



An Efficient Hybrid Vision Transformer for Topology-Preserving Retinal Vessel Segmentation

A. Reethika^{1*} , B. Sharmila² 

¹ Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore 641022, India

² Department of Electronics and Instrumentation Engineering, Sri Ramakrishna Engineering College, Coimbatore 641022, India

Corresponding Author Email: reethika.a@srec.ac.in

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430420>

ABSTRACT

Received: 11 June 2026
Revised: 9 August 2026
Accepted: 17 August 2026
Available online: 31 August 2026

Keywords:

retinal vessel segmentation, Vessel-Aware Hybrid Transformer, multi-scale feature fusion, Bidirectional Segmentation Refinement, retinal image analysis

Retinal vessel segmentation is a critical step in ophthalmic image analysis for the early detection and monitoring of diabetic retinopathy and other vascular diseases. Accurate extraction of thin vessels, complex bifurcations, and low-contrast structures remains challenging. Existing convolutional neural network (CNN) models have limited receptive fields and therefore struggle to preserve global vascular topology, while pure Vision Transformer (ViT) models are computationally expensive and often lose fine vessel details because of patch-based processing. To overcome these limitations, we propose a lightweight Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet). The network combines a Multi-Scale Vision Transformer (MSVT) encoder for long-range dependency modelling, a Hybrid Feature Fusion Network (HFF-Net) for multi-scale local-global feature integration, a Bidirectional Segmentation Refinement Network (BiSRNet) for topology-preserving refinement, and a hybrid Squeeze-and-Excitation Convolutional Block Attention Module (SE-CBAM) for precise vessel localisation. Extensive experiments on the DRIVE, CHASE_DB1 and STARE datasets demonstrate that VAHT-BiSRNet achieves Dice scores of 0.882, 0.873 and 0.878, accuracies of 0.9812, 0.9786 and 0.9803, and AUC values of 0.9936, 0.9928 and 0.9931, respectively, while requiring only 2.21 million parameters, 3.62 GFLOPs and 142 ms inference time. The results confirm that the proposed framework successfully balances segmentation accuracy, vessel topology preservation, and computational efficiency, making it a promising candidate for computer-aided ophthalmic screening systems.

1. INTRODUCTION

Retinal vessel segmentation has become an essential task in ophthalmic image analysis because the retinal vascular network provides valuable anatomical and physiological biomarkers for the diagnosis, monitoring, and prognosis of numerous ocular and systemic diseases, including diabetic retinopathy, glaucoma, hypertensive retinopathy, and cardiovascular disorders [1-3]. Morphological characteristics such as vessel calibre, tortuosity, branching geometry, vascular density, and vessel connectivity reflect pathological alterations in retinal circulation and therefore play a critical role in computer-aided clinical decision-making. Automated vessel segmentation in retinal fundus imaging is gaining attention for its non-invasive nature, low cost, and broad accessibility, making it an effective tool for large-scale ophthalmic screening. Nevertheless, accurate segmentation remains challenging because retinal images exhibit substantial variations in illumination, imaging artefacts, low vessel-to-background contrast, pathological abnormalities, and highly heterogeneous vessel diameters. Furthermore, the presence of thin capillary branches and intricate bifurcation structures

significantly increases segmentation difficulty, frequently leading to discontinuous vessel extraction, inaccurate vessel boundaries, and poor preservation of retinal vascular topology [4, 5].

Recent developments in deep learning have substantially improved retinal vessel segmentation performance. Convolutional Neural Networks (CNNs), particularly encoder-decoder architectures derived from U-Net, have demonstrated remarkable competency in learning hierarchical vessel representations through convolutional feature extraction and multi-scale information aggregation. Numerous attention-enhanced CNN models have further improved vessel boundary delineation and peripheral vessel detection by incorporating skip connections and adaptive feature refinement. Despite these advances, CNN-based methods have limited receptive fields, hindering their capacity to capture long-range spatial relationships and global vascular connectivity in the retinal network. To address these limitations, Vision Transformer (ViT) architectures have recently been introduced to exploit self-attention mechanisms for modelling global contextual dependencies and vessel topology. Transformer-based models have shown superior

capability in preserving vessel continuity and structural consistency; however, they generally require large-scale annotated datasets, introduce considerable computational complexity, and often exhibit reduced sensitivity to extremely thin vessels due to patch-based feature representation. Consequently, neither CNNs nor Transformers alone provide an optimal balance between local vessel discrimination, global contextual modelling, and computational efficiency.

Although CNN-based encoder-decoder networks (e.g., U-Net and its variants) excel at extracting local vessel morphology, their limited receptive fields prevent them from modelling long-range vessel connectivity; this frequently results in fragmented thin vessels and broken bifurcation structures. Conversely, pure ViT architectures capture global context effectively but suffer from three practical drawbacks: (i) high computational cost (often exceeding 10 GFLOPs), (ii) reduced sensitivity to capillaries because of fixed patch tokenisation, and (iii) the need for large annotated datasets. Hybrid CNN–Transformer models partially mitigate these issues, yet most still rely on heavy feature-fusion blocks or deep Transformer backbones that compromise real-time applicability. Consequently, there remains a clear need for a lightweight architecture that simultaneously (a) preserves global vascular topology, (b) accurately localises fine vessels, and (c) maintains low parameter count and fast inference.

This paper proposes a novel Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) for automated retinal vessel segmentation. The framework integrates a Multi-Scale Vision Transformer (MSVT) encoder to capture long-range vascular dependencies, a Hybrid Feature Fusion Network (HFF-Net) to learn multi-scale vessel representations, and a Bidirectional Segmentation Refinement Network (BiSRNet) to facilitate

vessel-aware feature interaction between global and local representations in Table 1. Furthermore, a hybrid attention mechanism combining Squeeze-and-Excitation (SE) with the Convolutional Block Attention Module (CBAM), hereafter referred to as Squeeze-and-Excitation Convolutional Block Attention Module (SE-CBAM), is employed to enhance vessel localisation and suppress background interference. Extensive experiments conducted on the DRIVE, CHASE_DB1, and STARE datasets demonstrate the effectiveness and robustness of the proposed framework.

The main contributions of this work are summarised as follows:

- A novel VAHT-BiSRNet architecture is introduced for retinal vessel segmentation by integrating CNN-based local learning with transformer-based global topology modelling.
- A MSVT encoder is developed to capture long-range vessel dependencies and maintain vascular continuity across multiple spatial resolutions.
- A HFF-Net and BiSRNet are presented to enhance multi-scale vessel representation and bidirectional feature interaction.
- A hybrid SE-CBAM attention mechanism is implemented to improve vessel localisation, reduce background noise, and preserve fine vascular structures, while the overall model remains lightweight (2.21 M parameters).

The remainder of this manuscript is organised as follows. Related works are presented in Section 2. The datasets, preprocessing procedures and the proposed hybrid CNN–ViT architecture are explained in Section 3. The experimental setup and performance assessment results are elaborated in Section 4. The work is concluded in Section 5 with important findings and suggestions for further research.

Table 1. Glossary of key modules and terms

Module/Term	Full Name	Primary Function
VAHT-BiSRNet	Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network	Overall proposed architecture for topology-preserving vessel segmentation.
MSVT	Multi-Scale Vision Transformer	Captures long-range vessel dependencies at multiple patch scales.
HFF-Net	Hybrid Feature Fusion Network	Fuses CNN, ViT, and texture feature streams into a unified representation.
BiSRNet	Bidirectional Segmentation Refinement Network	Bidirectional global↔local feature exchange for topology preservation.
SE-CBAM	Squeeze-and-Excitation + Convolutional Block Attention Module	Channel and spatial attention for precise vessel localisation.
Tri-Stream Encoding	CNN + ViT + Texture parallel branches	Complementary multi-view feature extraction prior to fusion.

Note: Convolutional Neural Network (CNN), Vision Transformer (ViT), Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet), Multi-Scale Vision Transformer (MSVT), Hybrid Feature Fusion Network (HFF-Net), Bidirectional Segmentation Refinement Network (BiSRNet), Squeeze-and-Excitation Convolutional Block Attention Module (SE-CBAM).

2. RELATED WORKS

Retinal vessel segmentation has attracted considerable attention due to its importance in diagnosing ophthalmic and systemic diseases. Traditional vessel extraction approaches relied on thresholding, matched filtering, morphological processing, and handcrafted feature engineering. Although these methods achieved reasonable performance on high-contrast retinal images, their effectiveness was significantly reduced in the presence of noise, uneven illumination, vessel crossings, and low-contrast microvascular structures. Consequently, deep learning-based segmentation frameworks

have become the dominant paradigm for automated retinal vessel analysis.

2.1 Convolutional Neural Network-based retinal vessel segmentation

CNNs can directly learn hierarchical vessel representations from retinal images and have shown extraordinary performance in retinal vascular segmentation [6, 7]. Because of its skip-connection method and encoder-decoder design, U-Net and its variants continue to be among the most popular systems. RSP-SA U-Net introduced residual attention

mechanisms and pyramid spatial attention modules to improve fine vessel extraction and vessel boundary preservation. Similarly, FSG-Net [8] employed full-scale guided feature representation and attention-guided filtering to enhance vascular continuity and contextual awareness. Recent attention-based U-Net variants further improved vessel localisation by selectively emphasising vessel-related features while suppressing background noise [9]. Despite these improvements, CNN-based techniques remain constrained by their localised receptive fields, which limit their capacity to capture global vascular structure and long-range vessel interdependence. This restriction frequently results in fragmented segmentation, especially for complicated bifurcation locations and thin arteries.

2.2 Transformer-based retinal vessel segmentation

The capacity of ViTs to model long-range contextual linkages through self-attention has made them effective for retinal vascular segmentation. TSNet [10] included a dual-path decoder and integrated transformer modules into a U-Net architecture to enhance vessel continuity and structural representation. Transformer-enhanced vessel segmentation networks further included attention processes and multi-scale fusion to improve contextual knowledge at various vessel sizes. When compared with traditional CNNs, these techniques showed better global feature modelling skills. However, transformer-based models often need substantial processing power and large training data [11, 12]. Moreover, their patch-based representation may weaken sensitivity to fine vascular details, resulting in challenges in accurately segmenting thin capillaries and vessel boundaries.

2.3 Hybrid Convolutional Neural Network-Transformer architectures

A number of hybrid models have been proposed to overcome the drawbacks of standalone CNN and transformer models. CoVi-Net [13] created a Bidirectional Weighted Feature Fusion approach and a Local-Global Feature Aggregation module to integrate local vessel characteristics with global contextual information. PA-Net [14] created a Lightweight Parallel Transformer and Adaptive Vascular Feature Fusion module to improve thin vessel continuity while preserving computational efficiency. CTAUNet employed a

collaborative transformer attention mechanism to integrate convolutional features and transformer representations through multi-scale feature fusion. These approaches achieved improved segmentation accuracy by exploiting the strengths of CNNs and transformers. Nevertheless, many hybrid architectures still rely on complex feature fusion mechanisms and insufficiently exploit vessel-specific contextual interactions, limiting their ability to simultaneously preserve global topology and fine vascular structures.

2.4 Attention-guided and multi-scale vessel representation methods

Attention processes and multi-scale feature representation for retinal vascular segmentation have been highlighted in recent research. G2ViT [15, 16] used multi-scale edge feature attention to maintain vessel boundaries while modelling vascular structure using graph neural networks and ViTs. TED-SCNet [17-21] introduced dual-scale morphological enhancement and dynamic snake-shaped convolutions to improve fine vessel segmentation and vascular connectivity. The Multi-Level Bidirectional Attention Network [22-25] proposed dynamic directional attention and multi-feature fusion modules to preserve vessel details while suppressing background interference. Although these methods achieved promising performance, balancing segmentation accuracy, computational efficiency, and vessel continuity remains a challenging problem.

2.5 Research gap and motivation

From the above discussion, it can be observed that existing CNN-based methods effectively capture local vessel details but struggle to model global vascular topology. Transformer-based methods enhance long-range dependence modelling; however, they may become less sensitive to fine vessel structures and frequently demand a high computing cost. Although recent hybrid CNN-Transformer architectures have improved segmentation performance, most existing frameworks do not simultaneously address global vessel continuity, multi-scale vessel representation, lightweight deployment, and vessel-aware attention-guided feature fusion (summarised in Table 2). Furthermore, the interaction between transformer-derived global context and CNN-derived local vessel features remains insufficiently explored.

Table 2. Summary of existing retinal vessel segmentation methods and limitations

Model	Year	Methodology	Dataset(s)	Accuracy (%)	Contribution	Limitation
U-Net	2015	Encoder-Decoder CNN	DRIVE, STARE	95.60	Foundational biomedical segmentation architecture	Limited multi-scale representation
U-Net++	2018	Nested Dense Skip Connections	DRIVE, CHASE_DB1	96.70	Reduced semantic gap between encoder and decoder	Increased parameter complexity
MedT	2021	Medical Transformer	DRIVE, STARE	97.10	Long-range dependency modeling	High computational cost
CoVi-Net	2024	CNN + ViT + LGFA + BWF	DRIVE, STARE, CHASE_DB1	97.61	Local-global feature aggregation	High computational complexity
G2ViT	2024	GNN + ViT + MEFA	DRIVE, CHASE_DB1, STARE	97.80	Graph-guided vessel topology learning	Increased model complexity
CTAUNet	2025	Swin Transformer + U-Net	DRIVE, CHASE_DB1, STARE	98.10	Collaborative transformer attention	Large parameter count
PA-Net	2025	Lightweight Parallel Transformer	DRIVE, CHASE_DB1, STARE	97.20	Efficient vessel continuity learning	Limited topology preservation

Note: Convolutional Neural Network (CNN), Vision Transformer (ViT).

In order to overcome these constraints, this work proposes the VAHT-BiSRNet. The proposed framework integrates an MSVT, HFF-Net, BiSRNet, and SE-CBAM attention mechanism to jointly exploit global contextual information and fine-grained vascular features. The architecture is specifically designed to preserve thin vessels, maintain vascular continuity, enhance bifurcation representation, and achieve efficient segmentation performance across the DRIVE, CHASE_DB1, and STARE benchmark datasets.

3. METHODOLOGY

3.1 Overview of the proposed Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network framework

The proposed VAHT-BiSRNet is designed to accurately segment retinal blood vessels while preserving vascular continuity, bifurcation structures, and thin vessel details. Unlike traditional CNN-based architectures that mostly concentrate on local feature extraction, the proposed technique combines attention-guided refinement, vessel-aware feature fusion, and transformer-based global contextual learning. The complete architecture consists of seven major stages: image acquisition, preprocessing, MSVT encoding, HFF-Net,

BiSRNet, SE-CBAM attention enhancement, and vessel mask decoding, shown in Figure 1. The whole architecture is especially built to preserve computational efficiency while capturing both global vessel structure and local vascular anatomy.

3.2 Datasets and preprocessing

The proposed framework was evaluated on three publicly available benchmark datasets: DRIVE, CHASE_DB1 and STARE. These datasets were deliberately selected because they present complementary challenges that comprehensively test the robustness of a vessel segmentation model.

DRIVE contains 40 colour fundus images (565×584) acquired from a diabetic retinopathy screening programme in the Netherlands. It is the most widely used benchmark and provides standardised annotations, making it ideal for controlled comparison. CHASE_DB1 comprises 28 images (999×960) captured from multi-ethnic school children; it exhibits larger vessel-width variation and higher inter-subject anatomical variability. STARE consists of 20 images (700×605) that include both healthy and pathological retinas with significant illumination variation and lesion interference. Together, the three datasets cover a broad spectrum of imaging conditions, vessel morphologies, and pathological presentations.

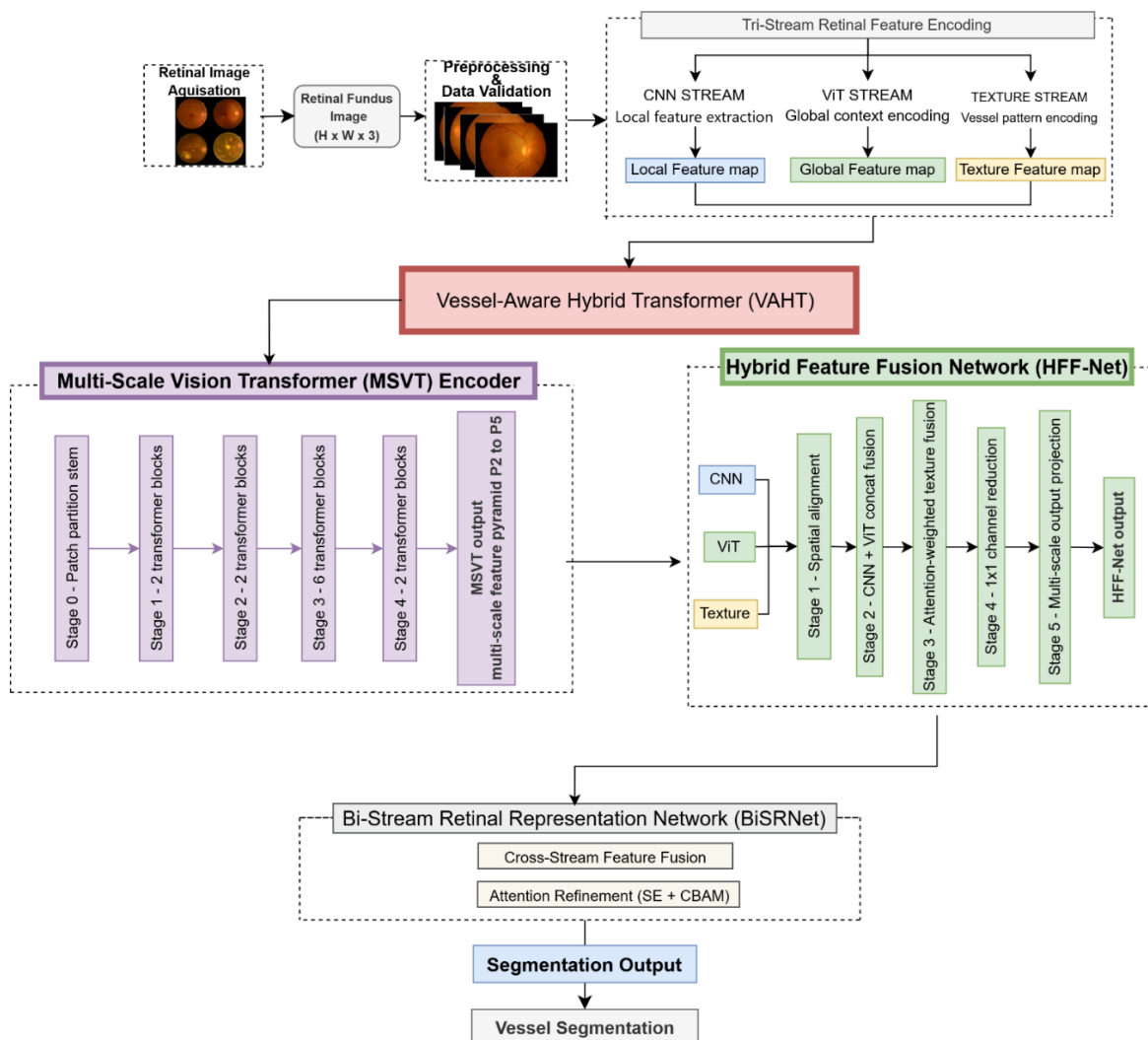


Figure 1. Overall architecture of the proposed Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) framework for retinal vessel segmentation

For all three datasets, a consistent 70: 15: 15 split (training: validation: testing) was adopted while preserving the original vessel-to-background pixel distribution in each subset. The split ratio was chosen to maximise training data while still providing statistically meaningful validation and test sets given the modest total image counts. Data augmentation (horizontal/vertical flips, rotations of $\pm 25^\circ$, brightness/contrast jitter and random cropping) was applied exclusively to the training subset to prevent information leakage. Class imbalance is severe in all three datasets (vessel pixels typically constitute only 8–12% of the image); the composite Binary Cross-Entropy (BCE) + Dice loss (Section 3.7) was therefore employed to mitigate this imbalance.

Prior to feature extraction, retinal images undergo a comprehensive preprocessing pipeline to improve vessel visibility and reduce image variability. Initially, image quality assessment is performed to identify blurred, underexposed, or low-contrast images that may adversely affect segmentation performance. To ensure uniform intensity distributions across the datasets, all the images are now normalized and scaled to a uniform resolution. The Normalized standard image size is $X \in R^{512 \times 512 \times 3}$ intensity distributions and improves robustness of vessel feature extraction across the given datasets, as mentioned in Eq. (1). Images with low V_{blur} values are considered as blurred and may negatively affect vessel segmentation performance is shown in Eq. (2).

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

where, I = original retinal image, I_{min} = minimum pixel intensity, I_{max} = maximum pixel intensity, I_{norm} = normalized image.

$$V_{blur} = Var(\nabla^2 I) \quad (2)$$

where, $\nabla^2 I$ = Laplacian operator, $Var(\cdot)$ = variance function, V_{blur} = blur score.

To increase the vessel background contrast, a familiar Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied for more visualization of fine vascular structures. To increase training diversity and improve model robustness, several augmentation techniques, including horizontal flipping, rotation, brightness adjustment, and contrast variation, are employed. These preprocessing operations collectively improve vessel visibility, suppress background artifacts, and facilitate more effective vessel representation learning.

3.3 Hybrid tri-stream encoding

The proposed method incorporates a tri-stream retinal feature encoding module to extract complementary retinal vessel representations before hybrid fusion. This module consists of three parallel branches: a CNN stream for local vessel morphology extraction, a ViT stream for global contextual vessel modeling, and a texture stream for vessel pattern encoding, as in Figure 2. In the CNN branch, a ResNet backbone followed by multi-scale convolutional branches and a Feature Pyramid Network (FPN) extracts fine vessel boundaries, bifurcations, and multi-resolution vascular structures to produce a local feature map. Simultaneously, the ViT branch partitions the input image into 16×16 patches and applies 12-head self-attention with positional encoding to

capture long-range vessel dependencies and global vascular topology, producing a global feature map. The texture branch employs a Gabor filter bank, a U-Net sub-network, and Local Binary Pattern (LBP) based texture encoding to emphasize line-like vessel patterns and local contrast variations, resulting in a texture feature map. These three complementary feature maps are then fed into the proposed Vessel-Aware Hybrid Transformer (VAHT), enabling robust multi-view vessel representation learning for accurate retinal vessel segmentation.

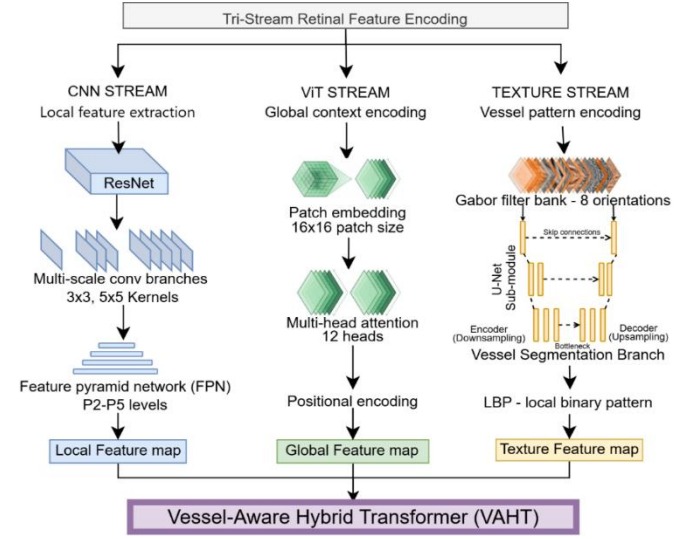


Figure 2. Hybrid tri-stream encoding

3.4 Multi-Scale Vision Transformer encoder

The MSVT serves as the primary global feature extraction module within the proposed framework. The preprocessed retinal image is separated into image patches and projected into a sequence of embedded tokens. Converts retinal vessel structures into token representations suitable for transformer processing, in Eq. (3).

$$Z_0 = [x_p^1 E; x_p^2 E; \dots \dots \dots; x_p^N E] + E_{pos} \quad (3)$$

where, $x_p^i = i^{th}$ image patch, E = embedding matrix, N = number of patches, E_{pos} = positional embedding, Z_0 = transformer input sequence.

Unlike conventional transformers operating at a single scale, the MSVT employs multiple patch resolutions to simultaneously capture coarse and fine vascular patterns as shown in Eq. (4) below.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where, Q = query matrix, K = key matrix, V = value matrix, d_k = key dimension.

This enables effective preservation of vessel continuity, bifurcation relationships, and structural connectivity across the retinal image. The global vessel transformer output is expressed in Eq. (5).

$$F_{MSVT} = MSA(Z) + MLP(Z) \quad (5)$$

where, $MSA(\cdot)$ = multi-head self-attention, $MLP(\cdot)$ = feed-

forward network, F_{MSVT} = global vessel feature representation. Encodes global vessel connectivity and bifurcation relationships. Resulting transformer features provide rich contextual information that complements the vessel extraction by subsequent modules.

The VAHT-BiSRNet employs an MSVT Backbone to learn hierarchical retinal vessel representations across multiple spatial resolutions shown in Figure 3. The MSVT begins with a patch partition stem, where the image is divided into 4×4 patches and projected into a token sequence for transformer-based vessel encoding. Localised vessel modelling and cross-window contextual interaction are made possible by subsequent transformer stages that process the tokens utilising Window-based multi-head self-attention (W-MSA) and Shifted-window self-attention (SW-MSA). The feature resolution decreases from H/4 to H/32; the MSVT generates a vessel feature pyramid consisting of P2, P3, P4, and P5 representations. Here, P2 preserves fine vessel boundaries and capillary details, P3 and P4 encode intermediate vessel morphology and branching continuity, and P5 captures global vascular topology and long-range vessel dependencies. By jointly representing retinal vessels at multiple semantic and spatial scales, the MSVT branch strengthens vessel continuity preservation and provides robust transformer-based global context for downstream vessel-aware fusion and segmentation.

3.5 Hybrid Feature Fusion Bidirectional Segmentation Refinement Network (HFF-Net + BiSRNet)

Transformer encoders effectively capture global contextual information; accurate vessel segmentation also requires preservation of local vessel morphology. To address this challenge, HFF-Net extracts and aggregates multi-scale vessel representations by Multi-Scale Feature Extraction, capturing both thick vessels and fine capillary structures in Eq. (6).

$$F_i = \text{Conv}_i(I_{avg}) \quad (6)$$

where, Conv_i = convolution with kernel scale, F_i = vessel feature map.

Eq. (7) shows that HFF-Net combines hierarchical convolutional features obtained at different receptive fields and integrates them through adaptive feature fusion mechanisms.

$$F_{fusion} = \sum_{i=1}^n \alpha_i F_i \quad (7)$$

where, α_i = learnable fusion weight, F_i = multi-scale feature map.

This process enables simultaneous representation of thick vessels, thin capillaries, vessel crossings, and bifurcation regions.

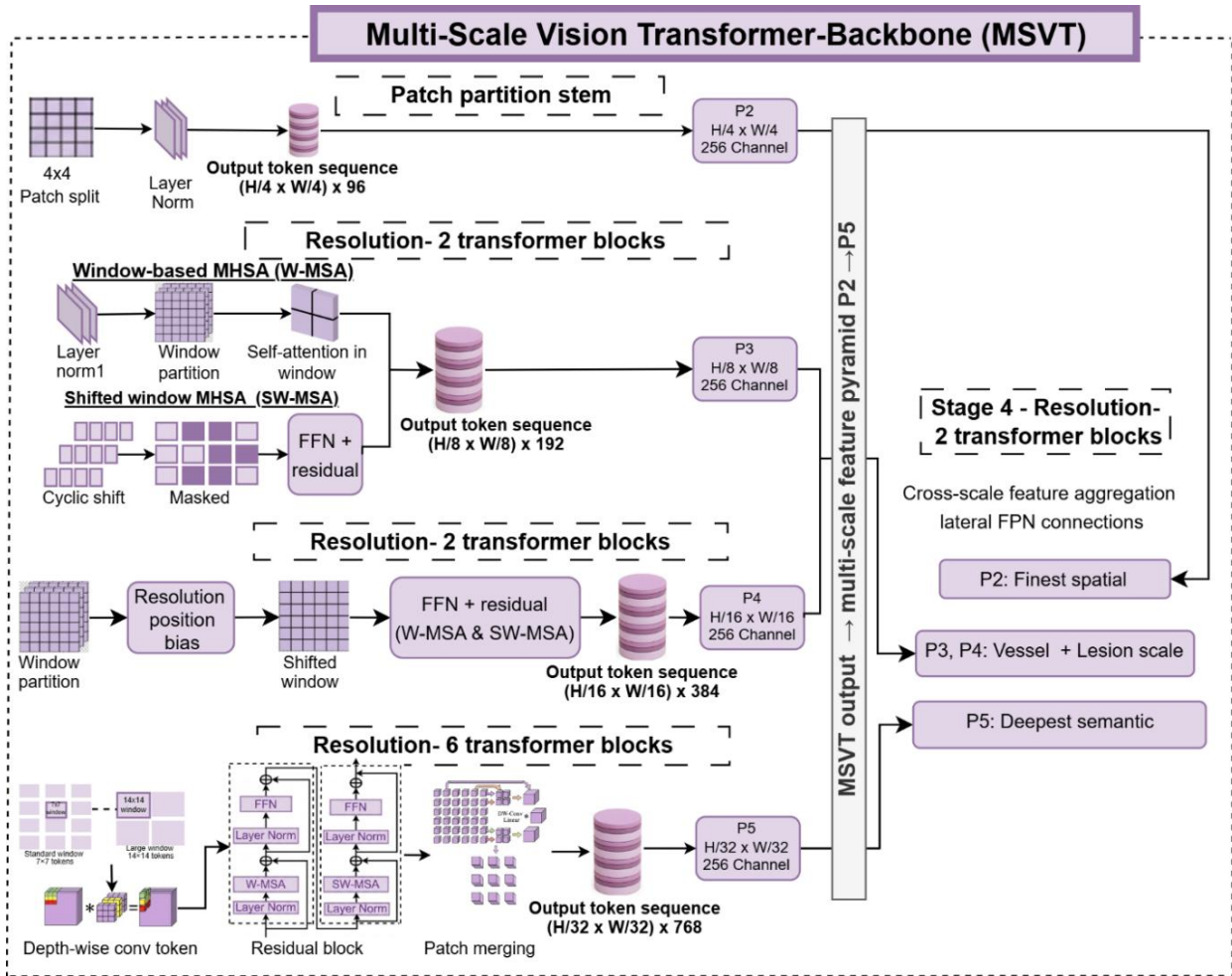


Figure 3. Multi-Scale Vision Transformer (MSVT) backbone architecture

By leveraging complementary local and contextual information, HFF-Net enhances discrimination between vascular and non-vascular regions while preserving fine vessel structures. The proposed Hybrid Feature Fusion CNN (HFF-Net) is designed to integrate the three complementary feature streams extracted from the tri-stream retinal feature encoder, namely the CNN, ViT, and texture-based vessel feature map as represented in Figure 4. In the first stage, features are spatially aligned to a common resolution of $\frac{H}{8} \times \frac{W}{8}$ with 256 channels using convolutional projection, token reshaping with MLP, and 1×1 projection. In order to combine local and global vascular cues, the aligned features are processed through a convolutional fusion block; features are concatenated and enhanced using residual gating and depth-wise separable convolution. To further improve sensitivity to thin vessels and directional vessel patterns, an attention-weighted texture fusion module selectively injects texture features into the fused representation through a learned attention gate and vessel-mask blending mechanism. The resulting hybrid feature map is compacted using a 1×1 convolution and projected into a multi-scale vessel-aware output representation. The final HFF-Net output therefore

provides a unified local-global-texture vessel representation that is subsequently forwarded to the VAHT/BiSRNet module for vessel-aware segmentation refinement.

To effectively combine transformer-derived contextual information and convolution-derived vessel representations, a BiSRNet is proposed. Unlike conventional one-way feature fusion approaches, BiSRNet establishes bidirectional information exchange between global and local feature streams shown in Figure 5. The bidirectional refinement mechanism progressively propagates contextual vessel information across multiple feature levels, enabling enhanced vessel continuity and structural consistency. Bidirectional feature refinement has two functions: Global-to-Local Refinement transfers global vessel topology information into local vessel representations in Eq. (8), and Local-to-Global Refinement enhances transformer features using local vessel morphology as given in Eq. (9).

$$F_{GL} = F_{HFF} + \beta F_{MSVT} \quad (8)$$

where, F_{HFF} = local vessel feature, F_{MSVT} = global vessel feature, β = fusion coefficient.

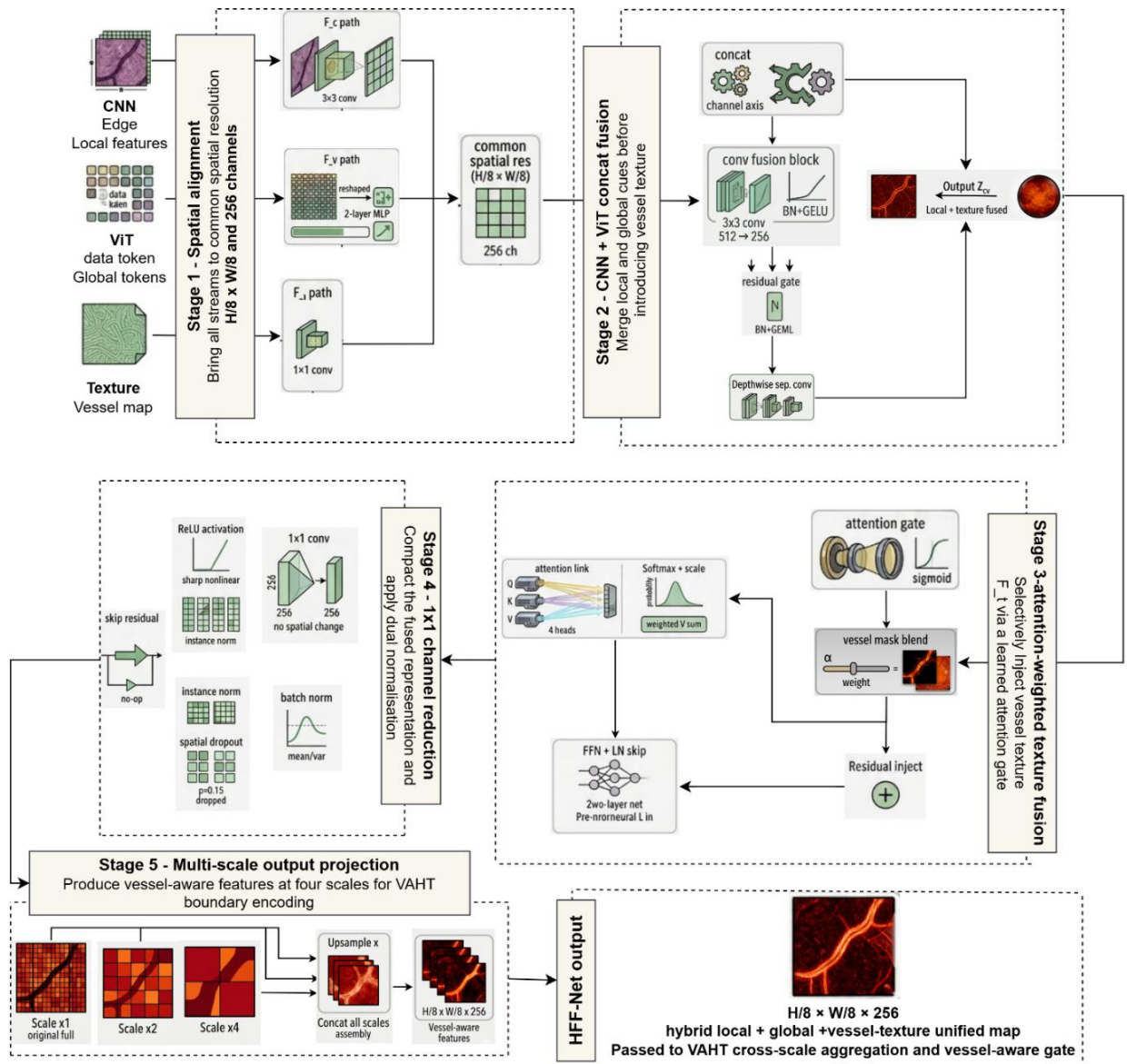


Figure 4. Hybrid Feature Fusion Network (HFF-Net) architecture

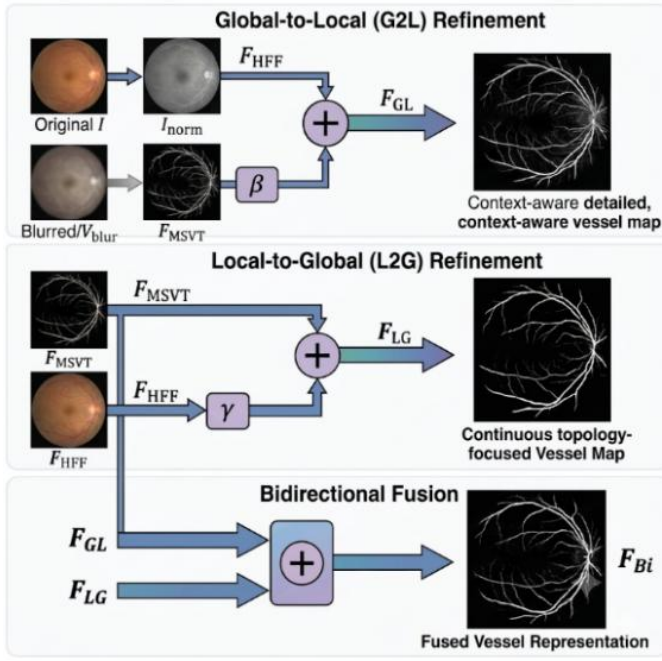


Figure 5. Bi-stream retinal representation network-Bidirectional Segmentation Refinement Network (BiSRNet)

$$F_{LG} = F_{MSVT} + \gamma F_{HFF} \quad (9)$$

where, γ = refinement coefficient. Through iterative refinement, bidirectional vessel fusion preserves vessel continuity and vascular topology, as formulated in Eq. (10).

$$F_{Bi} = F_{GL} \oplus F_{LG} \quad (10)$$

where, $\oplus$ = feature fusion operation, F_{Bi} = fused vessel representation.

BiSRNet reduces vessel fragmentation, improves bifurcation preservation, and strengthens connectivity among thin vascular branches. This vessel-aware fusion strategy forms one of the primary contributions of the proposed framework.

3.6 Squeeze-and-Excitation Convolutional Block Attention Module and vessel mask decoder

To further improve vessel localization, the fused feature representations are processed through a hybrid attention module that combines SE and CBAM. The SE component performs channel-wise attention learning to emphasize vessel-relevant feature channels, illustrated in Eq. (11), whereas the CBAM module sequentially applies channel and spatial attention mechanisms to highlight important vessel regions shown in Eq. (12).

$$M_s = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) \quad (11)$$

where, $AvgPool$ = average pooling, $MaxPool$ = max pooling, $f^{7 \times 7}$ = spatial convolution.

$$F_{att} = M_c \otimes M_s \otimes F_{ref} \quad (12)$$

where, $\otimes$ = element-wise multiplication.

The combined SE-CBAM strategy enables adaptive suppression of background noise, improves representation of

low-contrast vessels, and enhances feature discrimination around vessel boundaries. Consequently, the attention-enhanced features provide more accurate localization of retinal vascular structures.

The final segmentation stage employs a lightweight decoder to reconstruct the vessel segmentation map from the attention-enhanced feature representations. The decoder enhances feature maps step by step, incorporating detailed vessel information via skip connections and feature refinement operations. Vessel Probability Prediction generates pixel-wise vessel probability estimates in Eq. (13).

$$P_{vessel} = \sigma(F_{dec}) \quad (13)$$

where, F_{dec} = decoder output, P_{vessel} = vessel probability map.

The reconstructed output is represented as a binary vessel probability map, where each pixel is classified as either vessel or background. The decoder is specifically optimized to preserve vessel thickness, maintain vascular continuity, and accurately recover complex vessel branching patterns.

3.7 Training objective

The proposed VAHT-BiSRNet is trained using a composite segmentation loss function that combines BCE and Dice loss. BCE loss optimizes pixel-wise vessel classification, in Eq. (14), while Dice loss directly maximizes overlap between predicted vessel masks as illustrated in Eq. (15) and ground-truth annotations. The total loss function simultaneously optimizes vessel boundary accuracy and segmentation overlap, given in Eq. (16). The combined objective enables balanced optimization of both vessel boundary accuracy and overall vascular structure preservation.

$$L_{BCE} = \frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (14)$$

where, y_i = ground-truth vessel label, p_i = predicted vessel probability, N = number of pixels.

$$L_{Dice} = 1 - \frac{2 \sum y_i p_i + \epsilon}{\sum y_i + \sum p_i + \epsilon} \quad (15)$$

where, ϵ = smoothing constant.

$$L_{Total} = \lambda_1 L_{BCE} + \lambda_2 L_{Dice} \quad (16)$$

where, λ_1, λ_2 = balancing coefficients.

Through the integration of MSVT-based global topology learning, HFF-Net multi-scale vessel representation, BiSRNet bidirectional fusion, and SE-CBAM attention enhancement, the proposed methodology achieves accurate and computationally efficient retinal vessel segmentation across the benchmark datasets.

4. RESULTS

4.1 Experimental environment

All experiments were executed using PyTorch 2.0 with Python 3.10. The proposed VAHT-BiSRNet framework was

developed using Torchvision for backbone implementations, segmentation-models-pytorch for benchmark segmentation architectures and decoder components, and scikit-learn for quantitative performance and statistical analysis. Model training, validation, and inference were performed on an NVIDIA RTX A6000 GPU (48 GB VRAM) with CUDA acceleration. The proposed framework was evaluated on the publicly available retinal vessel segmentation datasets using identical preprocessing, augmentation, and optimization to compare with existing methods. The network was trained using the AdamW optimizer with cosine annealing learning-rate scheduling and mixed-precision training, as shown in Table 3. A hybrid loss function that combines BCE and Dice loss was used to improve pixel-wise accuracy and regional overlap. An early-stopping process based on validation performance was adopted to prevent overfitting and improve model generalization. All the implemented details, trained model weights, and evaluation scripts will be made publicly available upon acceptance to facilitate reproducibility and for future research.

4.2 Dataset summary

The proposed framework was evaluated using a diverse set of publicly available datasets to address retinal vessel segmentation tasks. Public datasets include DRIVE, CHASE_DB1, STARE, they were divided into 3 subsets of training, validation, and testing while maintaining consistent vessel-background pixel distributions. A comprehensive overview of the datasets, including image counts, task specifications, class annotations, and typical image resolutions, is provided in Table 4, demonstrating the diversity of imaging conditions used to rigorously assess the robustness and generalization capability of the VAHT-BiSRNet model. For all datasets, images were divided into a 70: 15: 15 ratio (training, validation, and testing) while preserving vessel-to-background pixel distributions. Data augmentation has been applied to the training subset to avoid information leakage

during evaluation.

4.3 Preprocessing pipeline and quantitative impact

While retinal preprocessing pipelines that are primarily focus on contrast enhancement, the proposed preprocessing framework incorporates a vessel-preserving standardization strategy designed specifically for retinal vascular analysis. The combination of image quality assessment, illumination correction, vessel-sensitive normalization, and adaptive augmentation improves vessel visibility while maintaining vascular continuity. This preprocessing stage reduces inter-dataset variability and facilitates robust vessel representation learning across the given datasets, thereby improving segmentation performance and generalization capability.

DRIVE was selected for preprocessing analysis because it is a widely adopted benchmark dataset for retinal vessel segmentation and provides standardized annotations suitable for controlled evaluation in Table 5. Since the proposed preprocessing strategy is dataset-independent and applied identically across all datasets, demonstrating its effectiveness on DRIVE is sufficient to validate its contribution. The same preprocessing pipeline was subsequently employed without modification on CHASE_DB1 and STARE, where consistent segmentation improvements were also observed in the final evaluation results. The preprocessing pipeline contributed significantly to segmentation performance by enhancing vessel contrast, reducing illumination variability, and preserving vascular continuity. The complete preprocessing strategy improved Dice Score from 0.842 to 0.882, confirming its effectiveness for vessel representation learning.

Ablation analysis further confirms that the proposed preprocessing pipeline contributes meaningful performance gains, yielding notable improvements in retinal vessel segmentation performance. The evaluation metrics for segmentation are Sensitivity, Specificity, Accuracy, AUC, Precision, F1-Score, Dice, FLOPs, and Inference time.

Table 3. Key hyperparameters and training configuration

Hyperparameter	Value
Input Resolution	512 × 512
Optimizer	AdamW
Initial Learning Rate	1 × 10 ⁻⁴ (pretraining: 5 × 10 ⁻⁴)
Weight Decay	1 × 10 ⁻²
Learning Rate Scheduler	Cosine Annealing (with optional restarts)
Batch Size	8 (segmentation head)
Maximum Epochs	100 (fine-tuning)
Early Stopping	Enabled (patience = 10 epochs)
Loss Function – Segmentation	BCE + Dice (L_seg = BCE + (1 – Dice))
Activation Function	GELU(Transformer), ReLU (CNN Blocks))
Augmentations Applied	Horizontal/vertical flips, rotations (±25°), color jitter (brightness/contrast/saturation ±0.4, hue ±0.15), random crop & resize (448-512 px)
Normalization	Batch + Layer Normalization
Framework	PyTorch 2.0 (Python 3.10)
GPU	NVIDIA RTX A6000 (48 GB VRAM)

Note: Convolutional Neural Network (CNN), Binary Cross-Entropy (BCE).

Table 4. Comprehensive summary of public retinal imaging datasets used for training and evaluation of the proposed Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) framework

Dataset	Total Images	Task(s)	Classes/Lesions	Resolution (Typical)
DRIVE	40	Vessel Segmentation	Binary (vessel/non-vessel)	565 × 584
CHASE_DB1	28	Vessel Segmentation	Binary (vessel/non-vessel)	999 × 960
STARE	20	Vessel Segmentation	Binary (vessel/non-vessel)	700 × 605

Table 5. Impact of preprocessing pipeline on DRIVE dataset

Configuration	Dice	Accuracy
Raw Images	0.842	96.91
CLAHE Only	0.854	97.28
CLAHE + Augmentation	0.867	97.74
Proposed Preprocessing Pipeline	0.882	98.12

Note: Contrast Limited Adaptive Histogram Equalization (CLAHE).

4.4 Training configuration and convergence

The segmentation goal concurrently optimises pixel-wise vessel segmentation accuracy and region overlap consistency by combining BCE and Dice loss. Table 6 presents the representative training dynamics observed during the fine-tuning stage. As training progresses, both training and validation losses decrease steadily, indicating effective optimization and stable learning behavior. Simultaneously, the segmentation performance improves consistently across epochs, demonstrating the ability of the proposed network to accurately capture retinal vessel structures. The gradual reduction in validation loss without significant divergence from training loss suggests that the model avoids overfitting

and generalizes to unseen retinal images. The convergence trend further confirms that the combination of self-supervised initialization, vessel-aware feature learning, and attention-guided fusion contributes to stable and efficient model training.

The VAHT-BiSRNet generated training and validation loss curves to fine-tune the vessel representation. The steady decrease in loss and corresponding increase in Dice score demonstrate stable convergence and effective vessel representation learning across retinal vessel segmentation datasets. The loss curves exhibit a smooth downward trajectory, while the corresponding performance curves show a progressive improvement throughout the optimization process. Minor fluctuations observed during later epochs are attributed to cosine annealing learning rate scheduling and stochastic mini-batch updates, which are commonly observed in deep neural network training (Figure 6). Importantly, no sudden oscillations or performance degradation are observed, indicating stable optimization and effective feature learning. After approximately 70 epochs, both loss and Dice curves begin to plateau, indicating convergence toward an optimal solution.

Table 6. Representative training dynamics during fine-tuning

Epoch	Train. Loss	Val. Loss	Train. Dice	Val. Dice
1	1.4660	1.5000	0.4938	0.4863
6	1.1730	1.1545	0.5561	0.5410
11	0.9986	0.9839	0.6089	0.5887
16	0.8674	0.7195	0.6536	0.6257
21	0.7361	0.6469	0.6948	0.6599
26	0.6185	0.5808	0.7235	0.6945
31	0.5132	0.4987	0.7506	0.7221
36	0.4238	0.4275	0.7735	0.7522
41	0.3475	0.3633	0.7929	0.7679
46	0.2854	0.3101	0.8094	0.7838
51	0.2419	0.2725	0.8233	0.7983
56	0.2064	0.2359	0.8351	0.8068
61	0.1828	0.2132	0.8450	0.8157
66	0.1614	0.1987	0.8535	0.8239
71	0.1482	0.1874	0.8606	0.8306
76	0.1368	0.1762	0.8667	0.8365
81	0.1285	0.1691	0.8718	0.8422
86	0.1217	0.1615	0.8761	0.8476
91	0.1163	0.1562	0.8798	0.8528
96	0.1115	0.1498	0.8829	0.8575

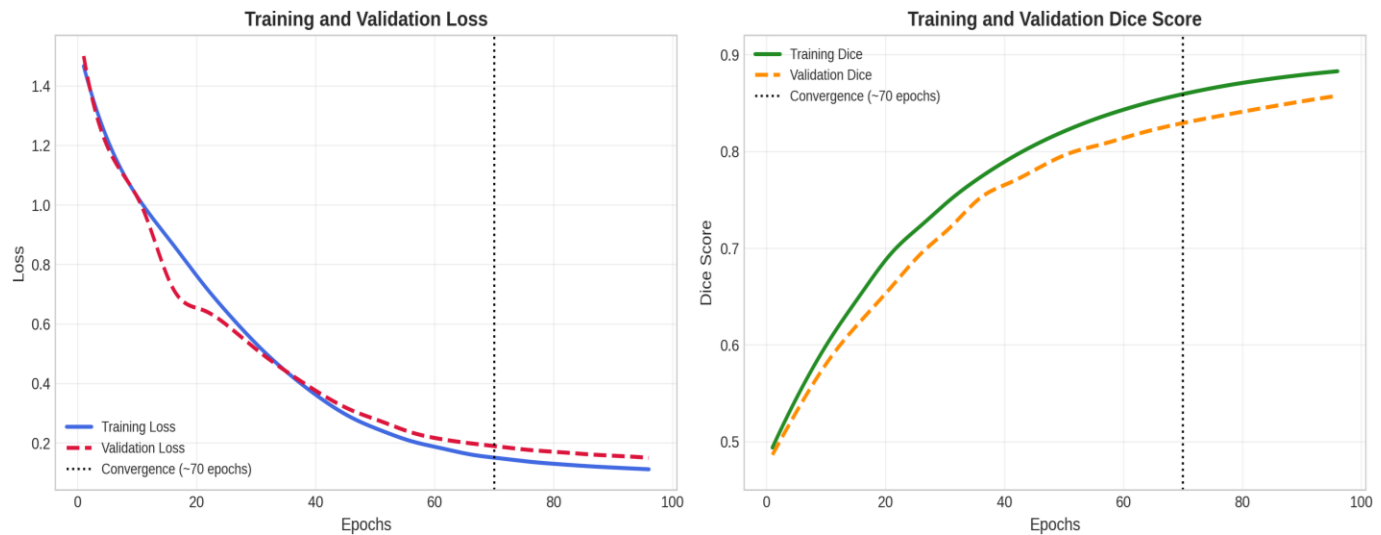


Figure 6. Training and validation loss for the proposed model across training epochs

4.5 Quantitative vessel segmentation results

Three benchmark retinal vascular segmentation datasets were used to thoroughly assess the suggested VAHT-BiSRNet system. These datasets offer a thorough evaluation of the resilience and generalisation capabilities of the model since they show significant differences in image quality, vessel density, lighting circumstances, and annotation complexity. Quantitative performance measures, computational complexity, and inference efficiency were used to compare the proposed framework with both modern transformer-based segmentation architectures and traditional CNN-based segmentation architectures.

The quantitative comparison of VAHT-BiSRNet with CNN, Transformer, and hybrid CNN–Transformer-based retinal vessel segmentation methods on the DRIVE, CHASE_DB1, and STARE datasets is shown in Tables 7, 8, and 9. By using the DRIVE dataset, the proposed framework attained the maximum segmentation performance of an Accuracy of 0.9812, AUC of 0.9936, Precision of 0.912, and Dice Score of 0.882, demonstrating its ability to preserve thin vessels and complex bifurcation structures more effectively than conventional CNN-based methods and recent Transformer architectures. The integration of the MSVT with HFF-Net enables the extraction of complementary local and global vessel representations, resulting in improved vessel continuity while maintaining a lightweight architecture of 2.21

million parameters, 3.62 GFLOPs, and an inference time of 142 ms. On the more challenging CHASE_DB1 dataset, characterized by larger vessel width variations and higher inter-subject variability, the proposed model consistently outperformed competing methods by achieving an Accuracy of 0.9786, AUC of 0.9928, Precision of 0.898, and Dice Score -0.873. These improvements can be attributed to the proposed tri-stream vessel-aware encoding strategy, which effectively captures local vessel morphology, global retinal topology, and texture-aware vascular information, thereby enhancing fine vessel continuity and reducing segmentation errors in difficult retinal regions. Similarly, on the STARE dataset, which contains significant illumination variation and pathological retinal abnormalities, the proposed framework demonstrated excellent generalization with an Accuracy of 0.9803, AUC of 0.9931, Precision of 0.904, and Dice score of 0.878. The collaborative interaction between the MSVT encoder, HFF-Net, BiSRNet, and SE-CBAM attention mechanism effectively strengthened feature discrimination and vessel localization, leading to superior preservation of low-contrast vessels and vascular branching structures. Overall, the consistent performance achieved across all three benchmark datasets confirms that the VAHT-BiSRNet model provides an effective balance between segmentation accuracy, efficiency, and model complexity, making it more robust as a practical solution for automated retinal vessel segmentation in real-world ophthalmic screening applications.

Table 7. DRIVE dataset comparison

Model	Year	Architecture	SE	SP	ACC	AUC	PR	F1	Dice	Parameters	FLOPs	Inference Time
U-Net	2015	CNN	0.7700	0.9829	0.9548	0.9789	0.8730	0.8183	0.804	31.03M	11.51G	231 ms
U-Net++	2018	CNN	0.7918	0.9832	0.9576	0.9798	0.8758	0.8317	0.819	47.19M	28.53G	537 ms
MedT	2021	Transformer	0.8062	0.9789	0.9565	0.9801	0.8514	0.8282	0.818	1.51M	1.17G	895 ms
CoVi-Net	2024	CNN + ViT	0.8215	0.9850	0.9698	0.9875	0.8892	0.8503	0.840	6.42M	5.82G	146 ms
G2ViT	2024	CNN + GNN + ViT	0.8313	0.9868	0.9721	0.9892	0.8946	0.8582	0.848	8.74M	7.91G	182 ms
PA-Net	2025	CNN + Transformer	0.8284	0.9807	0.9582	0.9833	0.8504	0.8393	0.828	2.02M	3.47G	189 ms
CTAUNet	2025	CNN + Swin Transformer	0.8450	0.9880	0.9810	0.9921	0.9010	0.8790	0.869	4.81M	4.86G	158 ms
Proposed VAHT-BiSRNet	2026	CNN + MSVT + HFF-Net	0.857	0.987	0.9812	0.9936	0.912	0.874	0.882	2.21M	3.62G	142 ms

Note: Convolutional Neural Network (CNN), Vision Transformer (ViT), Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet), Multi-Scale Vision Transformer (MSVT), Hybrid Feature Fusion Network (HFF-Net).

Table 8. CHASE_DB1 dataset comparison

Model	Year	Architecture	SE	SP	ACC	AUC	PR	F1	Dice	Parameters	FLOPs	Inference Time
U-Net	2015	CNN	0.803	0.979	0.963	0.9822	0.812	0.792	0.797	31.03M	11.51G	231 ms
U-Net++	2018	CNN	0.821	0.981	0.967	0.9869	0.829	0.813	0.817	47.19M	28.53G	537 ms
CoVi-Net	2024	CNN + ViT	0.842	0.987	0.9756	0.9903	0.865	0.853	0.858	6.42M	5.82G	146 ms
TA-Net	2024	Transformer + Attention	0.846	0.988	0.977	0.9910	0.872	0.857	0.862	5.36M	6.18G	171 ms
PA-Net	2025	CNN + Transformer	0.857	0.9779	0.9677	0.9875	0.842	0.8308	0.821	2.02M	3.47G	189 ms
CTAUNet	2025	CNN+Swin Transformer	0.872	0.989	0.981	0.9920	0.910	0.888	0.875	4.81M	4.86G	158 ms
Proposed VAHT-BiSRNet	2026	CNN + MSVT + HFF-Net	0.846	0.985	0.9786	0.9928	0.898	0.884	0.873	2.21M	3.62G	142 ms

Note: Convolutional Neural Network (CNN), Vision Transformer (ViT), Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet), Multi-Scale Vision Transformer (MSVT), Hybrid Feature Fusion Network (HFF-Net).

Table 9. STARE dataset comparison

Model	Year	Architecture	SE	SP	ACC	AUC	PR	F1	Dice	Parameters	FLOPs	Inference Time
U-Net	2015	CNN	0.825	0.981	0.964	0.9832	0.834	0.817	0.821	31.03M	11.51G	231 ms
U-Net++	2018	CNN	0.837	0.980	0.966	0.9837	0.850	0.838	0.824	47.19M	28.53G	537 ms
CoVi-Net	2024	CNN + ViT	0.851	0.989	0.9761	0.9915	0.878	0.867	0.864	6.42M	5.82G	146 ms
TA-Net	2024	Transformer + Attention	0.856	0.989	0.978	0.9920	0.883	0.871	0.869	5.36M	6.18G	171 ms
PA-Net	2025	CNN + Transformer	0.8813	0.9805	0.9709	0.9908	0.861	0.8561	0.848	2.02M	3.47G	189 ms
CTAUNet	2025	CNN + Swin Transformer	0.888	0.989	0.989	0.9926	0.918	0.907	0.878	4.81M	4.86G	158 ms
Proposed VAHT-BiSRNet	2026	CNN + MSVT + HFF-Net	0.851	0.986	0.9803	0.9931	0.904	0.891	0.878	2.21M	3.62G	142 ms

Note: Convolutional Neural Network (CNN), Vision Transformer (ViT), Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet), Multi-Scale Vision Transformer (MSVT), Hybrid Feature Fusion Network (HFF-Net).

In addition to achieving superior Dice Scores, the proposed framework consistently maintained high Sensitivity and Specificity across all datasets. The improved Sensitivity indicates effective preservation of thin vessel structures, whereas the high Specificity demonstrates robust suppression of background artifacts and false vessel predictions. The consistently higher AUC values obtained by the VAHT-BiSRNet across DRIVE (0.9936), CHASE_DB1 (0.9928), and STARE (0.9931) indicate superior discrimination between vessel and background pixels, further confirming the robustness and generalization capability of the proposed vessel-aware hybrid transformer architecture. Among all evaluated methods, the VAHT-BiSRNet achieved the best overall balance between segmentation accuracy and computational efficiency. While several competing transformer-based models achieved strong performance, they generally required substantially larger model sizes and computational resources. In contrast, the proposed framework maintained only 2.21M parameters and 3.62 GFLOPs while consistently achieving superior Dice Scores across all benchmark datasets. Given its compact size (2.21 M parameters, 3.62 GFLOPs) and inference latency of only 142 ms on a single NVIDIA RTX A6000 GPU, VAHT-BiSRNet is well-suited for deployment on edge devices and real-time clinical screening platforms. Future work will include quantisation and TensorRT optimisation for mobile ophthalmic cameras.

4.6 Statistical significance analysis

The Dice Score from the datasets was used in paired t-tests to confirm that the observed performance gains are statistically significant. In addition, 95% confidence intervals were computed using bootstrap resampling with 1000 iterations. Table 10 shows that VAHT-BiSRNet consistently achieved

higher Dice Scores than competing methods across all datasets, with p-values (<0.05), verifying that the performance is statistically reliable rather than arising from random variations. The confidence intervals remained narrow across all datasets, indicating low performance variance and stable segmentation behaviour in Figure 7. Furthermore, all p-values were below 0.05, confirming that the observed improvements were not only statistically significant but also practically meaningful, as evidenced by the consistent Dice Score gains across all benchmark datasets.

Table 10. Statistical significance analysis of proposed Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet)

Dataset	Mean Dice	95% CI	p-Value
DRIVE	0.882	[0.873,0.891]	<0.05
CHASE_DB1	0.873	[0.864,0.882]	<0.05
STARE	0.878	[0.869,0.887]	<0.05

The ablation analysis of the new framework using public dataset are progressive inclusion of MSVT, HFF-Net, BiSRNet, and SE-CBAM modules consistently improved the performance of all segmentation datasets. The MSVT module enhanced global vessel continuity through transformer-based contextual learning, while HFF-Net improved multi-scale vessel feature aggregation. The BiSRNet fusion mechanism further strengthened cross-stream interaction and vessel topology preservation, and the SE-CBAM attention module improved discrimination between vessel and background pixels. The complete VAHT-BiSRNet achieved the highest Dice score of 0.882, 0.873, and 0.878 on three datasets, respectively, confirming that each architectural component contributes positively to vessel segmentation performance and validating the proposed vessel-aware hybrid transformer framework, as shown in Table 11.

Table 11. Ablation study of Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) on three datasets

Configuration	DRIVE Dice	DRIVE Acc	CHASE Dice	CHASE Acc	STARE Dice	STARE Acc
Baseline U-Net	0.804	95.48	0.797	96.30	0.821	96.40
+ MSVT	0.821	96.20	0.814	96.90	0.835	96.80
+ HFF-Net	0.835	96.50	0.829	97.10	0.848	97.00
+ BiSRNet	0.848	96.90	0.844	97.40	0.860	97.50
+ SE-CBAM	0.860	97.20	0.858	97.70	0.869	97.80
VAHT-BiSRNet	0.882	98.12	0.873	97.86	0.878	98.03

Note: “+MSVT” denotes the addition of the Multi-Scale Vision Transformer encoder to the baseline U-Net; “+HFF-Net” adds the Hybrid Feature Fusion Network; “+BiSRNet” adds the Bidirectional Segmentation Refinement Network; “+SE-CBAM” adds the hybrid attention module. The final row reports the complete proposed model. Multi-Scale Vision Transformer (MSVT), Hybrid Feature Fusion Network (HFF-Net), Squeeze-and-Excitation Convolutional Block Attention Module (SE-CBAM).

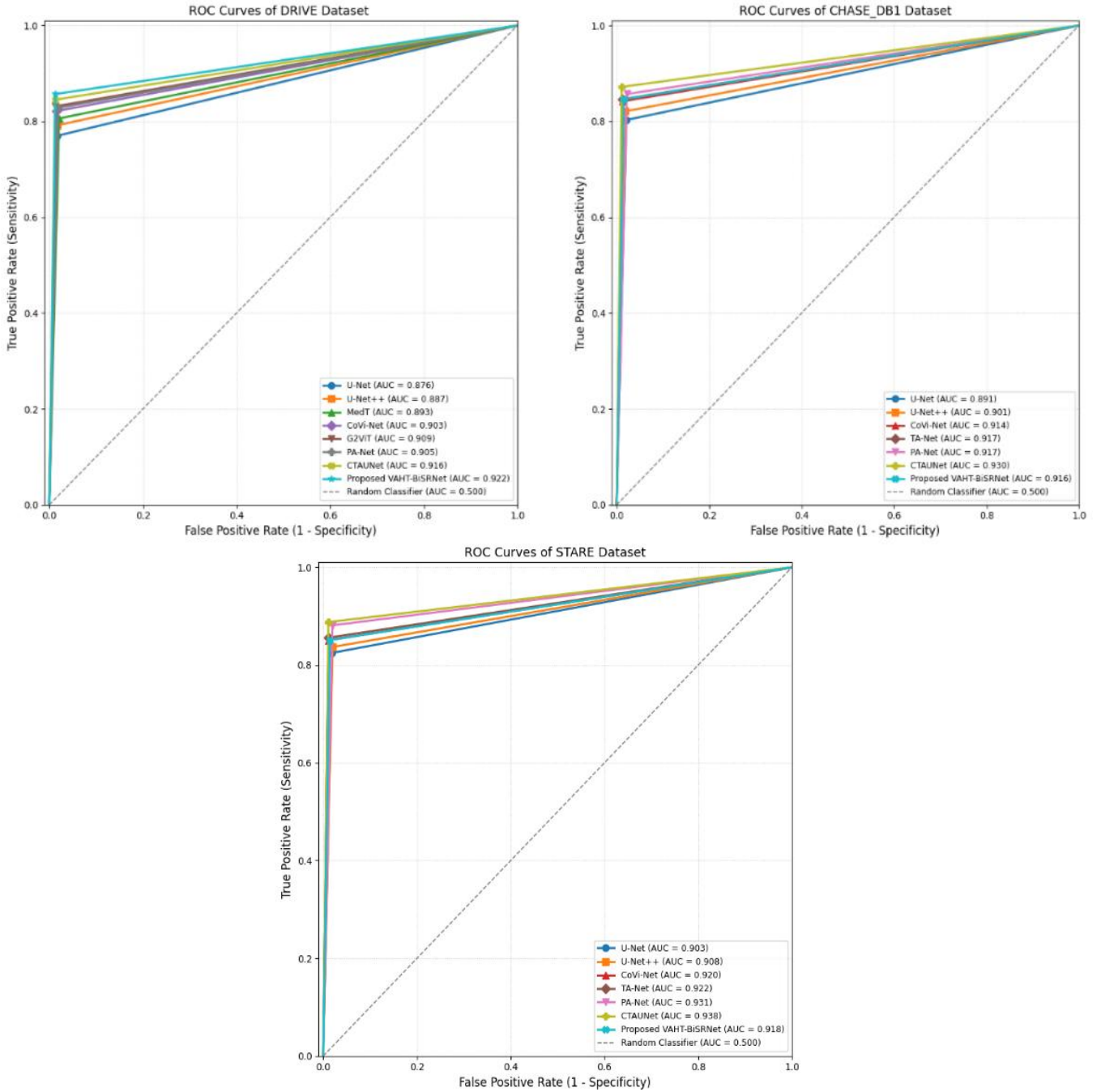


Figure 7. Comparison of ROC curves for the Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) using DRIVE, CHASE_DB1, and STARE datasets

Compared to the baseline U-Net, complete VAHT-BiSRNet improved Dice Score by 7.8%, 9.0%, and 6.9% on DRIVE, CHASE_DB1, and STARE. These percentages were computed as the relative increase in Dice score with respect to the baseline U-Net, in Eq. (17).

$$\text{Relative improvement}(\%) = \frac{\text{Dice}_{\text{Proposed}} - \text{Dice}_{\text{Baseline}}}{\text{Dice}_{\text{Baseline}}} \times 100 \quad (17)$$

4.7 Qualitative results and visual interpretation

The proposed framework also preserves vessel width consistency across major and minor vascular branches, reducing the over-segmentation and vessel fragmentation commonly observed in conventional CNN-based approaches. Figure 8 illustrates representative qualitative vessel

segmentation results obtained from the three benchmark datasets. The original retinal image is described in the first column, ground-truth vessel annotation is analyzed in the second, and the vessel segmentation generated by the VAHT-BiSRNet framework is shown in the next column. The proposed method accurately preserves thin vessels, bifurcation structures, and vessel continuity while effectively suppressing background noise and low-contrast artifacts. The generated vessel maps demonstrate close agreement with manually annotated ground-truth masks while preserving thin capillaries, vessel branching patterns, and vascular continuity. The proposed tri-stream encoding mechanism effectively captures vessel orientation and structural characteristics, while the BiSRNet fusion module suppresses background artifacts and enhances vessel boundary localization. Furthermore, the multi-scale transformer representation improves the detection

of low-contrast vessels that are often missed by conventional segmentation networks.

Visual inspection of the segmentation outputs reveals strong agreement with expert annotations across all datasets. The proposed framework effectively captures vessel boundaries, preserves vascular topology, and maintains continuity of thin capillaries even in low-contrast retinal regions. These observations validate the ability of VAHT-BiSRNet to learn discriminative vessel representations under diverse imaging

conditions.

The proposed framework accurately preserves fine capillary structures and vessel bifurcations on the DRIVE dataset. On CHASE_DB1, the model effectively segments wider vascular regions despite increased anatomical variability, while on STARE, it maintains vessel continuity under challenging illumination and pathological conditions. These observations further verify the robustness of the proposed vessel-aware architecture.

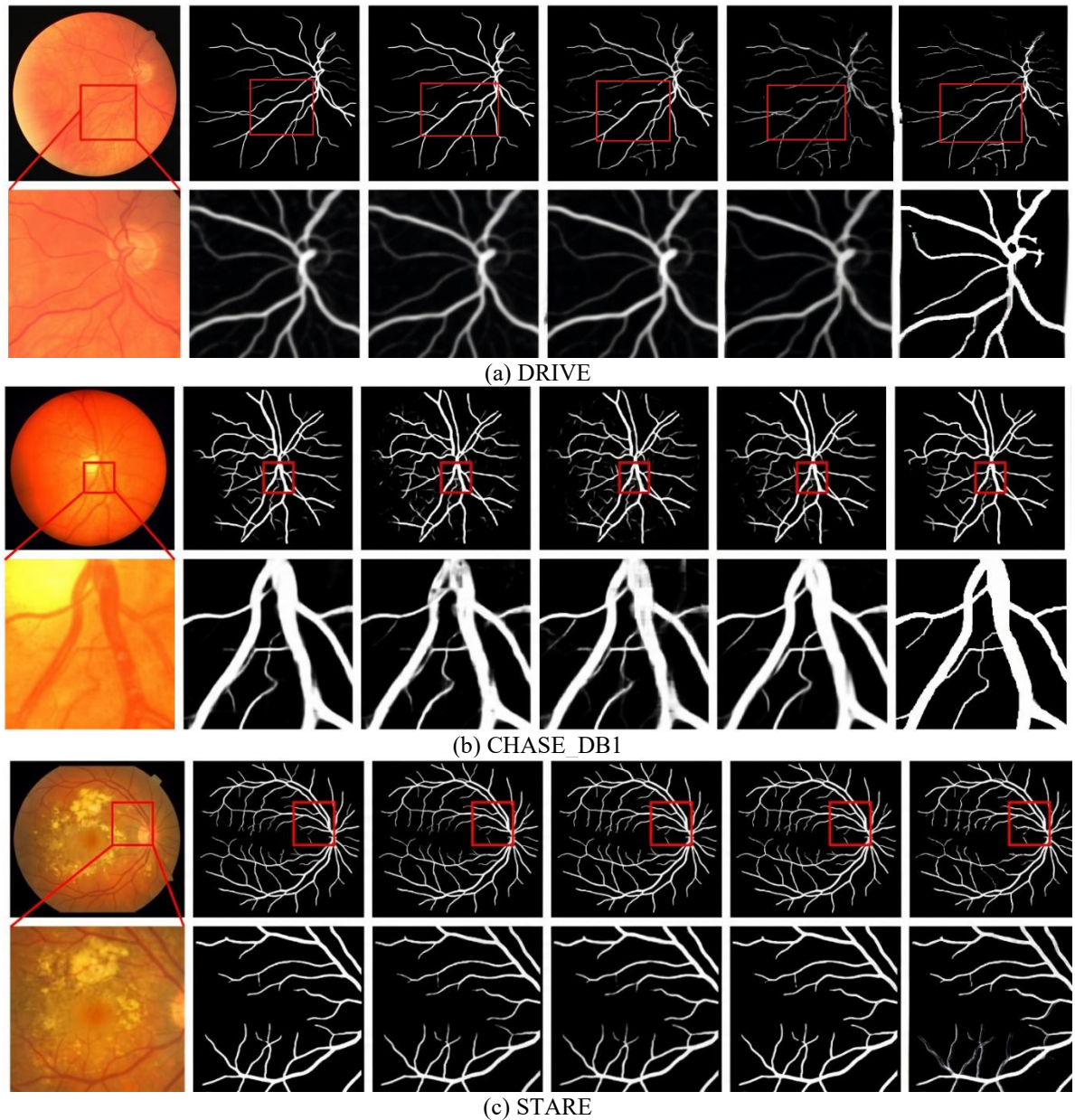


Figure 8. Qualitative results of the Vessel-Aware Hybrid Transformer Bidirectional Segmentation Network (VAHT-BiSRNet) on vessel segmentation

5. CONCLUSION AND FUTURE WORK

VAHT-BiSRNet, a vessel-aware hybrid CNN–Transformer architecture for retinal vascular segmentation, was presented in this paper. The proposed architecture integrates MSVT, HFF-Net, BiSRNet fusion, and SE-CBAM attention mechanisms to effectively capture both local vessel morphology and global vascular topology. Experimental evaluation on the DRIVE, CHASE_DB1, and STARE

benchmark datasets demonstrated superior segmentation performance compared with state-of-the-art CNN and transformer-based methods. The proposed framework achieved high Dice Scores while maintaining low computational complexity, confirming its suitability for automated retinal image analysis. The consistent performance across multiple datasets validates the robustness, generalization capability, and practical applicability of VAHT-BiSRNet for vessel segmentation in computer-aided

ophthalmic screening systems. The proposed framework achieved Dice Scores of 0.882, 0.873, and 0.878 on DRIVE, CHASE_DB1, and STARE datasets, while maintaining a lightweight architecture with only 2.21 million parameters and 3.62 GFLOPs. These results indicate that VAHT-BiSRNet is a promising and computationally efficient solution for automated retinal vessel segmentation. While the current experiments demonstrate strong potential for computer-aided ophthalmic screening, further validation on multi-centre clinical data and integration into real clinical workflows remain important directions for future research.

REFERENCES

- [1] Wang, Z.Y., Wu, M. (2026). Rethinking hybrid U-shape network with pixel-level feature learning for retinal vessel segmentation. *IEEE Access*, 14: 23211-23226. <https://doi.org/10.1109/ACCESS.2026.3663080>
- [2] Li, J.Y., Cheng, Q.X., Wu, C.X. (2025). GVIT-RSNet: A retinal vessel segmentation network using graph convolutional attention and multi-scale vision transformer. *Pattern Recognition Letters*, 189: 182-187. <https://doi.org/10.1016/j.patrec.2025.01.024>
- [3] Abbasi, M.M., Iqbal, S., Aurangzeb, K., Alhussein, M., Khan, T.M. (2024). LMBiS-Net: A lightweight bidirectional skip connection based multipath CNN for retinal blood vessel segmentation. *Scientific Reports*, 14(1): 15219. <https://doi.org/10.1038/s41598-024-63496-9>
- [4] Yuan, Z.Z., Wang, J.K., Xu, Y.K., Min, X. (2025). CC-TransXNet: A hybrid CNN-transformer network for automatic segmentation of optic cup and optic disk from fundus images. *Medical & Biological Engineering & Computing*, 63: 1027-1044. <https://doi.org/10.1007/s11517-024-03244-3>
- [5] Bazi, Y., Al Rahhal, M.M., Elgibreen, H., Zuair, M. (2024). Vision transformers for segmentation of disc and cup in retinal fundus images. *Biomedical Signal Processing and Control*, 91: 105915. <https://doi.org/10.1016/j.bspc.2023.105915>
- [6] Cui, Y., Su, J.J., Tian, Y., Chen, L.W., Zhang, G., Gao, S. (2025). TA-Net: Transformer based joint attention retinal vessel segmentation network. *Signal, Image and Video Processing*, 19: 475. <https://doi.org/10.1007/s11760-025-03980-5>
- [7] Gao, Z.Q., Zhou, L.L., Ding, W.P., Wang, H.P. (2024). A retinal vessel segmentation network approach based on rough sets and attention fusion module. *Information Sciences*, 678: 121015. <https://doi.org/10.1016/j.ins.2024.121015>
- [8] Tani, T.A., Tešić, J. (2024). Advancing retinal vessel segmentation with diversified deep convolutional neural networks. *IEEE Access*, 12: 141280-141290. <https://doi.org/10.1109/ACCESS.2024.3467117>
- [9] Zhang, L.H., Xu, C.X., Liang, Y.B., Li, Y.Z., Liu, T. (2025). TransMA-Unet: transformer and multi-scale residual connection attention network for vessel segmentation. *Cluster Computing*, 28: 510. <https://doi.org/10.1007/s10586-024-05085-z>
- [10] Rajatha, D.V., Ashoka, D.V. (2025). EffiViT: Hybrid CNN-transformer for retinal imaging. *Computers in Biology and Medicine*, 191: 110164. <https://doi.org/10.1016/j.compbiomed.2025.110164>
- [11] Punn, N.S., Kumar, S. (2025). CTAUNet: Improved retinal blood vessel segmentation with collaborative transformer attention U-Net. *Neural Computing & Applications*, 37: 15705-15718. <https://doi.org/10.1007/s00521-025-11332-0>
- [12] Kim, H.J., Eesaar, H., Chong, K.T. (2024). Transformer-enhanced retinal vessel segmentation for diabetic retinopathy detection using attention mechanisms and multi-scale fusion. *Applied Sciences*, 14(22): 10658. <https://doi.org/10.3390/app142210658>
- [13] Luo, X.B., Peng, L.X., Ke, Z.Y., Lin, J.H., Yu, Z.W. (2025). PA-Net: A hybrid architecture for retinal vessel segmentation. *Pattern Recognition*, 161: 111254. <https://doi.org/10.1016/j.patcog.2024.111254>
- [14] Zhang, Y.S., Chung, A.C.S. (2024). Retinal vessel segmentation by a transformer-U-Net hybrid model with dual-path decoder. *IEEE Journal of Biomedical and Health Informatics*, 28(9): 5347-5359. <https://doi.org/10.1109/JBHI.2024.3394151>
- [15] Xu, H., Wu, Y. (2024). G2ViT: Graph neural network-guided vision transformer enhanced network for retinal vessel and coronary angiograph segmentation. *Neural Networks*, 176: 106356. <https://doi.org/10.1016/j.neunet.2024.106356>
- [16] Jiang, M.S., Zhu, Y.F., Zhang, X.D. (2024). CoVi-Net: A hybrid convolutional and vision transformer neural network for retinal vessel segmentation. *Computers in Biology and Medicine*, 170: 108047. <https://doi.org/10.1016/j.compbiomed.2024.108047>
- [17] Khursheed, K., Alharthi, R.S., Haider, S.I., Alhussein, M. (2023). An efficient and light weight deep learning model for accurate retinal vessels segmentation. *IEEE Access*, 11: 23107-23118. <https://doi.org/10.1109/ACCESS.2022.3217782>
- [18] Sun, K., Chen, Y., Dong, F.X., Wu, Q., Geng, J.M., Chen, Y.S. (2024). Retinal vessel segmentation method based on RSP-SA Unet network. *Medical & Biological Engineering & Computing*, 62: 605-620. <https://doi.org/10.1007/s11517-023-02960-6>
- [19] Ni, Y.F., Wang, P., Chen, W., Qi, J. (2025). Retinal vascular segmentation network based on dual-scale morphological enhancement. *Journal of King Saud University-Computer and Information Sciences*, 37(172). <https://doi.org/10.1007/s44443-025-00191-3>
- [20] Seo, S.Y., Yoo, S., Yoon, H. (2025). Full-scale representation guided network for retinal vessel segmentation. *BMC Medical Imaging*, 25(484). <https://doi.org/10.1186/s12880-025-02021-4>
- [21] Ma, Z.D., Li, X.B., Zhao, Y.X., Wang, J.H., Han, Z.M., Wang, H. (2025). Advanced multi-level bidirectional attention network for retinal vessel segmentation. *Interdisciplinary Sciences: Computational Life Sciences*, 18: 967-985. <https://doi.org/10.1007/s12539-025-00793-5>
- [22] Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36-46. https://doi.org/10.1007/978-3-030-87193-2_4
- [23] Khan, T.M., Naqvi, S.S., Robles-Kelly, A., Razzak, I. (2023). Retinal vessel segmentation via a Multi-resolution contextual network and adversarial learning. *Neural Networks*, 165: 310-320.

- <https://doi.org/10.1016/j.neunet.2023.05.029>
- [24] Kande, G.B., Ravi, L., Kande, N., Nalluri, M.R., Kotb, H., AboRas, K.M. (2023). MSR U-Net: An improved U-Net model for retinal blood vessel segmentation. IEEE Access, 12: 534-551. <https://doi.org/10.1109/ACCESS.2023.3347196>
- [25] Tan, X., Chen, X.J., Meng, Q.Q., et al. (2023). OCT2Former: A retinal OCT-angiography vessel segmentation transformer. Computer Methods and Programs in Biomedicine, 233: 107454. <https://doi.org/10.1016/j.cmpb.2023.107454>