightweight Parallel Transformer (LPT), which addresses the shortcomings of the standard Transformer that is highly dependent on large datasets and computational resources, and captures long-distance dependencies and prevents fine-vessel breaks.

3 Methods

In this work, we propose a novel Wavelet and Hybrid Attention Network for Vessel Segmentation in OCTA fundus images, named WHANet, which is based on spatial-frequency-domain joint optimization. Fig. 2 illustrates the architecture of WHANet. It consists of two branches: the Hybrid Attention Depthwise Convolutional Branch (HADCB) and the Multi-scale Wavelet Feature Fusion Branch (MWFFB).

Firstly, the original image is fed into the backbone U-Net model, which can effectively learn image features with different scale spatial resolution through its encoder–decoder architecture. These preliminary feature maps are then fed into two specialized branches for further optimization.

The hybrid attention deep convolutional branch (HADCB) can mine the salient features with different channels and captures multi-scale spatial information from images. Meanwhile, the Multi-scale Wavelet Feature Fusion Branch (MWFFB) can decompose the image into multiple frequency sub-bands, which can extract fine-grained details at frequency levels. A multi-scale feature fusion strategy is employed to integrate information with different resolutions. This approach facilitates the fusion of low-level, high-resolution features with high-level, semantically rich representations.

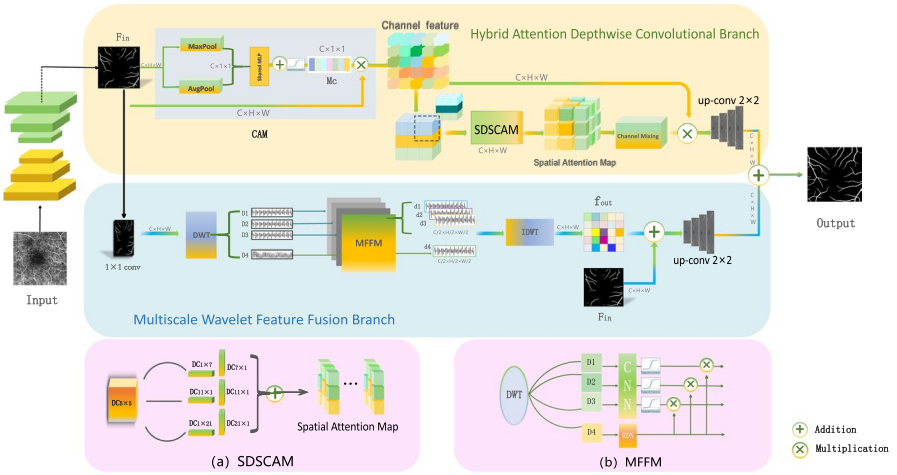

Fig. 2 The architecture of WHANet

Finally, the outputs from both branches are integrated using an element-wise addition to produce the final segmentation. This fusion can preserve the details of tiny vessels to some extent.

3.1 Hybrid attention deep convolutional branch

To capture salient features while suppressing irrelevant information of complex vessel structures under varying noise conditions, we propose a hybrid attention deep convolutional branch (HADCB). This branch can integrate a multi-level attention mechanism with depthwise convolution. The HADCB comprises two key components: a Channel Attention Module (CAM) and a Spatial Depth-Separable Convolutional Attention Module (SDSCAM), which work together to selectively effective features with both channel and spatial dimensions.

3.1.1 Channel attention module

The Channel Attention Module (CAM) is similar to the Convolutional Block Attention Module (CBAM). Given the feature map F_{in}, which is the preliminary segmentation output from U-Net, the module first generates two distinct spatial context descriptors with Global Average Pooling (GAP) and Global Max Pooling (GMP). These pooled features are then passed through a shared Multi-Layer Perceptron (MLP) for nonlinear transformation.

The outputs of the MLP are subsequently combined through element-wise summation and activated using a Sigmoid function to generate the channel attention map M_c. Finally, this attention map is multiplied element-wise with the original input feature map F_{in} to produce the refined channel-weighted feature map F_c.

This process can be formally expressed as:

$$M_c = \sigma\left(\mathrm{MLP}\left(\mathrm{GAP}\left(F_{\mathrm{in}}\right)\right) + \mathrm{MLP}\left(\mathrm{GMP}\left(F_{\mathrm{in}}\right)\right)\right). \tag{1}$$

$$F_c = M_c \otimes F_{\mathrm{in}}. \tag{2}$$

where $F_{\mathrm{in}} \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map. GAP(·) and GMP(·) denote the global average pooling and global maximum pooling operations, respectively. MLP(·) is the shared multilayer perceptron. $\sigma(\cdot)$ denotes the Sigmoid activation function, which is used to generate the channel attention graph. M_c is the computed channel attention weighted graph. F_c denotes the computed weighted feature map. $\otimes$ represents element-by-element multiplication.

3.1.2 Spatial depth separation convolutional attention module

We construct the spatial attention map based on the principle of avoiding rigid consistency constraints across channels, instead adopting a strategy that dynamically allocates attention weights across both channel and spatial dimensions. Fig. 2a illustrates the Spatial Depth-Separable Convolutional Attention Module (SDSCAM).

In this module, depthwise separable convolutions are to capture spatial features at multiple scales with kernels of varying sizes (e.g., 5×5, 1×7, 7×1, 1 × 11, 11 × 1, 1 × 21, 21 × 1.). The DC5×5 convolution, with a kernel size of 5×5, covers a relatively small local region and is primarily used to capture fine-grained spatial features, making it suitable for extracting detailed information at smaller scales. The DC1×7 and DC7×1 convolutions are designed to capture elongated features in the horizontal and vertical directions, respectively. The kernel with 1×7 is used to detect horizontal edges and textures, while the kernel with 7×1 is designed for vertical edge and texture extraction. These two kernels, with extended coverage in specific directions, help capture strongly directional features. The DC1×11, DC1×21, DC11×1, and DC21×1 convolutions further enlarge the receptive field, enabling the extraction of longer-range horizontal and vertical features, and are suitable for a broader range of structural and textural patterns. Larger kernels are more effective at capturing features of wide vessels, while smaller kernels are better at focusing on fine vessels and edge details. This multi-scale design effectively addresses the complex topology of retinal vasculature while mitigating the information loss that may occur with single-scale convolutions. These multi-scale operations are crucial for detecting blood vessels of different sizes and orientations. Notably, each input channel undergoes convolution independently, preserving inter-channel independence while allowing for the specialized extraction of channel-specific features.

Compared to the traditional convolutional operations, depthwise separable convolution offers a significant reduction in computational cost, making the model more efficient and better suited for deployment in resource-constrained environments. At the end of the SDSCAM, a channel-mixing operation is performed using a 1×1 convolution to generate the refined spatial attention map M_s. This process can be formally represented as:

$$M_s = W_{1\times1}\left(N\left(\sum_{i=0}^{3}\lambda_i \cdot g\left(\beta_i\left(DC\left(F_c\right)\right)\right)\right)\right). \tag{3}$$

$$F_s = M_s \otimes F_c. \tag{4}$$

where, $W_{1\times1}$ denotes the 1×1 convolution operation, which can perform the linear transformation. $N(\cdot)$ represents a normalization operation, Layer Normalization. λ_i is a learnable weight parameter to adjust the contribution of each branch. $g(\cdot)$ denotes the nonlinear activation function, ReLU$(\cdot)$. $\beta_i(\cdot)$ represents the computation process of the i^{th} branch. $DC(\cdot)$ denotes Depthwise Convolution. F_c is the feature map processed by the channel attention mechanism as input to the spatial depth-separated convolution attention module. M_s represents the computed spatial depth-separated convolutional attention weight map. $\otimes$ denotes element-by-element multiplication.

The traditional spatial attention mechanisms typically learn features at a single scale. While the Spatial Depth-Separable Convolutional Attention Module (SDSCAM) utilizes the convolutional kernels of multiple sizes (e.g., 5×5, 1×7, 7×1, 1×11, 11×1, 1×21, 21×1.) to capture multi-scale spatial features. This module can effectively identify vascular structures of diverse sizes and orientations in the image. In addition, depthwise separable convolution significantly reduces the number of parameters and computational complexity by decomposing standard convolution into depthwise convolution and pointwise (1×1) convolution, thereby greatly improving the model's operational efficiency. Meanwhile, the SDSCAM module introduces a dynamic feature fusion mechanism, which is a more flexible strategy compared to traditional spatial attention mechanisms. This mechanism dynamically assigns attention weights to different regions of the input feature map, enabling more precise capture of the complex anatomical structures in OCTA images. By processing multi-scale feature representations and dynamically modulating feature fusion, this module significantly enhances the model's adaptability to different vessel sizes and types, making it more robust against noise and variations in imaging conditions. Therefore, it provides strong support for the performance stability and reliability of OCTA retinal vessel segmentation tasks.

3.2 Multi-scale wavelet feature fusion branch

To learn effective features across different frequency bands from fine vessels and maintain the global morphology and continuity of the vessels, we design a Multi-scale Wavelet Feature Fusion Branch (MWFFB). MWFFB consists of three key components: the Discrete Wavelet Transform module (DWT), the multi-frequency feature fusion module (MFFM), and the Inverse Discrete Wavelet Transform module (IDWT).

3.2.1 Discrete wavelet transform module

The Discrete Wavelet Transform (DWT) is an effective technique for feature extraction in vessel image analysis. In this section, we design the Discrete Wavelet Transform (DWT) module to learn robust features of vessel, which can perform

multi-scale and multi-resolution decomposition. By decomposing an image into low-frequency (approximation) and high-frequency (detail) sub-bands, DWT captures both the global structure and fine details of vascular patterns. This is particularly beneficial for distinguishing vessels of varying sizes and enhancing contrast between vessels and the background.

The feature map F_{in} is first decomposed into multi-scale feature by the Discrete Wavelet Transform (DWT), which generates high-frequency (detail) components D_1, D_2, D_3 and a low-frequency (approximation) component D_4. This decomposition leverages the multi-resolution analysis capability of the wavelet transform, which enables the examination of signals at various scales. Such analysis facilitates the capture of vascular structures of varying sizes and orientations, thereby providing rich and complementary information for subsequent feature fusion. This process can be formally expressed as:

(1) The low-frequency feature D_4 can be expressed as:

$$D_4(x) = \sum_k c_{j,k}\phi_{j,k}(x). \tag{5}$$

where, $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ is scale function, which is used to capture the low-frequency features of the image. j is the decomposition scale (larger means rougher). k is the translation parameter, which controls the spatial position of the basis function. $c_{j,k}$ is the Scale coefficients, which reflects the low-frequency energy distribution of the input image at the scale j and position k.

(2) The high-frequency feature can be expressed as:

$$D_j(x) = \sum_k d_{j,k}\psi_{j,k}(x). \tag{6}$$

where, $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ is the Wavelet function, which is used to capture the high-frequency features of the image. $d_{j,k}$ denotes the Wavelet coefficient, which reflects the high-frequency energy distribution of the input image at scale j and position k.

3.2.2 Multi-frequency feature fusion module

As illustrated in Fig. 2b, we design a frequency-based hierarchical feature enhancement architecture for retinal vessel analysis. The input image initially undergoes four-channel discrete wavelet decomposition (DWT), generating three directional high-frequency sub-bands (D_1, D_2, D_3 corresponding to horizontal, vertical, and diagonal orientations) and one low-frequency approximation component (D_4).

For high-frequency processing, we developed a multi-stage convolutional neural network with directional sensitivity. A cascade of deformable convolutional layers with adaptive receptive fields is employed to capture long-range spatial dependencies along vascular structures. Subsequently, a Sigmoid gating mechanism dynamically

integrates these multi-frequency features through learnable channel-wise attention weights, effectively enhancing edge responses for fine vascular branches and complex topological patterns.

The low-frequency components are processed by a residual dense network (RDN) composed of stacked residual dense blocks to obtain d_4. This architecture facilitates deep feature learning through local feature fusion and global residual connections, thereby constructing semantically rich representations that preserve structural integrity of main vessels. The proposed multi-scale fusion strategy enables cross-frequency feature interaction through hierarchical concatenation and adaptive weighting. This synergistic combination achieves coordinated optimization between global vascular distribution patterns and local textural details, ultimately improving topological completeness.

Subsequently, the low-frequency feature map is learned by element-wise multiplied with each of the high-frequency feature maps to produce the fused feature maps d_1, d_2, d_3. This fusion process preserves global structural information while emphasizing fine-grained vascular details. The process can be formally described as:

$$c_j = \sigma\Big(\mathrm{CNN}_{\theta_j}(D_j)\Big), (j = 1, 2, 3). \tag{7}$$

$$d_4 = \big(\mathrm{RDN}(D_4)\big). \tag{8}$$

$$d_j = d_4 \odot c_j, (j = 1, 2, 3). \tag{9}$$

$$\mathrm{Output} = \left\{ d_4 \odot \sigma\Big(\mathrm{CNN}_{\theta_j}(D_j)\Big) | j = 1, 2, 3 \right\} \cup \{d_4\}. \tag{10}$$

$$\mathrm{Output} = \{d_1, d_2, d_3, d_4\}. \tag{11}$$

where, $\mathrm{CNN}_{\theta_j}(\cdot)$ denotes Convolutional Neural Network for high-frequency features D_j, θ_j as its parameters. $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid function for enhancing the nonlinear representation of features. RDN $(\dot{})$ refers to a residual dense network composed of stacked residual dense blocks.d_j denotes the high-frequency vector processed by CNN and Sigmoid function. $\odot$ denotes element-by-element multiplication operation. $\cup$ denotes the concatenation operation of sets.

3.2.3 Inverse wavelet discrete module

The Inverse Discrete Wavelet Transform (IDWT) is the process of reconstructing a signal or image that has been previously decomposed using the Discrete Wavelet Transform (DWT). In this step, the wavelet coefficients are recombined to recover the original signal or image. IDWT enables the reconstruction of the image while

allowing for the selective removal of noise or irrelevant details. In our framework, the fused feature maps are transformed back into the original image space through the Inverse Wavelet Transform Module, yielding the final segmentation result. This process leverages the reversibility of the wavelet transform, enabling accurate reconstruction of the image from the decomposed feature maps. The reconstruction process can be formally expressed as:

$$A(x) = \sum_{k} d_{4,k} \phi_{J,k}(x). \tag{12}$$

$$D_j(x) = \sum_{k} c_{j,k} \psi_{j,k}(x), \quad j = 1, 2, 3. \tag{13}$$

$$f_{\text{out}}(x) = A(x) + \sum_{j=1}^{3} D_j(x). \tag{14}$$

where, $A(x)$ is the low-frequency signal obtained. $d_{4,k}$ is the coefficient of the low-frequency feature map d_4, which represents the low-frequency component of the signal at scale J. $\phi_{j,k}(x) = 2^{J/2}\phi(2^J x - k)$ is the basis function of the scale function at scale J. J is the maximum decomposition scale, which represents the lowest frequency resolution level. $c_{j,k}$ is the coefficient of the high frequency feature map c_j, which represents the high-frequency component of the signal at scale j. $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ is the basis function of the wavelet function at scale j. j is the scale of the high-frequency feature map. Parameters $j = 1, 2, 3$ denote the high-frequency components at different scales, respectively. $f_{\text{out}}(x)$ is the complete reconstructed signal obtained. x is the pixel position of the image, which is used to describe the distribution of the image in the spatial domain.

Traditional wavelet transforms may present certain limitations when applied to retinal blood vessel images, such as increased sensitivity to noise and insufficient consideration of the interrelationships between different features during the fusion process. The multi-scale wavelet feature fusion branch (MWFFB) proposed in this study effectively addresses these limitations. By incorporating convolutional neural networks (CNNs) and sigmoid functions to enhance high-frequency features, and employing element-wise multiplication for feature fusion, the proposed method significantly improves the model's performance in retinal blood vessel image segmentation.

3.3 Overview of algorithm

Algorithm 1 Training Process of WHANet

Require:
Input: OCTA image training set $\mathcal{D}$, batch size B, learning rate η, maximum number of training rounds T
Ensure:
Output: Parameters of the trained model Θ_{WHANet}
1. Initialisation Parameters
- Parameters of the skeleton network (UNet): Θ_{UNet}
- Hybrid Attention Deep Convolutional Branch (HADCB) parameters: $\Theta_{\text{HADCB}}, \Theta_{\text{CAM}}, \Theta_{\text{SDSCAM}}, \Theta_{\text{Mixing}}$
- Multi-scale Wavelet Feature Fusion Branch (MWFFB) parameters: $\Theta_{\text{MWFFB}}, \Theta_{\text{DWT}}, \Theta_{\text{MFFM}}, \Theta_{\text{IDWT}}$
- Optimiser: Adam (initial learning rate η, weight decay 0.001)

2. Iterative training ($t = 1$ to T):
for each training epoch t **do**
  for each batch of input images $X \in \mathbb{R}^{B\times H\times W\times C}$ **do**
    a. Forward propagation:
      Step 1: Backbone Feature Extraction: $F_{\text{UNet}} \leftarrow \text{UNet}(X; \Theta_{\text{UNet}})$
      Step 2: Hybrid Attention Deep Convolutional Branch (HADCB):
        - Generate Channel Attention Feature Map: $M_c \leftarrow \text{CAM}(F_{\text{UNet}}; \Theta_{\text{CAM}})$
        - Generate Spatial Attention Weight Map: $M_s \leftarrow \text{SDSCAM}(F_c; \Theta_{\text{SDSCAM}})$
        - Feature Refinement, Generate Refined Feature Map: $F_s \leftarrow F_c \otimes \text{Channel Mixing}(M_s; \Theta_{\text{Mixing}})$
      Step 3: Multi-scale Wavelet Feature Fusion Branch (MWFFB):
        - Multi-scale Decomposition: $D_1, D_2, D_3, D_4 \leftarrow \text{DWT}(\text{Conv}_{2\times2}(F_{\text{UNet}}); \Theta_{\text{DWT}})$
        - Feature enhancement processing with feature fusion: $d_1, d_2, d_3, d_4 \leftarrow \text{MFFM}(D_1, D_2, D_3, D_4; \Theta_{\text{MFFM}})$
        - The fused feature map is converted back to the original image space: $F_{\text{wavelet}} \leftarrow \text{IDWT}(d_1, d_2, d_3, d_4; \Theta_{\text{IDWT}}), F_f = F_{\text{UNet}} + F_{\text{wavelet}}$
      Step 4: Dual-branch feature aggregation: $out = \text{Upsample}(F_s) + \text{Upsample}(F_f)$
      return *out*
    b. Loss Calculation:
      - Dice loss: $\mathcal{L}_{\text{Dice}} = 1 - \frac{2|Y_{\text{pred}} \cap Y_{\text{gt}}|}{|Y_{\text{pred}}| + |Y_{\text{gt}}|}$
      - Total loss: $\mathcal{L} = \mathcal{L}_{\text{Dice}}$
    c. Backpropagation with parameter update:
      - Calculating the gradient: $\nabla\Theta = \frac{\partial \mathcal{L}}{\partial \Theta}$
      - Updating parameter: $\Theta \leftarrow \Theta - \eta \cdot \nabla\Theta$
  end for
end for
3. Output: parameters of the trained model $\Theta_{\text{WHANet}} = \{\Theta_{\text{UNet}}, \Theta_{\text{HADCB}}, \Theta_{\text{MWFFB}}\}$

Algorithm 2 The Testing Process for WHANet

Require:
Unseen OCTA test image $X \in \mathbb{R}^{H \times W \times C}$
Trained model parameters $\Theta_{\text{WHANet}} = \{\Theta_{\text{UNet}}, \Theta_{\text{HADCB}}, \Theta_{\text{MWFFB}}\}$
(Optional) Preprocessing parameters (μ, σ)
Ensure:
Predicted vessel segmentation mask $\hat{Y} \in [0,1]^{H \times W}$
Preprocessing:
$X_{\text{normalized}} \leftarrow (X - \mu)/\sigma$ ▷ Normalize with dataset statistics
1. Backbone Feature Extraction:
$F_{\text{UNet}} \leftarrow \text{UNet}(X_{\text{normalized}}; \Theta_{\text{UNet}})$
2. Hybrid Attention Deep Convolutional Branch (HADCB):
Channel Attention Module:
$M_C \leftarrow \sigma\Big(\text{MLP}\big(\text{GAP}(F_{\text{UNet}})\big) + \text{MLP}\big(\text{GMP}(F_{\text{UNet}})\big)\Big)$ ▷ Eq.1
$F_C \leftarrow M_C \otimes F_{\text{UNet}}$ ▷ Channel-refined features (Eq.2)
Spatial Depth-Separable Conv Attention Module:
$M_S \leftarrow \text{ChannelMixing}\Big(\sum_{i=1}^{7} \lambda_i \cdot \text{ReLU}\big(\text{DWConv}_i(F_C)\big)\Big)$ ▷ Multi-scale DW conv (Eq.3)
$F_S \leftarrow M_S \otimes F_C$ ▷ Spatial-refined features (Eq.4)
3. Multi-scale Wavelet Feature Fusion Branch (MWFFB):
Discrete Wavelet Transform:
$F_{\text{conv}} \leftarrow \text{Conv}_{1\times1}(F_{\text{UNet}})$ ▷ Channel adjustment
$\{D_1, D_2, D_3, D_4\} \leftarrow \text{DWT}(F_{\text{conv}})$ ▷ High/low-frequency decomposition
Multi-frequency Feature Fusion:
$d_j \leftarrow \sigma\big(\text{CNN}_{\theta_j}(D_j)\big), \quad j = 1,2,3$ ▷ High-freq (Eq.7)
$d_4 \leftarrow \text{CNN}_{\theta_4}(D_4)$ ▷ Low-freq
$c_j \leftarrow d_4 \odot d_j, \quad j = 1,2,3$ ▷ Cross-frequency fusion (Eq.8-10)
Inverse DWT:
$F_{\text{wavelet}} \leftarrow \text{IDWT}(c_1, c_2, c_3, d_4)$ ▷ Reconstruct (Eq.11-13)
$F_f \leftarrow F_{\text{UNet}} + F_{\text{wavelet}}$ ▷ Skip connection
4. Dual-Branch Feature Aggregation:
if $\text{size}(F_S) \neq \text{size}(F_f)$ **then**
  $F_S \leftarrow \text{BilinearInterpolate}(F_S, \text{size}(F_f))$
end if
$\hat{Y}_{\text{raw}} \leftarrow F_S + F_f$ ▷ Element-wise sum
5. Postprocessing:
$\hat{Y} \leftarrow \sigma\big(\text{Conv}_{1\times1}(\hat{Y}_{\text{raw}})\big)$ ▷ Sigmoid activation
$\hat{Y}_{\text{binary}} \leftarrow \mathbb{I}(\hat{Y} > 0.5)$ ▷ Binary thresholding
return $\hat{Y}_{\text{binary}}$

Algorithm 1 concludes the training procedure of WHANet. It provides a comprehensive overview of the entire training process, including parameter initialization, iterative training, and updating parameters. The algorithm details the operations of both the hybrid attention deep convolutional branch (HADCB) and the multi-scale wavelet feature fusion branch (MWFFB), as well as the processes involved in loss computation and model optimization.

Algorithm 2 summarizes the inference procedure of the WHANet model. This process encompasses all critical steps, including input image normalization, backbone feature extraction, branch-wise feature enhancement, feature fusion, and final segmentation prediction.

4 Experiments

4.1 Datasets

In this study, there exist three OCTA datasets, including the OCTA-RV, OCTA-500-3 M, and OCTA-500-6 M datasets. Each dataset consists of with different imaging scales and subjects, which are detailed as follows.

- OCTA-RV: OCTA-RV was collected using a Swept-Source Optical Coherence Tomography (SS-OCT) system and comprises retinal vascular images from 62 distinct subjects, including both diabetic retinopathy patients and healthy controls. Each subject contributed one original OCTA image, covering a 12×12 mm^2 field of view with a resolution of 304×304 pixels, providing detailed visualization of the retinal microvasculature. The dataset contains a diverse spectrum of retinal vascular pathologies, such as choroidal neovascularization (CNV), microaneurysms (MA), retinal non-perfusion (RNP), and intraretinal microvascular abnormalities (IMA). To improve model generalization, standard data augmentation techniques were applied to the original images, resulting in a total of 372 images. Importantly, we employed a subject-level splitting strategy, ensuring that all images (including augmented versions) from the same subject were assigned exclusively to either the training or testing set, thus effectively preventing data leakage. Moreover, the OCTA-RV dataset is completely independent from the OCTA-500 datasets in terms of subject source, with no overlap, guaranteeing the scientific rigor and fairness of cross-dataset evaluations.
- OCTA-500-3 M: OCTA-500-3 M is a publicly available subset of the OCTA-500 dataset, focusing on the central retinal region with a 3×3 mm^2 field of view. This subset contains 200 OCTA images, each considered an independent sample. Most images were acquired from healthy subjects, while a portion were obtained from patients diagnosed with common retinal diseases, including DR, age-related macular degeneration (AMD), and retinal vein occlusion (RVO). To ensure consistency with prior studies utilizing the OCTA-500 dataset, a random image-level splitting strategy was employed, in which images were randomly assigned to the training, validation, and testing sets. Ensure that one image does not appear in the same subset.
- OCTA-500-6 M: OCTA-500-6 M complements the OCTA-500-3 M subset to complete the full OCTA-500 dataset. It focuses on a broader retinal area with a 6×6 mm^2 field of view. This subset contains 300 OCTA images, which are collected from patients and diagnosed with retinal diseases such as DR, AMD, and RVO. Consistent with OCTA-500-3 M, an image-level random partitioning method was used to divide the data into training, validation, and testing sets, ensuring no image overlap across subsets.

4.2 Evaluation metrics

In order to comprehensively and objectively evaluate the segmentation performance of the proposed method as well as the comparative methods, four evaluation indexes, namely, Sensitivity (SEN), Specificity (SPE), Dice Coefficient (DICE), and Jaccard Index (JAC), are calculated.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{16}$$

$$\text{Dice Coefficient} = 2 \times \frac{\text{TP}}{(\text{FP} + \text{FN} + 2 \times \text{TP})}, \tag{17}$$

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \tag{18}$$

where, *TP* stands for True Positive indicates the number of pixels correctly identified as blood vessels. *FP* denotes False Positive indicates the number of pixels incorrectly identified as blood vessels. *TN* stands for True Negative indicates the number of pixels correctly identified as non-vascular. *FN* represents False Negative indicates the number of pixels incorrectly identified as non-vascular.

4.3 Experimental setup and parameter configuration

All methods were implemented on the PyTorch platform. To comprehensively evaluate model performance, the dataset was randomly divided into training, testing, and validation sets with a ratio of 8:1:1. For model optimization, we adopted the Dice loss as the objective function and used the Adam optimizer for parameter updates. The initial learning rate was set to 0.0005, and the weight decay was set to 0.001 to balance training efficiency and generalization capability. The batch size was set to 8. The learning rate was decayed by a factor of 0.1 every 50 epochs to fine-tune the model and mitigate overfitting. To enhance data diversity and model robustness to variations in vessel orientation and shape, data augmentation techniques were applied to the training set, including random rotations (within $\pm 15^o$), random horizontal and vertical flips, and elastic deformations. All experiments were conducted on a workstation with an NVIDIA RTX A4000 GPU (16GB VRAM), running Windows 11 and Python 3.8. To ensure sufficient convergence, the epochs of training were set to 200. For fair comparison, five-fold cross-validation was employed for all methods. Each fold was independently trained and validated, and the final results were reported as the average performance across the five folds. All baseline models were trained under the same experimental settings, including optimizer configuration, batch size, number of epochs, and data augmentation strategies, to ensure a fair and consistent evaluation.

4.4 Experimental results

Figure 3 presents a qualitative comparison of the segmentation results obtained by WHANet and other representative methods on three datasets: OCTA-RV, OCTA500-3 M, and OCTA500-6 M. For each dataset, the first row displays the full-field segmentation outputs, while the second row provides magnified views of the regions of interest (ROIs), highlighted by red rectangles. These ROIs allow for a detailed inspection of model behavior in handling vessel bifurcations, fine capillaries, and low-contrast regions.

The classical U-Net, as a typical encoder–decoder architecture, has been widely adopted in medical image segmentation due to its ability to fuse low-level spatial details and high-level semantic features via skip connections. However, as illustrated in Fig. 3, U-Net tends to exhibit topological discontinuities and false-positive predictions (marked by blue dashed circles), particularly when dealing with complex vascular bifurcations and low-contrast microvessels. These issues are mainly attributed to its limited capacity in capturing multi-scale vascular morphologies and modeling contextual dependencies.

Attention U-Net, built upon the U-Net architecture, introduces attention gating mechanisms to enhance the response to vascular regions and suppress background noise. Although it improves performance to some extent, Fig. 3 reveals that this method may over-suppress weakly activated areas (e.g., capillaries with diameters less than 3 pixels), leading to missed detections (highlighted by yellow dashed

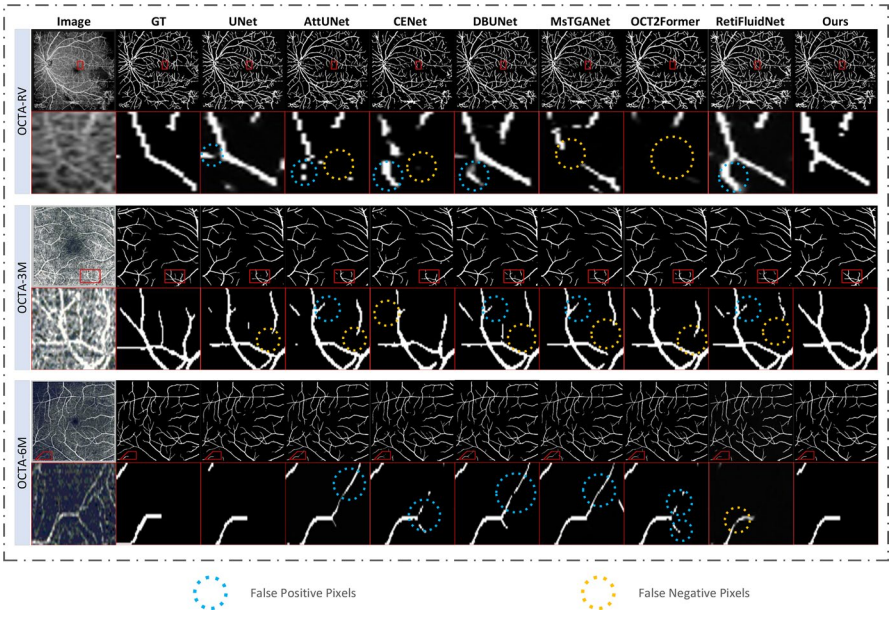

Fig. 3 The retinal vessel segmentation results of the proposed WHANet and compared methods. From top to bottom are the predicted segmentation results on OCTA500-3 M, OCTA500-6 M, and OCTA-RV datasets

circles). Furthermore, under pathological interference, such as hemorrhagic or exudative lesions, its attention distribution is easily distorted, resulting in false segmentation, especially around vascular intersections and edges.

OCT2Former, as a Transformer-based model, demonstrates promising performance in modeling global context. However, it shows suboptimal ability in preserving fine anatomical structures. In Fig. 3, OCT2Former frequently misses small vessels and exhibits discontinuities in vessel morphology across multiple datasets. This can be attributed to its heavy reliance on large-scale annotated data for effective learning. In practice, however, medical imaging datasets–especially OCTA images–are often limited in size and costly to annotate, which hampers convergence and weakens generalization performance.

Compared to Transformer-based models, the proposed WHANet is a dual-branch segmentation network architecture that integrates wavelet transform and hybrid attention mechanisms, inherently addressing the shortcomings of Transformer architectures in few-shot scenarios. Unlike Transformers, which lack inductive bias and heavily depend on large-scale annotated data, WHANet fully leverages the inherent inductive biases of convolutional operations and wavelet decomposition, thereby demonstrating stronger generalization ability under limited data conditions. WHANet consists of two key components: the hybrid attention deep convolutional branch (HADCB) and the multi-scale wavelet feature fusion branch (MWFFB). The HADCB enhances the model's sensitivity to fine vascular structures (e.g., edges and bifurcations) through multi-scale depthwise separable convolutions combined with both channel and spatial attention mechanisms, significantly improving the preservation of spatial details. On the other hand, the MWFFB introduces the Discrete Wavelet Transform (DWT) to decompose feature maps into multiple frequency bands, enabling the extraction of critical high-frequency vascular information while effectively suppressing noise. This joint optimization strategy in both spatial and frequency domains allows WHANet to maintain global structural consistency while more precisely capturing fine, grained local features–achieving a seamless integration of global semantic modeling and local detail enhancement. Moreover, compared to Transformer models relying on self-attention mechanisms, WHANet offers lower computational complexity and reduced hardware demands, making it more suitable for deployment in resource-constrained clinical environments.

Figure 3 demonstrates the superior segmentation performance of our WHANet on all three datasets. Notably, it significantly reduces both false positives and false negatives, as indicated by the blue and yellow dashed circles, respectively. WHANet can improve the accuracy of capillaries and bifurcation regions, maintaining the continuity and structural integrity of the vascular network, even in the more complex OCTA500-6 M dataset. The OCTA500-60 M dataset provides a broader perspective and greater pathological changes. WHANet is more effective in maintaining topological consistency than comparative methods, resulting in anatomically faithful vascular reconstruction. It is worth noting that due to its enhanced edge and microstructure representation ability, it exhibits strong performance in detail preservation, accurately depicting capillary and other fine scale vascular structures.

These advantages of our WHANet is that the novel dual-branch architecture and comprising the Hybrid Attention Deep Convolutional Branch (HADCB) and the

Multi-scale Wavelet Feature Fusion Branch (MWFFB). HADCB enhances the model's sensitivity to microstructures, such as edges and bifurcations, through multi-scale depthwise separable convolutions integrated with both channel and spatial attention mechanisms. Meanwhile, MWFFB leverages the Discrete Wavelet Transform (DWT) to decompose feature maps into multiple frequency bands. Through high-frequency feature enhancement and multi-frequency fusion, this branch improves edge representation and suppresses background noise, further boosting the model's precision, and robustness in retinal vessel segmentation.

4.5 Quantitative analysis

To validate the effectiveness of the proposed method, we conducted extensive experiments on three datasets: OCTA-RV, OCTA-500-3 M, and OCTA-500-6 M. Our approach was thoroughly compared with several state-of-the-art retinal vessel segmentation methods. Specifically, we evaluated seven deep learning-based models: U-Net, Attention U-Net, CENet, DBU-Net, MsTGANet, OCT2Former, and RetiFluidNet. To ensure a fair and rigorous evaluation, all comparative methods were carefully fine-tuned to achieve their optimal performance on the respective datasets. For each model, hyperparameters were manually adjusted to maximize segmentation accuracy. Experimental results demonstrate that our WHANet consistently outperforms existing approaches across multiple evaluation metrics in Tables 1, 2, and 3.

To evaluate whether the improvements are statistically significant, the DICE scores from all comparative experiments were subjected to the Wilcoxon signed-rank test. As shown in Tables 1, 2, and 3, all p values are below 0.05, indicating that our WHANet achieves significant improvements over other networks across all three datasets.

WHANet outperforms the baseline and state-of-the-art methods across the majority of evaluation metrics in Table 1. Compared to the latest OCTA vessel segmentation model, OCT2Former, WHANet achieves notable improvements in sensitivity (SEN), Dice coefficient (DICE), and Jaccard index (JAC), with increases of 10.22%, 1.7%, and 0.25%, respectively. Furthermore, when compared

Table 1 Results on the OCTA-RV dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC(%)	p-value
U-Net	72.11 ± 1.24	94.46 ± 0.71	70.41 ± 2.37	56.40 ± 3.32	p=0.003
Attention U-Net	78.86 ± 1.56	93.62 ± 0.30	72.41 ± 2.45	57.02 ± 2.47	p=0.038
CENet	66.83 ± 1.74	94.92 ± 0.23	67.47 ± 2.14	51.12 ± 2.73	p< 0.001
DBU-Net	68.53 ± 1.83	95.36 ± 0.67	69.90 ± 1.97	55.62 ± 3.12	p = 0.002
MsTGANet	72.10 ± 2.41	**95.53 ± 0.98**	72.16 ± 2.35	55.98 ± 2.34	p=0.031
OCT2Former	73.69 ± 1.59	95.56 ± 0.31	71.10 ± 2.51	57.22 ± 2.43	p=0.023
RetiFluidNet	79.17 ± 2.34	91.94 ± 1.21	69.10 ± 1.98	53.68 ± 2.78	p=0.002
Ours	**79.88 ± 2.74**	93.52 ± 0.79	**72.80 ± 1.78**	**57.47 ± 2.45**	–

Best results are in bold

Table 2 Results on the OCTA-500-3 M dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC(%)	*p* value
U-Net	85.26 ± 1.76	98.54 ± 0.72	81.89 ± 2.35	69.49 ± 2.01	$p< 0.001$
Attention U-Net	85.90 ± 1.79	99.09 ± 0.25	85.85 ± 1.91	75.37 ± 2.35	$p =0.003$
CENet	83.47 ± 2.49	99.18 ± 0.17	84.50 ± 2.13	74.04 ± 2.57	$p =0.002$
DBU-Net	86.57 ± 2.31	99.06 ± 0.29	86.06 ± 2.35	75.69 ± 2.33	$p =0.024$
MsTGANet	83.36 ± 2.46	99.36 ± 0.16	86.17 ± 1.84	75.85 ± 2.48	$p =0.032$
OCT2Former	83.16 ± 2.37	**99.49 ± 0.13**	86.10 ± 2.58	75.75 ± 1.84	$p =0.029$
RetiFluidNet	83.58 ± 1.86	99.38 ± 0.30	86.48 ± 2.21	76.31 ± 2.04	$p =0.041$
Ours	**87.27 ± 1.98**	99.08 ± 0.37	**86.56 ± 2.34**	**76.45 ± 2.12**	–

Best results are in bold

Table 3 Results on the OCTA-500-6 M dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC (%)	*p*-value
U-Net	87.73 ± 1.75	98.64 ± 0.56	86.51 ± 2.32	76.54 ± 1.67	$p =0.017$
Attention U-Net	86.30 ± 2.01	99.14 ± 0.26	88.28 ± 1.67	79.07 ± 2.34	$p=0.027$
CENet	85.22 ± 1.74	98.90 ± 0.56	86.41 ± 1.65	76.17 ± 1.79	$p =0.012$
DBU-Net	87.25 ± 1.55	99.01 ± 0.36	88.17 ± 2.07	78.90 ± 2.21	$p =0.031$
MsTGANet	85.56 ± 2.14	99.12 ± 0.36	87.70 ± 1.62	78.19 ± 2.03	$p =0.024$
OCT2Former	**88.80 ± 1.98**	98.72 ± 0.37	87.74 ± 1.39	78.21 ± 1.83	$p =0.022$
RetiFluidNet	87.48 ± 1.88	98.78 ± 0.48	87.02 ± 2.06	77.34 ± 1.67	$p =0.019$
Ours	86.62 ± 1.58	**99.15 ± 0.32**	**88.46 ± 1.78**	**79.42 ± 2.02**	–

Best results are in bold

with RetiFluidNet, WHANet shows improvements of 0.71% in SEN, 1.58% in specificity (SPE), 3.70% in DICE, and 3.79% in JAC.

Table 2 shows the segmentation results on the OCTA-500-3 M dataset. There exist a significant improvement for WHANet in key evaluation metrics, particularly in sensitivity (SEN), Dice coefficient (DICE), and Jaccard index (JAC), compared to previously proposed models. Specifically, compared with DBU-Net, WHANet achieves an increase of 3.69% in SEN, 0.08% in DICE, and 0.14% in JAC. When compared to the recently introduced OCT2Former model, WHANet shows even more substantial gains, with improvements of 4.11% in SEN, 0.46% in DICE, and 0.70% in JAC. Similarly, in comparison to RetiFluidNet, WHANet again outperforms it with the same margins: 3.69% in SEN, 0.08% in DICE, and 0.14% in JAC.

Table 3 shows the segmentation results on the OCTA-500-6 M dataset. One can see that WHANet obtains the significant improvements in key evaluation metrics, including specificity (SPE), Dice coefficient (DICE), and Jaccard index (JAC). Specifically, compared to the latest OCT2Former, WHANet achieves an improvement of 0.43% in SPE, 0.72% in DICE, and 1.21% in JAC. In comparison with RetiFluidNet, WHANet shows further improvement of 0.37% in SPE, 1.44%

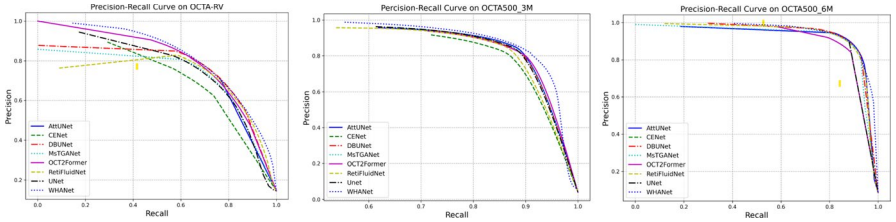

Fig. 4 Precision-Recall Curve for WHANet and the compared models. From left to right are results on OCTA500-3 M, OCTA500-6 M, and OCTA-RV datasets

Table 4 Model efficiency comparison in terms of Params and FLOPs

Model	Params(M)	FLOPs(G)
U-Net	7.29	10.62
Attention U-Net	34.88	103.99
CENet	10.13	12.36
DBU-Net	24.79	102.86
RetiFluidNet	8.56	25.22
MstGANet	11.60	27.24
OCT2Former	7.35	102.20
Ours (WHANet)	9.26	23.71

in DICE, and 2.08% in JAC. These comparative experimental results validate that the proposed WHANet framework is an effective solution for OCTA vessel segmentation, offering both high accuracy and robust performance across diverse datasets.

By comparing the Precision-Dice curves of different models across various datasets, the performance advantages and limitations of each model in segmentation tasks can be more intuitively and accurately assessed in Fig. 4.

4.5.1 Model performance comparison analysis

Besides segmentation accuracy, model efficiency is also a critical factor for clinical deployment, especially in scenarios under limited computational resources. To this end, we comprehensively evaluate the proposed WHANet from the perspectives of model complexity and computational cost, comparing it with several representative baseline methods. The analysis includes the number of parameters (Params, unit: million) and Floating-point Operations Per Seconds (FLOPs, unit: GFLOPs). The results are summarized in Table 4.

Despite WHANet's dual-branch architecture and the integration of both spatial and frequency-domain operations (including DWT and IDWT), it maintains a relatively low parameter count of 9.26M and a moderate computational cost of 23.71G FLOPs. This is significantly lower than several recent state-of-the-art models such as Attention U-Net (34.88M / 103.99G), DBU-Net (24.79M / 102.86G), and OCT-2Former (7.35M / 102.20G), all of which exhibit higher complexity.

Compared to the widely used U-Net baseline (7.29M / 10.62G), WHANet introduces a modest increase in model parameters (w.r.t Param), and computing expenditure (w.r.t Flops), yet achieves substantially better segmentation performance, as shown in Fig. 4. This demonstrates WHANet's strong performance-to-complexity ratio.

4.6 Ablation studies

4.6.1 Ablation study of the two branches in the WHANet model

To evaluate the effectiveness of each branch within the WHANet model, ablation studies were conducted on three datasets: OCTA-RV, OCTA-500-3 M, and OCTA-500-6 M. In these experiments, WHANet/up means the model without the Multi-scale Wavelet Feature Fusion Branch, retaining only the Hybrid Attention Deep Convolutional Branch. Conversely, WHANet/down indicates the removal of the Hybrid Attention Deep Convolutional Branch, preserving only the Multi-scale Wavelet Feature Fusion Branch. The segmentation results of the ablation experiments are presented in Tables 5, 6, and 7, respectively.

In Table 5, WHANet without the Multi-scale Wavelet Feature Fusion Branch from WHANet results in a performance decline of 2.29% in SEN, 1.23% in SPE,

Table 5 The segmentation results of ablation on the OCTA-RV dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC(%)
WHANet/up	77.59 ± 1.44	**94.54 ± 1.32**	68.22 ± 0.78	51.93 ± 1.45
WHANet/down	76.34 ± 1.65	92.47 ± 1.56	70.94 ± 0.97	55.98 ± 1.01
Ours	**79.88 ± 1.75**	93.52 ± 1.23	**72.80 ± 0.78**	**57.47 ± 1.34**

Best results are in bold

Table 6 The segmentation results of ablation on the OCTA-500-3 M dataset. Best results are in bold

Methods	SEN(%)	SPE (%)	DICE(%)	JAC (%)
WHANet/up	79.76 ± 1.47	97.85 ± 0.76	80.67 ± 1.23	67.91 ± 1.22
WHANet/down	85.15 ± 1.45	**99.26 ± 0.24**	84.68 ± 1.56	72.55 ± 1.52
Ours	**87.27 ± 0.78**	99.08 ± 0.32	**86.56 ± 1.32**	**76.45 ± 1.67**

Table 7 The segmentation results of ablation on the OCTA-500-6 M dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC (%)
WHANet/up	84.43 ± 1.34	97.05 ± 1.53	85.54 ± 0.98	76.88 ± 1.78
WHANet/down	**87.37 ± 1.45**	96.10 ± 1.45	87.31 ± 1.04	76.11 ± 1.87
Ours	86.62 ± 1.28	**99.15 ± 0.33**	**88.46 ± 1.23**	**79.42 ± 0.78**

Best results are in bold

5.89% in DICE, and 8.54% in JAC on the OCTA-RV dataset. Similarly, removing the Hybrid Attention Deep Convolutional Branch leads to decreases of 3.54% in SEN, 1.05% in SPE, 1.86% in DICE, and 1.49% in JAC.

From Table 6, it can be observed that on the OCTA-500-3 M dataset, WHANet without the Multi-scale Wavelet Feature Fusion Branch causes a significant drop in performance 7.51% in SEN, 1.23% in SPE, 5.89% in DICE, and 8.54% in JAC. Removing the hybrid attention deep convolutional branch results in a decrease of 2.12% in SEN, 1.88% in DICE, and 3.90% in JAC.

Table 7 shows the segmentation results on the OCTA-500-6 M dataset. One can observe that the removal of the multi-scale wavelet feature fusion branch leads to a reduction of 2.19% in SEN, 2.10% in SPE, 2.92% in DICE, and 2.54% in JAC. Removing the hybrid attention deep convolutional branch causes a decrease of 3.05% in SPE, 1.15% in DICE, and 3.31% in JAC.

4.6.2 Model complexity and efficiency analysis

To evaluate the contribution of each key component in WHANet to overall model efficiency, we conducted ablation experiments focusing on three critical aspects: parameter count (Params), floating-point operations (FLOPs), and inference speed (FPS). The results are presented in Fig. 5. One can see that the baseline U-Net model, owing to its relatively simple architecture, has the lowest parameter count (8.29M) and the lowest computational complexity (10.62G FLOPs), thereby achieving the fastest inference speed (49.2 FPS). This efficiency makes it well suited for real-time applications in resource-constrained environments. When the Hybrid Attention Deep Convolutional Branch (HADCB) is incorporated to form WHANet/up, the parameter count increases slightly to 8.62M, and the FLOPs rise to 21.43G. The results in a moderate reduction in inference speed to 39.7 FPS. While the additional computational cost is noticeable, the branch significantly enhances the model's ability to capture microvascular details, leading to improved segmentation performance (in Tables 5, 6 and 7). In contrast, WHANet/down which integrates the Multi-Scale Wavelet Feature Fusion Branch (MWFFB) exhibits the highest computational load among the configurations (19.84G FLOPs) and the slowest inference speed (4.1 FPS). This is indicative of the high resource demand associated with

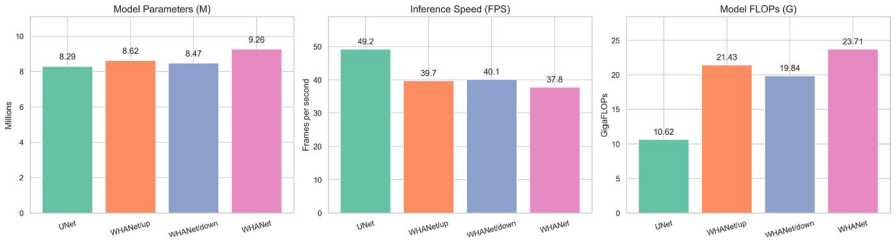

Fig. 5 Model performance evaluation. (*Model Parameters* denotes the number of parameters in the model, **Inference Speed** denotes the evaluation of average frame-per-second (FPS), and *Model FLOPs* denotes the computational cost)

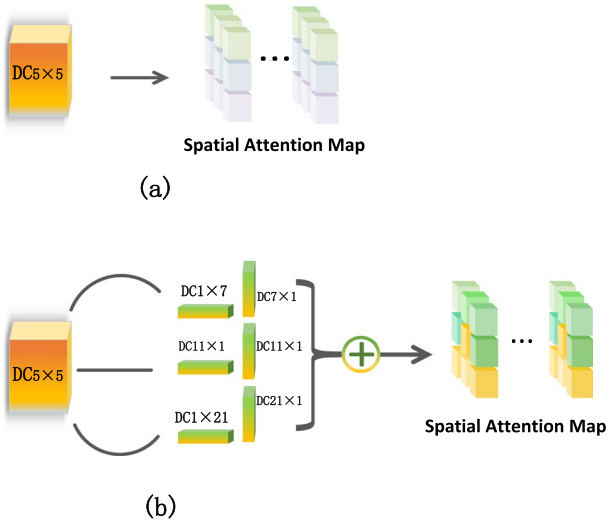

Fig. 6 Ablation of different convolutional kernels in SDSCAM

frequency-domain processing. However, it also empowers the model with stronger structural modeling capabilities in the frequency domain. Finally, the complete WHANet model, which combines both HADCB and MWFFB branches, has a total of 9.26M parameters and 23.71G FLOPs, with an inference speed of 37.8 FPS. Despite the increasing computational cost, it achieves the best performance across multiple evaluation metrics, demonstrating the effectiveness of the space-frequency co-optimization strategy in balancing segmentation accuracy and computational efficiency.

4.6.3 Ablation study of convolution kernel size

To systematically validate the effectiveness of the multi-scale depthwise separable convolution kernel in the Spatial Depth-Separable Convolutional Attention Module (SDSCAM) of our WHANet model, we conducted a series of ablation experiments focused on different kernel settings. These experiments were carried out on three representative public OCTA fundus datasets, i.e., OCTA-RV, OCTA500-3 M and OCTA500-6 M. The results are summarized in Tables 8, 9 and 10, respectively. The

Table 8 Ablation study of different convolution kernel sizes on the OCTA-RV dataset

Convolution kernel configuration	SEN (%)	SPE(%)	DICE (%)	JAC(%)
only5×5	69.78 ± 3.17	82.13 ± 2.79	61.83 ± 1.77	51.35 ± 2.13
Ours(WHANet)	**79.88 ± 1.75**	**93.52 ± 1.23**	**72.80 ± 0.78**	**57.47 ± 1.34**

Best results are in bold

Table 9 Ablation study of different convolution kernel sizes on the OCTA500-3 M dataset

Convolution kernel configuration	SEN (%)	SPE(%)	DICE(%)	JAC(%)
only5×5	56.74 ± 3.67	97.83 ± 0.63	73.10 ± 3.46	62.47 ± 7.59
Ours(WHANet)	**87.27 ± 0.78**	**99.08 ± 0.32**	**86.56 ± 1.32**	**76.45 ± 1.67**

Best results are in bold

Table 10 Ablation study of different convolution kernel sizes on the OCTA500-6 M dataset

Convolution kernel configuration	SEN(%)	SPE(%)	DICE (%)	JAC(%)
only5×5	70.34 ± 2.45	86.32 ± 0.65	76.33 ± 1.89	69.34 ± 1.65
Ours(WHANet)	**86.62 ± 1.28**	**99.15 ± 0.33**	**88.46 ± 1.23**	**79.42 ± 0.78**

Best results are in bold

extensive experimental results clearly demonstrate that the combination of convolution kernels with varying receptive fields (i.e., the full settings WHANet in Fig. 6b) consistently yields superior performance on all datasets, compared to using a single convolution kernel (Fig. 6a). Specifically, on the OCTA-RV dataset, WHANet achieves 79.88% in sensitivity, 93.52% in specificity, 72.80% in DICE, and 57.47% in JAC, significantly outperforming the configuration that uses only a 5×5 kernel. Similar performance advantages are also observed on the OCTA500-3 M and OCTA500-6 M datasets, where WHANet achieves 87.27%, 99.08%, 86.56%, and 76.45%, and 86.62%, 99.15%, 88.46%, and 79.42%, respectively. These results strongly support the conclusion that combining multi-scale convolution kernels is highly effective in capturing vascular features of varying scales. In particular, smaller kernels (e.g., 5×5) are more adept at extracting fine-grained details such as microvasculature and edge textures, while larger kernels (e.g., 11×1, 1×11, 21×1, and 1×21) are better suited for recognizing broader vascular structures and global topology. By integrating convolutions across different orientations and scales, the model achieves a more comprehensive fusion of local detail and global spatial context, thereby significantly enhancing segmentation accuracy and structural completeness.

4.7 Generalization experiment

4.7.1 Cross-dataset generalization

To further evaluate the cross-domain generalization capability of WHANet, we conducted cross-dataset experiments using two publicly available OCTA datasets: OCTA-RV and OCTA500-3 M. Specifically, the experiments were configured under two settings: (1) training on OCTA-RV and testing on OCTA500-3 M; and (2) training on OCTA500-3 M and testing on OCTA-RV. These two datasets were selected because they share the same image resolution (304×304 pixels), which

helps eliminate interference caused by differences in image size, ensuring the fairness and consistency of generalization performance evaluation. In contrast, the OCTA500-6 M dataset has an image resolution of 400×400 pixels. Such discrepancies in spatial dimensions can affect key components of convolutional neural networks, including receptive field alignment, feature scaling, and patch extraction, potentially introducing biases in performance evaluation due to size mismatch and compromising the validity and comparability of the generalization results. In addition to resolution differences, there are also variations in case distribution. All samples in OCTA500-6 M are derived from subjects diagnosed with retinal diseases (e.g., diabetic retinopathy [DR], age-related macular degeneration [AMD], and retinal vein occlusion [RVO]), whereas both OCTA500-3 M and OCTA-RV contain a mix of normal and pathological cases. This class imbalance across datasets may adversely impact the fairness of generalization assessment between the source and target domains.

The results ,Table 11, demonstrate that under the setting where the model is trained on OCTA-RV and tested on OCTA500-3 M, WHANet achieves the best performance in both Dice (41.92%) and Jaccard (27.65%) metrics, significantly outperforming other baseline methods such as DBU-Net, RetiFluidNet, and OCT2Former. Moreover, WHANet maintains high stability in terms of Sensitivity (93.81%) and Specificity (83.46%), indicating strong segmentation capability even when facing shifts in data distribution. In the reverse setting, where the model is trained on OCTA500-3 M and tested on OCTA-RV(see Table 12), WHANet again demonstrates strong cross-domain adaptability. It achieves the highest performance with a Dice score of 61.21% and a Jaccard index of 44.56%. Although MsTGANet shows a slight advantage in Specificity (99.61% vs. WHANet's 98.81%), its overall segmentation accuracy (Dice = 59.01%) is slightly inferior to that of WHANet. These results strongly confirm that, under the joint spatial-frequency modeling strategy, WHANet not only delivers excellent performance on in-domain tasks but also exhibits outstanding generalization capability and structural preservation in cross-domain scenarios.

Table 11 Results on Cross-dataset generalization results: Training on OCTA-RV and testing on OCTA500-3 M

Methods	SEN(%)	SPE(%)	DICE(%)	JAC (%)
U-Net	96.07 ± 3.52	81.24 ± 2.13	39.62 ± 3.15	24.81 ± 2.43
Attention U-Net	92.94 ± 2.66	82.15 ± 1.79	40.47± ±1.09	25.48 ± 1.07
CENet	93.30 ± 0.90	81.71 ± 3.10	39.26 ± 2.15	24.53 ± 3.67
DBU-Net	93.29 ± 1.03	82.43 ± 0.99	40.52 ± 2.31	27.15 ± 4.18
MsTGANet	96.12 ± 0.64	81.64 ± 0.84	40.18 ± 3.01	25.26 ± 1.43
OCT2Former	93.12 ± 2.17	80.33 ± 2.64	38.43 ± 3.11	23.92 ± 2.33
RetiFluidNet	**96.15 ± 3.42**	81.07 ± 2.21	39.49 ± 0.06	24.72 ± 2.19
Ours	93.81 ± 0.88	**83.46 ± 3.22**	**41.92 ± 2.78**	**27.65 ± 2.91**

Best results are in bold

Table 12 Results on Cross-dataset generalization results: Training on OCTA500-3 M and testing on OCTA-RV

Methods	SEN(%)	SPE(%)	DICE(%)	JAC (%)
U-Net	44.57 ± 3.06	97.67 ± 0.43	55.76 ± 1.59	39.00 ± 1.37
Attention U-Net	44.66 ± 2.31	97.86 ± 0.34	56.16 ± 2.89	39.47 ± 2.32
CENet	41.30 ± 1.45	98.42 ± 0.67	54.50 ± 1.81	37.72 ± 3.75
DBU-Net	42.85 ± 2.32	**99.61 ± 0.04**	58.17 ± 2.74	39.50 ± 3.12
MsTGANet	47.56 ± 1.51	97.98 ± 0.73	59.01 ± 2.63	42.30 ± 3.26
OCT2Former	42.81 ± 1.47	98.96 ± 0.71	56.93 ± 1.10	40.29 ± 0.99
RetiFluidNet	44.01 ± 3.01	97.94 ± 0.40	55.62 ± 3.03	39.02 ± 2.06
Ours	**48.90 ± 3.41**	98.81 ± 0.18	**61.21 ± 0.98**	**44.56 ± 3.55**

Best results are in bold

4.7.2 Mixed-domain training evaluation

To further evaluate the model's generalization ability across domains with different data distributions, we conducted a mixed-domain experiment. Since the OCTA-500-6 M dataset images have a resolution of 400×400 pixels, while the OCTA-500-3 M and OCTA-RV datasets consist of images at 304×304 pixels, this spatial resolution discrepancy may affect key components of convolutional neural networks, including receptive field alignment, feature scale consistency, and patch extraction. To ensure a fair and reliable assessment of the model's generalization capability, we constructed a mixed dataset consisting of images from OCTA-RV and OCTA-500-3 M. All models were trained and tested on this merged dataset, which was split into training, validation, and test sets in a standard 8:1:1 ratio.

As shown in Table 13, WHANet outperforms most competing methods across multiple metrics. Our approach achieves a Dice score of 77.51% and a Jaccard index of 65.66%, which indicates superior accuracy for complex retinal vessels. Additionally, WHANet obtains a high sensitivity of 89.27% and the highest specificity

Table 13 Performance comparison of different models on the mixed OCTA-RV and OCTA500-3 M Dataset

Methods	SEN(%)	SPE(%)	DICE(%)	JAC (%)
U-Net	89.19 ± 2.53	95.15 ± 0.74	69.73 ± 3.11	53.78 ± 2.34
Attention U-Net	85.97 ± 0.99	95.95 ± 0.31	76.61 ± 2.05	62.41 ± 2.45
CENet	86.49 ± 1.59	94.33 ± 0.38	66.45 ± 3.22	50.11 ± 3.08
DBU-Net	88.81 ± 1.84	96.74 ± 0.79	76.94 ± 2.43	62.84 ± 4.11
MsTGANet	85.92 ± 2.45	97.29 ± 0.46	77.31 ± 1.89	64.09 ± 2.59
OCT2Former	89.09 ± 2.09	95.91 ± 0.80	73.04 ± 2.86	57.78 ± 2.05
RetiFluidNet	87.59 ± 2.02	96.99 ± 0.19	67.39 ± 3.06	51.08 ± 1.91
Ours	**89.27 ± 2.10**	**97.51 ± 0.09**	**77.51 ± 2.17**	**65.66 ± 2.84**

Best results are in bold

among all models at 97.51%, demonstrating a balanced ability to detect true vessels while minimizing false positives.

These results indicate that WHANet's joint spatial-frequency optimization design effectively mitigates domain shift issues, enabling robust generalization under mixed-distribution scenarios while maintaining high segmentation accuracy and topological integrity. This robustness highlights the strong potential of WHANet for practical clinical applications, especially in scenarios involving heterogeneous OCTA images from multiple sources.

These results suggest that WHANet's spatial-frequency collaborative optimization design effectively mitigates domain shifts, enabling robust generalization in mixed-distribution scenarios while preserving high segmentation accuracy and topological integrity. This robustness highlights WHANet's strong potential for real-world clinical applications, especially in heterogeneous OCTA images from multiple sources.

4.7.3 Robustness on color OCTA images

To further assess the robustness and generalization capability of WHANet across different imaging modalities, we conducted additional experiments on the DRIVE dataset, which contains color fundus images exhibiting significant visual differences from the grayscale OCTA images primarily used in this study. Given the differences in image structure, color space, and vascular presentation, this dataset provides a valuable benchmark for evaluating cross-modality performance.

As shown in Table 14, WHANet achieves strong performance across all key metrics. Specifically, it achieves a sensitivity of 80.21%, the highest among all compared methods, indicating superior capability in detecting true vessel pixels–particularly critical for identifying fine microvascular structures. Additionally, WHANet achieves the highest specificity at 98.55%, reflecting its effectiveness in suppressing background noise and non-vascular regions, thereby minimizing false positives. Although NFN+ achieves slightly higher accuracy (96.68%), WHANet still delivers a highly competitive accuracy of 95.24%, striking a better balance between sensitivity and specificity. In contrast to NFN+, which tends to miss fine vessels and

Table 14 Performance comparison on the DRIVE dataset

Model	Sen	Spe	ACC (%)
U-net	79.15 ± 0.23	98.08 ± 0.31	96.40 ± 0.13
Attention U-Net	78.82 ± 0.17	98.48 ± 0.35	96.54 ± 0.07
CENet	80.15 ± 0.22	98.16 ± 0.19	96.49 ± 0.19
R2U-Net	79.23 ± 0.56	98.03 ± 0.48	96.59 ± 0.16
IterNet	79.95 ± 0.26	98.26 ± 0.08	96.57 ± 0.17
NFN+	80.02 ± 0.19	97.90 ± 0.27	**96.68 ± 0.09**
Ours (WHANet)	**80.21 ± 0.43**	**98.55 ± 0.09**	95.24 ± 0.15

Best results are in bold

boundary structures, WHANet offers more comprehensive vascular recognition, making it especially suitable for clinical applications where high recall is essential.

These results demonstrate that WHANet not only excels in grayscale OCTA segmentation tasks but also maintains strong performance under modality shifts to color fundus images. This cross-modality generalization can be attributed to its dual-branch spatial-frequency co-optimization design, which enables the model to learn modality-invariant vascular representations, thereby enhancing adaptability to varying imaging devices and characteristics.

4.8 Heat map visualization for WHANet

The heat map is a visualization tool that represents the intensity or distribution of data using a gradient in different colors. In deep learning, heat maps are commonly employed to illustrate the activation levels of feature maps or attention maps, where blue typically indicates lower intensity (cold regions) and red denotes higher intensity (hot regions). By visualizing heat maps, one can intuitively observe the model's feature extraction performance across different regions, providing deeper insights into its attention distribution and the underlying basis for its decision-making.

From left to right, Fig. 7 displays the following components: the original OCTA image (Image), the heatmap output from the backbone network (Backbone), the output of the proposed WHANet model (WHANet), and the final vessel segmentation results (Result). Specifically, image represents the original retinal OCTA scan, illustrating the vascular structures to be segmented. Backbone shows the heatmap generated by the base U-Net architecture without any additional branches, reflecting its initial ability to perceive vascular features. The WHANet panel presents the heatmap produced by our proposed model, which integrates a hybrid attention mechanism with multi-scale wavelet feature fusion. This output reveals richer color variations and clearer structural details in the feature representations. Finally, the Result panel depicts the binary vessel segmentation masks predicted by each model, where white indicates the segmented vessels and black denotes the background, providing an intuitive visualization of segmentation performance.

One can observe that WHANet produces heatmaps with richer color variations and clearer structural details. The advantages are two-fold: the hybrid attention deep convolutional branch (HADCB) and the multi-scale wavelet feature fusion branch (MWFFB). The Channel Attention Module (CAM) within HADCB enhances vascular feature channels while suppressing background channels, resulting in more vivid red responses in vascular regions and deeper blue in non-vascular regions. Furthermore, the Spatial Depth-Separable Convolutional Attention Module (SDSCAM) employs multi-scale convolutional kernels to capture vessels of various sizes, strengthening the activation of fine vessels and complex branching structures. Meanwhile, the Discrete Wavelet Transform (DWT) component in MWFFB performs multi-scale decomposition, enabling the model to capture both global vascular structure and local details. The low-frequency sub-bands preserve the continuity of main vessels, while the high-frequency sub-bands enhance edge-level features. The multi-frequency feature fusion module (MFFM) further integrates these components

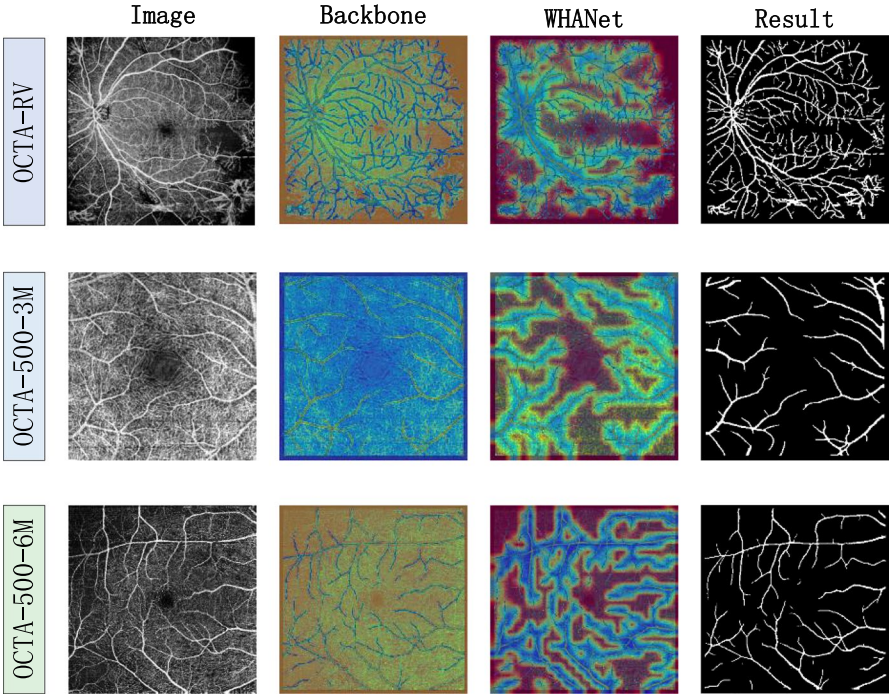

Fig. 7 Qualitative heat map visualization at different model stages. From left to right: original OCTA image, heat map of U-Net, heat map of WHANet, and final segmentation result. From top to bottom: the OCTA-RV, OCTA500-3 M, and OCTA500-6 M datasets

to emphasize vascular boundaries and fine structures. Together, these architectural elements allow WHANet to generate heatmaps that more accurately reflect vascular characteristics, ultimately leading to improved segmentation performance.

5 Discussion

Although WHANet performs well in capturing vascular details through its hybrid attention mechanism and wavelet-based feature fusion strategy, its segmentation accuracy remains suboptimal for certain low-contrast and ultrafine blood vessels–particularly capillaries with diameters less than 2 pixels. This limitation primarily arises from the fact that, in OCTA images, such vessels often exhibit intensity characteristics similar to the surrounding background tissue, making them difficult for the model to distinguish accurately. Future work may focus on enhancing feature extraction modules and integrating advanced contrast enhancement techniques or super-resolution reconstruction methods to improve segmentation performance for these challenging vascular structures. Moreover, the effectiveness of WHANet

heavily relies on the quality and diversity of the training data. In the absence of sufficient variation in retinal vascular morphology or pathological conditions within the dataset, the model's generalization ability and robustness may be limited. To address this issue, semi-supervised learning approaches could be explored to reduce dependence on large-scale annotated datasets and improve the model's adaptability to diverse data distributions. In the clinical practice, future efforts will explore the integration of WHANet with multimodal imaging techniques, such as fluorescein fundus angiography (FFA) and color fundus photography, to provide a more comprehensive assessment of retinal vasculature. For instance, combining the vascular leakage information from FFA with the high-resolution three-dimensional structural data from OCTA may enable more accurate identification and quantification of pathological features. Additionally, considering the significant inter-patient variability in retinal anatomy and disease presentation, we aim to develop a personalized model adaptation strategy. This would enable WHANet to rapidly learn and adjust to individual patient data, thereby improving segmentation accuracy and clinical applicability.

6 Conclusion

In this work, we propose a novel dual-branch framework that integrates wavelet transform and hybrid attention mechanisms, named WHANet, combining spatial and frequency-domain optimization for retinal vessel segmentation in OCTA images. The dual-branch architecture of WHANet facilitates the collaborative extraction and fusion of spatial and frequency features, substantially enhancing its ability to segment fine vessels and complex vascular topologies. Notably, WHANet achieves a significant improvement in sensitivity for detecting small vessels in the OCTA retinal datasets, offering stronger technical support for the early diagnosis of diabetic retinopathy. By leveraging discrete wavelet transform (DWT) and multi-scale feature fusion strategies, WHANet effectively addresses the traditional trade-off between noise suppression and the preservation of vascular boundary details. In summary, the proposed approach obtains the fine segmentation outputs that not only retain the integrity of the global vascular structure but also capture vessel edge information with higher precision.

Appendix A

To provide a clearer illustration of the feature processing flow within the WHANet architecture, this section offers a detailed description of the input and output feature map dimensions for its major modules. By presenting the dimensional changes

Table 15 Components and Descriptions

Component	Description
HADCB	
CAM	Input feature map F_{in} has dimensions $C \times H \times W$, where C is the number of channels, H is the height, and W is the width. Through global average pooling (GAP) and global max pooling (GMP), the dimensions become $C \times 1 \times 1$. By using a multi-layer perceptron (MLP) for nonlinear transformation, the output dimensions remain $C \times 1 \times 1$. Finally, through a Sigmoid activation function, the attention map M is generated, with dimensions $C \times 1 \times 1$. After element-wise multiplication with F_{in}, the resulting feature map F has dimensions $C \times H \times W$
SDSCMD	Input feature map F has dimensions $C \times H \times W$. Through multi-scale feature fusion (e.g., 5×5, 1×7, 7×1), multiple feature maps are generated, each with dimensions $C \times H \times W$
Channel Mixing	Through 1x1 convolution, the feature map is transformed into a new feature map M, with dimensions $1 \times H \times W$. Subsequently, element-wise multiplication with the channel feature map results in the output feature map, with dimensions $C \times H \times W$
MWFFB	
DWT	Input feature map has dimensions $C \times H \times W$. Through DWT decomposition, four sub-bands are generated: one low-frequency sub-band D_1 and three high-frequency sub-bands D_2, D_3, D_4, each with dimensions $C/2 \times H/2 \times W/2$ (using 2x2 down-sampling)
MFFM	Input feature maps D_1, D_2, D_3, D_4 have dimensions $C/2 \times H/2 \times W/2$. Through multi-scale feature fusion (MFFM), the output feature maps d_1, d_2, d_3 and d_4 are generated, each with dimensions $C/2 \times H/2 \times W/2$
IDWT	The four sub-bands are fused to reconstruct the original feature map, with dimensions $C \times H \times W$

in the Channel Attention Module (CAM), Spatial Depth-Separable Convolutional Attention Module (SDSCAM), Discrete Wavelet Transform (DWT) module, multi-frequency feature fusion module (MFFM), and Inverse Discrete Wavelet Transform (IDWT) module, we aim to help readers gain a deeper understanding of the individual functions and collaborative mechanisms of these components in feature extraction and fusion. The detailed information is shown in Table 15.

Appendix B

To enhance the rigor and clarity of the paper, we have provided detailed definitions and explanations for all variables involved in the mathematical expressions. Table 16, Table 17 present a comprehensive list of all equations used in the paper, along with corresponding descriptions of each variable.

Table 16 Descriptions for each variable used in the mathematical formulas

Formula	Variable explanation
(1) $M_C = \sigma\big(\mathrm{MLP}\big(\mathrm{GAP}(F_{\mathrm{in}})\big) + \mathrm{MLP}\big(\mathrm{GMP}(F_{\mathrm{in}})\big)\big)$	GAP(·) and GMP(·) denote the global average and maximum pooling. MLP(·) is the shared multilayer perceptron. σ is the sigmoid activation. F_{in} is the input feature map
(2) $F_C = M_C \odot F_{\mathrm{in}}$	$\odot$ denotes element-wise multiplication. M_C is the computed channel attention weighted graph. F_C denotes the computed weighted feature map
(3) $M_s = W_{1\times1}\Big(\mathbb{N}\Big(\sum_{i=0}^{3} \lambda_i \cdot g\big(\beta_i\big(\mathrm{DConv}_i(F_C)\big)\big)\Big)\Big)$	$W_{1\times1}$ denotes the 1×1 convolution operation, which performs a linear transformation. $\mathbb{N}(\cdot)$ represents a normalization operation, specifically Layer Normalization. λ_i is a learnable weight parameter used to adjust the contribution of each branch. $g(\cdot)$ denotes the nonlinear activation function, ReLU(·). $\beta_i(\cdot)$ represents the computation process of the i-th branch. DC(·) denotes depthwise convolution
(4) $F_S = M_S \odot F_C$	F_C is the feature map processed by the channel attention mechanism as input to the spatial depth-separated convolution attention module. M_s represents the computed spatial depth-separated convolutional attention weight map. $\odot$ denotes element-wise multiplication
(5) $D_4(x) = \sum_k c_{j,k}\phi_{j,k}(x)$	$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ is the scale function, which is used to capture the low-frequency features of the image. j is the decomposition scale (larger means rougher). k is the translation parameter, which controls the spatial position of the basis function. $c_{j,k}$ is the scale coefficients, which reflects the low-frequency energy distribution of the input image at the scale j and position k

Table 17 Descriptions for each variable used in the mathematical formulas

Formula	Variable explanation
(6) $D_j(x) = \sum_k d_{j,k}\psi_{j,k}(x)$	$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ is the Wavelet function, which is used to capture the high-frequency features of the image. $d_{j,k}$ denotes the Wavelet coefficient, which reflects the high-frequency energy distribution of the input image at scale j and position k
(7) $c_j = \sigma\left(\mathrm{CNN}_{\theta_j}(D_j)\right), (j = 1, 2, 3)$	$\mathrm{CNN}_{\theta_j}(\cdot)$ denotes Convolutional Neural Network for high-frequency features D_j, θ_j as its parameters. $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid function for enhancing the nonlinear representation of features
(8) $d_j = \mathrm{RDN}(D_j)$	$\mathrm{RDN}(\cdot)$ refers to a residual dense network composed of stacked residual dense blocks
(9) $d_4 = d_4 \odot c_j, (j = 1, 2, 3)$	d_j denotes the high-frequency vector processed by CNN and Sigmoid function. $\odot$ denotes element-by-element multiplication operation. $\cup$ denotes the concatenation operation of sets
(10) Output $= \{d_4 \odot \sigma(\mathrm{CNN}_{\theta_j}(D_j))\} \cup \{d_j\}$	d_j denotes the high-frequency vector processed by CNN and Sigmoid function. $\odot$ denotes element-by-element multiplication operation. $\cup$ denotes the concatenation operation of sets
(11) Output $= \{d_1, d_2, d_3, d_4\}$	-
(12) $A(x) = \sum_k d_{j,k}\phi_{j,k}(x)$	J is the maximum decomposition scale, which represents the lowest frequency resolution level. $A(x)$ is the low-frequency signal obtained. $d_{4,k}$ is the coefficient of the low-frequency feature map d_4, which represents the low-frequency component of the signal at scale J. $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ is the basis function of the scale function at scale J
(13) $D_j(x) = \sum_k c_{j,k}\psi_{j,k}(x), \quad j = 1, 2, 3$	j is the scale of the high-frequency feature map. $c_{j,k}$ is the coefficient of the high-frequency feature map c_j, which represents the high-frequency component of the signal at scale j. $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ is the basis function of the wavelet function at scale j. Parameters $j = 1, 2, 3$ denote the high-frequency components at different scales, respectively
(14) $f_{\mathrm{out}}(x) = A(x) + \sum_{j=1}^{3} D_j(x)$	$f_{\mathrm{out}}(x)$ is the complete reconstructed signal obtained. x is the pixel position of the image, which is used to describe the distribution of the image in the spatial domain

Author Contributions Shuxin Xue and Fei Ma wrote the paper and conceptualized and designed the model. Fei Ma and Jing Meng managed the project and obtained funding. Xiaofen Ai, Yuefeng Ma collected and monitored the data. Shuxin Xue, Yanfei Guo, Fen Yan, Zhaohui Zhang, Guangmei Jia analyzed and processed the data. Shuxin Xue, Yanfei Guo, Fen Yan, Zhaohui Zhang, Guangmei Jia performed data analysis and processing.

Funding This work was supported by Natural Science Foundation of Shandong Province (No:ZR2020MF105), Guangdong Provincial Key Laboratory of Biomedical Optical Imaging Technology (No:2020B121201010), the Natural National Science Foundation of China (62175156,61675134),

Science and technology innovation project of Shanghai Science and Technology Commission (19441905800, 22S31903000), Qufu Normal University Foundation for High Level Research (116-607001), and Academic-Enterprise Joint Research and Development Project(KJ2025HX019).

Data availability No datasets were generated or analyzed during the current study.

Code availability The codes used during the study are available from the corresponding author by request.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Almotiri J, Elleithy K, Elleithy A (2018) Retinal vessels segmentation techniques and algorithms: a survey. Appl Sci 8(2):155
2. Srinidhi L, Aparna C, Rajan P (2017) Recent advancements in retinal vessel segmentation. J Med Syst 41:1–22
3. Or C, Sabrosa AS, Sorour O, Arya M, Waheed N (2018) Use of octa, fa, and ultra-widefield imaging in quantifying retinal ischemia: a review. Asia-Pacific J Ophthalmol 7(1):46–51
4. Li M, Chen Y, Ji Z, Xie K, Yuan S, Chen Q, Li S (2023) Corrections to image projection network: 3d to 2d image segmentation in octa images. IEEE Trans Med Imaging 42(1):329–329
5. Jiang Z, Huang Z, Qiu B, Meng X, You Y, Liu X, Geng M, Liu G, Zhou C, Yang K (2020) Weakly supervised deep learning-based optical coherence tomography angiography. IEEE Trans Med Imaging 40(2):688–698
6. Kadry S, Dhanaraj RK, Manthiramoorthy C (2024) Res-unet based blood vessel segmentation and cardio vascular disease prediction using chronological chef-based optimization algorithm based deep residual network from retinal fundus images. Multimed Tools Appl 83(40):87929–87958
7. Shin Y-I, Nam KY, Lee SE, Lim H-B, Lee MW, Jo Y-J, Kim J-Y (2019) Changes in peripapillary microvasculature and retinal thickness in the fellow eyes of patients with unilateral retinal vein occlusion: an octa study. Investig Ophthalmol Visual Sci 60(2):823–829
8. Novais E, Waheed N (2016) Oct angiography in retinal and macular diseases. Am Orthop J 56:132–138
9. Nugroho HA, Aras RA, Lestari T, Ardiyanto I (2017) Retinal vessel segmentation based on frangi filter and morphological reconstruction. In: 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), pp 181–184. IEEE
10. Aguirre-Ramos H, Avina-Cervantes JG, Cruz-Aceves I, Ruiz-Pinales J, Ledesma S (2018) Blood vessel segmentation in retinal fundus images using gabor filters, fractional derivatives, and expectation maximization. Appl Math Comput 339:568–587
11. Annunziata R, Trucco E (2016) Accelerating convolutional sparse coding for curvilinear structures segmentation by refining scird-ts filter banks. IEEE Trans Med Imaging 35(11):2381–2392
12. Hoover A, Kouznetsova V, Goldbaum M (2000) Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans Med Imaging 19(3):203–210
13. Jiang X, Mojon D (2003) Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images. IEEE Trans Pattern Anal Mach Intell 25(1):131–137
14. Saleh MD, Eswaran C, Mueen A (2011) An automated blood vessel segmentation algorithm using histogram equalization and automatic threshold selection. J Digit Imaging 24:564–572
15. Lesage D, Angelini ED, Bloch I, Funka-Lea G (2009) A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. Med Image Anal 13(6):819–845
16. Staal J, Abràmoff MD, Niemeijer M, Viergever MA, Van Ginneken B (2004) Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 23(4):501–509

17. Soares JV, Leandro JJ, Cesar RM, Jelinek HF, Cree MJ (2006) Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. IEEE Trans Med Imaging 25(9):1214–1222
18. Lupascu CA, Tegolo D, Trucco E (2010) Fabc: retinal vessel segmentation using adaboost. IEEE Trans Inf Technol Biomed 14(5):1267–1274
19. Carlo TE, Romano A, Waheed NK, Duker JS (2015) A review of optical coherence tomography angiography (octa). Int J Retina Vitreous 1:1–15
20. Ma Y, Hao H, Xie J, Fu H, Zhang J, Yang J, Wang Z, Liu J, Zheng Y, Zhao Y (2020) Rose: a retinal oct-angiography vessel segmentation dataset and new model. IEEE Trans Med Imaging 40(3):928–939
21. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al (2018) Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999
22. Mou L, Zhao Y, Chen L, Cheng J, Gu Z, Hao H, Qi H, Zheng Y, Frangi A, Liu J (2019) Cs-net: Channel and spatial attention network for curvilinear structure segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, pp 721–730. Springer
23. Guo C, Szemenyei M, Yi Y, Wang W, Chen B, Fan C (2021) Sa-unet: Spatial attention u-net for retinal vessel segmentation. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 1236–1242. IEEE
24. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp 205–218. Springer
25. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021) Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306
26. Zhong Y, Li B, Tang L, Kuang S, Wu S, Ding S (2022) Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4504–4513
27. Chanda PB, Sarkar SK (2020) Discrete wavelet transform based segmentation approach for identification of cancer diseases from mammogram images. In: 2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), pp 1–6. IEEE
28. Wang C, Zhan N, Jia L, Zhang J, Li Y (2018) Dwt-based adaptive decomposition method of electrostatic signal for dilute phase gas-solid two-phase flow measuring. Powder Technol 329:199–206
29. Frangi AF (1998) Multiscale vessel enhancement filtering. Medical Image Comput Comput-Assist Interv 1496:130–137
30. Al-Diri B, Hunter A, Steel D (2009) An active contour model for segmenting and measuring retinal vessels. IEEE Trans Med Imaging 28(9):1488–1497
31. Otsu N (1975) A threshold selection method from gray-level histograms. Automatica 11(285–296):23–27
32. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp 234–241. Springer
33. Cao Y, Liu S, Peng Y, Li J (2020) Denseunet: densely connected unet for electron microscopy image segmentation. IET Image Proc 14(12):2682–2689
34. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J (2019) Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans Med Imaging 38(10):2281–2292
35. Gao Y, Ai D, Wang Y, Cao K, Song H, Fan J, Xiao D, Zhang T, Wang Y, Yang J (2025) Spatio-temporal correspondence attention network for vessel segmentation in x-ray coronary angiography. Biomed Signal Process Control 99:106792
36. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
37. Zhen Z, Hu Y, Feng Z (2024) Freqmamba: viewing mamba from a frequency perspective for image deraining. arXiv preprint arXiv:2404.09476
38. Wang M, Zhu W, Shi F, Su J, Chen H, Yu K, Zhou Y, Peng Y, Chen Z, Chen X (2021) Mstganet: automatic drusen segmentation from retinal oct images. IEEE Trans Med Imaging 41(2):394–406
39. 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. Comput Methods Programs Biomed 233:107454

40. Luo X, Peng L, Ke Z, Lin J, Yu Z (2025) Pa-net: a hybrid architecture for retinal vessel segmentation. Pattern Recogn 161:111254

Authors and Affiliations

Shuxin Xue[1] · Zhaohui Zhang[1] · Fen Yan[2] · Fei Ma[1] · Guangmei Jia[1] · Yanfei Guo[1] · Yuefeng Ma[1] · Xiaofei Ai[1] · Jing Meng[1]

✉ Fei Ma
mafei0603@163.com

✉ Jing Meng
jingmeng@qfnu.edu.cn

Shuxin Xue
15006538205@163.com

Zhaohui Zhang
zhaohuizhang@qfnu.edu.cn

Fen Yan
17753701819@163.com

Guangmei Jia
2554795524@qq.com

Yanfei Guo
guoyanfei2022@qfnu.edu.cn

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

[2] Ultrasound Medicine Department, Qufu People's Hospital, Shandong, China