Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SRD U-Net: Spatial Refined Depthwise U-Net forRetinal Vessel Segmentation

Srinivas Rangu

srigana.57@gmail.com

Central University of Karnataka

Nagaraj Yamanakkanavar

Central University of Karnataka

Narasimhadhan A V

National Institute of Technology Karnataka

Research Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11327124/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SRD U-Net: Spatial Refined Depthwise U-Net for Retinal Vessel Segmentation

Srinivas Rangu[1*†], Nagaraj Yamanakkanavar[1†] and Narasimhadhan A V[2†]

[1*]Department of E.C.E, Central University of karnataka, Kadaganchi, Kalaburagi, 585367, karnataka, India.
[2]Department of E.C.E, National Institute of Technology, Surathkal, Mangalore, 575025, Karnataka, India.

*Corresponding author(s). E-mail(s): srigana.57@gmail.com;
Contributing authors: nagrajpy@cuk.ac.in; nagraj.p.y@gmail.com;
†These authors contributed equally to this work.

Abstract

Precise pixel level categorization of retinal vessels requires both deep contextual awareness and distortion free spatial representation. Traditional encoder-decoder models often suffer from excessive parameter counts and an inability to prioritize significant feature spaces selectively. Consequently sub optimal boundaries and noise propagation are presents in the retinal vascular fundus images. To address these limitations, this paper introduces the contrast-aware retinal preprocessor block(CARP) and spatial refined depthwise U-Net (SRD U-Net). The CARP serves as a retinal image preprocess and augmentation module that enhances contrast and sharpness. The CARP uses contrast-limited adaptive histogram equalization and TensorFlow-based luma?chroma sharpening to generate preprocessed datasets for segmentation tasks. The SRD U-Net is a lightweight highly discriminative segmentation network with three blocks including conventional filter depthwise unit (CFDU) encoder, channel attention bottleneck(CAB) and CFDU decoder. . The core of our proposed architecture lies in a novel hybrid feature-extraction sequence, CFDU. The CFDU block initially captures primitive spatial patterns through a standard 2D convolution. The convolution output transitioning to a manual depthwise separable configuration, decoupling spatial filtering from channel mixing to dramatically compress the network parameter. Furthermore, we combat feature redundancy in the latent space by integrating a channel attention bottleneck(CAB). The CAB block aggregates global context

through spatial pooling to dynamically compute non-linear inter channel dependencies. The bottleneck attention mechanism explicitly recalibrates channel-wise responses with highly informative feature maps. In addition, the bottleneck systematically suppresses attention, filtering out irrelevant background noise. Furthermore, we evaluated our framework with DRIVE, HRF, STARE, CHASE and FIVES datasets. The performance metrics used for assessment of proposed framework are accuracy, sensitivity, F1 score, recall and precision. On average across all datasets, our proposed method afforded improvements of approximately 2 % among all metrics compared to the lowest-performing baseline. Finally, our model achieves comparable accuracy with significantly lower computational overhead, featuring reduced parameter size and faster execution speeds. This portrays our approach as a highly efficient alternative for resource-constrained environments.

Keywords: Retinal Vessel Segmentation, Multi-Scale Information, Deep Learning, Attention Mechanisms.

1 Introduction

Eye is one of the most vital organs in human body. Every year, numerous people globally become blind due to different types of ophthalmic diseases. Generally, it is estimated that around 26 % people worldwide affected from eye and vision impairments [1]. Changes in retinal vascular morphology can lead to ophthalmic and cardiovascular issues, making accurate vessel segmentation significant for diagnose related diseases. However, segmentation task remains challenging due to the wide range of retinal vessel variations and the low contrast between vessels and the background. Zhenchao *et al.* [2] developed an MF2ResU-Net to refine small vessels and segment retinal vessels. He *et al.* [3] introduced an algorithm that extracts multi-scale vascular features and performs segmentation in an end-to-end manner to improve the low contrast of the original image. Li *et al.* [4] proposed a model to enhance feature extraction that uses attention U-Net as the baseline network replaces all conventional convolutions in the encoding-decoding process with dynamic convolution decomposition. Cai *et al.* [5] developed a multidimensional collaborative attention encoder to strengthen the capacity for comprehensive detailing, ensure that sophisticated features are systematically sustained during the segmentation of tiny vessels. Li *et al.* [6] introduced a GViT-RSNet, an efficient up-convolution and a graph convolutional attention module that robustly enhances feature maps. Furthermore, a gated fusion module utilized to generate refined segmentation outputs. Zhang *et al.* [7] used a multi-scale feature input module to acquire multilevel semantics information from the source image. Jiang *et al.* [8] introduced the dynamic convolution and attention mechanisms for the retinal vessel segmentation task. Khawaja *et al.* [9] found that lesion quantity and characteristics on the retinal surface play a crucial role in describe disease severity and diagnostic accuracy procedures. Subsequently, the efficacy of an automated screening system for thorough examination link on the accuracy of blood vessel segmentation,retinal lesion identification, and optical cup delineation. Zhao *et*

al. [10] proposed a modified encoder-decoder framework that uses a pyramid pooling module to facilitate the integration of multi-scale information which enhances comprehensive global context. Mu *et al.* [11] used an attention-driven residual U-Net model for the precise segmentation of cerebral aneurysm . This approach combines specialized preprocessing and postprocessing techniques, which enables exceptional precision in automated identification of intracranial aneurysms. Long *et al.* [12] introduced the fully convolutional network (FCN),which has revolutionized semantic segmentation through convolutional neural networks. Numerous segmentation models following an encoder-decoder architecture have emerged. Ronneberger *et al.* [13] introduced the U-Net network, an extension of FCN, that integrates additional upsampling procedures and skip connections to incorporate contextual cues, enabling the capture of more comprehensive scale details than FCN. Jin *et al.* [14] developed a U-shaped architecture for retinal vessel segmentation that incorporates deformable convolutions to extract nuanced local features and integrates low- and high-level features to improve spatial identification accuracy. Poynton *et al.* [15] proposed that the image is converted from RGB to a luminance-chrominance color space to isolate spatial frequencies. Zuiderveld *et al.* [16] introduced the clip limit, which constrains the vertical extent of individual histograms and facilitates even dispersion of truncated pixel values across the entire intensity span, namely contrast-limited adaptive histogram equalization (CLAHE). .Although these methods extract some contextual information, the vascular structures in fundus images occupy only a small portion of the image, making it difficult for network models to perceive blood vessels accurately. Secondly, ordinary skip connections at each stage often directly combine local information, introducing excessive irrelevant background noise and making it difficult to distinguish retinal vessels from noise, especially capillaries. Theoretically, the high-level features extracted from deeper stages contain rich semantic information but low resolution. In comparison the low-level features from shallower stages have abundant spatial details but lack global semantic context. Additionally, it is crucial to extract semantic representations while intelligently suppressing artifacts and noise. To address the above problems, we propose a hybrid convolutional method, which integrates depthwise convolution and attention mechanisms to achieve accurate retinal vessel segmentation. Firstly, we designed a new encoding module, called the conventional filter depthwise unit (CFDU), to extract vessel features. The CFDU module effectively captures vessel irregularities using depthwise convolutions, deepening feature extraction, and reducing feature loss from convolutional operations. Secondly, we introduce a channel attention bottleneck (CAB) module to dynamically recalibrate channel-wise feature dependencies by amplifying highly informative feature maps and suppressing irrelevant noise. Finally, the proposed framework leverages deep supervision to fully exploit feature information at multiple levels, thereby improving the networks segmentation and generalization capabilities, including through attention mechanisms and deeper feature fusion. Our main work can be outlined as follows:

- We proposed the CARP block functions as a preprocessing and augmentation unit for retinal images, employing CLAHE and TensorFlow-driven luma chroma sharpening to improve contrast and sharpness. It also produces augmented datasets tailored for segmentation purposes.

- We introduced the CFDU encoder network that fully integrates conventional convolution to depthwise separable convolution focusing on the features of slender and tortuous vessels, thereby improving the segmentation accuracy and robustness of the model.
- We proposed a CAB module to enhance the interaction of features between the encoder and decoder. This module employs channel and spatial attention to extract more comprehensive representations of low-level and high-level features enabling interaction between features with different semantics, thereby reducing feature loss.
- We introduce the CFDU decoder, a channel-wise attention mechanism to highlight important features and suppress background noise significantly.
- Extensive experiments are performed on five public datasets, DRIVE, FIVES, CHASE, HRF, and STARE. The results demonstrate that our framework achieves good performance and a good segmentation in comparison with the other baseline frameworks.

The article are structured as follows: Section 2 presents an overview of previous research performed on retinal vessel segmentation, evaluating established methodologies and their limitations. In Section 3 the proposed architecture is presented, explaining the basic components and the justifications behind design of the framework. In Section 4, the experimental setup use din this study is described with the use of datasets that were used and a comparison with state-of-art-methods. Finally, the main findings summarized and future research directions are suggested in Section 5.

2 Related Work

This section provides a comprehensive review of preprocessing methods, deep learning approaches and Transformer-based methods for medical image segmentation.

2.1 Preprocessing Methods

Standardizing and refining raw data plays a vital role for the purpose of data alignment with the selected framework. Additionally, preprocess improves feature quality, thereby enhancing the accuracy and efficiency of models, particularly in intricate assessment such as medical image interpretation. The process of converting an RGB image to a luminance-chrominance color model involves separating the color channels into various components such as an intensity component and two color-difference components. The partition facilitates isolation of spatial frequencies as the human eye is highly sensitive to intricate structural elements in luminance and less sensitive to high frequencies in color [17]. Fundus image preprocessing increases the diagnostic accuracy by enhancing contrast and balancing color to highlight critical structures such as blood vessels and lesions. Subsequently, it reduces artifacts and corrects uneven illumination to acquire uniform image clarity for further analysis [18]. Patch-based extraction to mitigate computational complexity and enhance local feature learning intensity normalization for training stability[19]. The preprocessing method enhances retinal images by isolating the green channel, applying CLAHE, and utilizing Gamma correction to reduce noise. Additionally, the approach implements an improved Black Hat

transformation and multiscale matched filtering to highlight tiny vessels and reduce background noise [20]. Intensity transformation, in-painting, out-painting, random pixel filling and a vector classification module is used to enhance structural and topological feature learning [21]. Fine vascular structures are highlighted by green channel extraction followed by CLAHE procedure. Further steps include grayscale standardization for training stability and data augmentation to reduce overfitting [22–25]. A perfect preprocessing is crucial for increasing fundus image quality before extracting retinal vessels accurately. Sophisticated methods, such as contrast enhancement and noise reduction, are used to address issues including inconsistent lighting, limited visibility in certain areas, and diminished contrast in intricate vascular structures.

2.2 Deep Learning architectures for Retinal Vessel Segmentation

The deep learning techniques enables automated extraction of image features from large datasets without human intervention. These methods have increased to prominence within medical field. The restricted spatial scope of convolutional operations obstructs their ability to accommodate the intricate and several structures of retinal vessels [26–28]. A distinct variety of methods have been suggested to overcome this challenge. Gu *et al.* [29] proposed a context encoder network for improvement of capturing high level spatial information and reduce the grid problem in expansion convolution. Jiang *et al.* [30] introduced a down-sampling coefficient and joint expansion convolution for segmentation task. Feng *et al.* [31] formulated a cross-stitching network to amalgamate multi-scale features, enhancing robustness and obviating the need for pre-processing hyperparameters. Kamran *et al.* [32] introduced the multi-scale RVGan to enhance the extraction of retinal microvessel structures. Wang *et al.* [33] proposed a dual-channel encodernetwork featuring a feature fusion module, an attention jump module, and a structure loss to refine micro vessel segmentation and boundary detection. Zhang *et al.* [34] introduced Bridge-net model combines recurrent neural networks (RNNs) with convolutional neural networks (CNNs) and integrates a patch-based loss-weighting technique to address imbalances in vessel morphology.The model uses multi-head attention, scaled dot-product attention, and positional encodings, setting a new standard for sequence-to-sequence modeling [35]. Mineault *et al.* [36] highlight how the transformer remixes concepts by shifting from a selective, bottlenecked, and dynamic biological attention mechanism to an exhaustive, static, and parallelized matrix. Tong *et al.* [37] introduce a lightweight transformer network tailored for segmenting retinal blood vessels. The proposed network combines elements from MobileViT+ with a unique local representation in the encoder module to prioritize computational efficiency. Lin *et al.* [38] utilize a dual-scale encoder for simultaneous local and global feature extraction combined with a transformer interactive fusion module, which effectively addresses limitations in capturing both detailed anatomy and contextual background. Chen *et al.* [39] utilize a hybrid architecture integrating CNNs and Transformers to enhance medical image segmentation by capturing long-range dependencies through a specialized self-aware attention module at the bottleneck. Cao *et al.* [40] developed a fully Transformer-based architecture that

relies on no standard convolutions. Liu *et al.*[41] developed IMFF-Net, a segmentation network that uses APF blocks to prevent spatial information loss and UDFF blocks to segment fine structures precisely. Ding *et al.* [42] proposed a retinal vessel segmentation method using a rough channel attention U-Net to handle feature uncertainty. Although it is incredibly difficult for the model to capture pixel-level context, resulting in blurred segmentation boundaries around small tumors, lesions, or organs. This paper introduces a novel hybrid network framework, SRD U-Net, that combines convolutional and Transformer architectures for improved performance. Researchers have introduced approaches such as attention modules and patch-based loss-weighting to specifically address concerns about spatial information loss, inadequate use of global data, and morphological differences between thick and thin vessels. Nonetheless, opportunities persist to develop more powerful techniques to address additional issues and complexities in this domain.

3 Methodology

The proposed method, derived from the U-shaped model, comprises an encoder and a decoder.Figure 1 shows that the comprehensive procedural and architectural outline of proposed work. The workflow is divided in two stages. The first stage is CARP block deals with preprocessing steps. The second stage is the major part of the segmentation task indicated as SRD U-Net. The detailed explanation of each stage as follows.

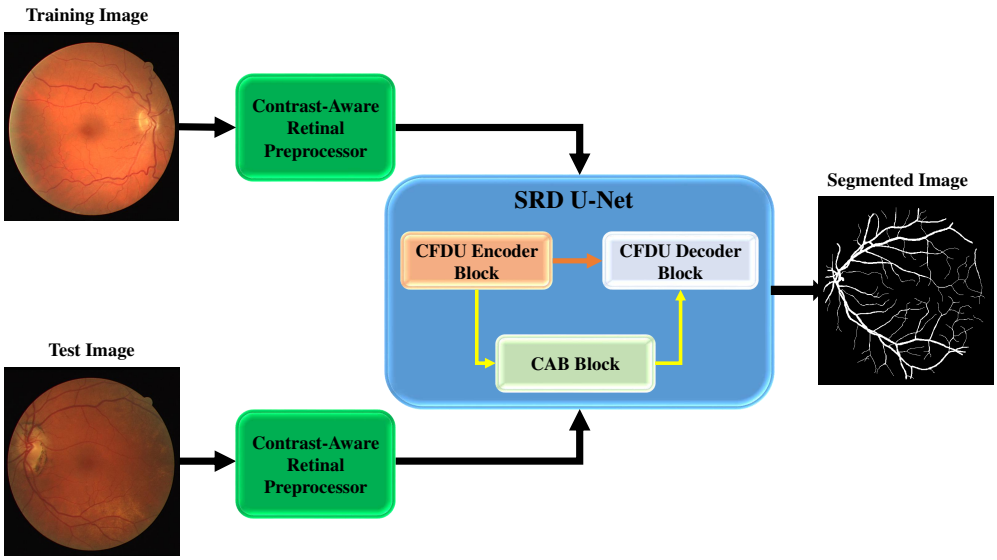


Fig. 1 Overview of the proposed method.

3.1 CARP Block

The preprocessing stage was designed to improve the contrast and sharpness of the input images. The pipeline used to preprocess maintains colour fidelity and improves the visibility of the fine patterns without causing chromatic aberration. The first step in the process is to transform the input image from RGB colour space to LAB colour space. The transform separates lightness channel from the chromatic channels, allowing for selective contrast improvement of brightness information. CLAHE operation is performed on the lightness channel. CLAHE effectively applied locally to redistribute the intensities of the pixels within small tiles,with clip limit setting controlling the amount of over amplification of the noise. The increased light channel is then fused to the original chromatic channels, and the image is then transformed back to RGB, maintaining the natural colour balance of the image. Finally, adaptive style sharpening is performed for leverages edge-aware logic similar to that of attention mechanisms. Edge strength can be calculated using Sobel gradients, contrast using the local standard deviation, and a smooth baseline using a Gaussian blur. These are all aggregated into an adaptation weight function which is able to enhance edges and dampen noise in flat areas. The increased luma is recalculated with chroma to produce the final image. This pipeline combines adaptive style sharpening and CLAHE contrast enhancement to create edge?aware images with preserved color information.Adaptive sharpening based on gradients and local contrast. Augmentation used for the diversity, and splitting the dataset into structures helps with reproducible training and evaluation. These steps create a strong pre-processing procedure that can be applied in medical imaging applications like retinal vessel analysis.The subsequent steps involve supplying the processed images to the SRD U-Net for inference and then reconstructing the resultant outputs.

3.2 SRD U-Net Block

The SRD U-Net consists of three major blocks including the CFDU encoder, the CAB, and the CFDU decoder that shows as in Figure 2. The detailed functions of each block are explained as follows.

3.2.1 CFDU Encoder Block

The proposed encoder architecture uses a custom block, the CFDU block. It sequentially combines a standard convolutional layer for initial feature extraction with a manual depthwise separable convolution. The?MobileNet inspires the framework framework inspired by MobileNet to optimize parameter efficiency while maintaining high-fidelity spatial and channel-wise representations. The model receives the raw input image tensor. The input tensor has a shape of $512 \times 512 \times 3$, corresponding to a height and width of 512 pixels with 3 colour color channels (RGB). A standard 2D convolutional layer applies 4 small learnable filters across the input image. This expands the channel dimension from 3 to 4. It captures primitive low-level features such as edges, lines, and simple textures while maintaining a spatial resolution of 512×512. Batch The batch normalization layer normalizes the activations of the previous convolution layer across the current batch. Enforcing a zero mean and unit

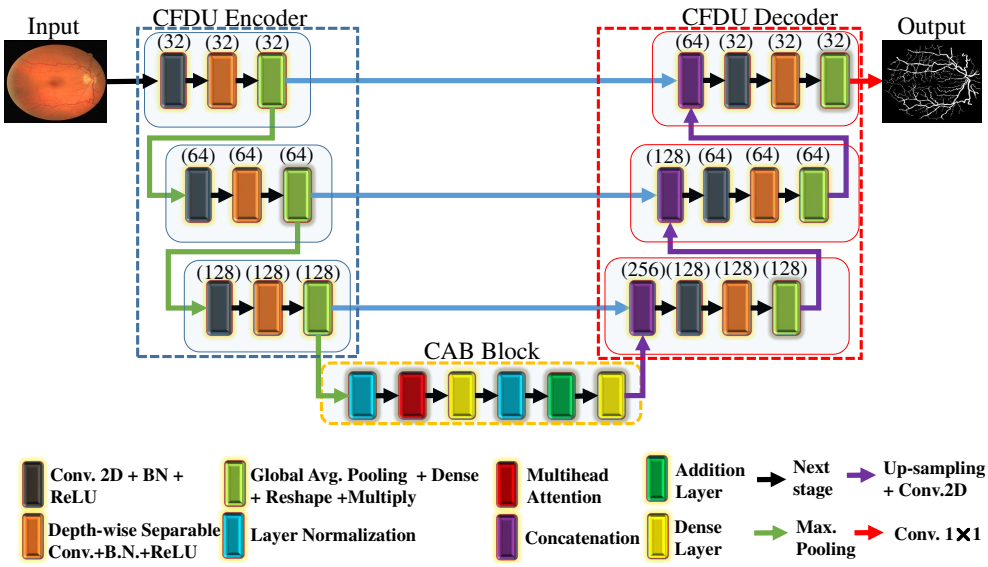


Fig. 2 Architecture of the SRD U-Net.

variance mitigates internal covariate shift, speeds up model convergence, and acts as a mild regularize regularizer to prevent overfitting. The rectified linear unit (ReLU) activation function sets all negative pixel values in the feature maps to zero. This allows the network to learn complex, non-linear representations of the data without altering the shape or parameter counts. The process is expressed as,

$$X_{\text{conv}} = \text{ReLU}(\text{BatchNormalization}(W_{\text{init}} \otimes X_{\text{input}} + b_{\text{init}})) \tag{1}$$

where $X_{\text{conv}}$is the convolution output, $\otimes$ denotes the convolution operation, X_{input} input denotes the input , W_{init} and b_{init} are the initial weights and biases of the first stage. To reduce computational complexity and parameter count, a depthwise separable convolution is used instead of a standard convolution. This is the first half of the depthwise separable convolution. Instead of mixing channels, it applies a single spatial filter to each channel independently. The standard convolution acts as the second half of the depthwise separable convolution. It linearly combines the spatial maps generated by the depthwise layer. In this specific configuration, it maps 4 channels back to 4 channels, completing a highly efficient feature-extraction cycle. As in the first block, these layers stabilize the pointwise convolutional output and introduce nonlinearity, preparing the 4-channel feature maps for the attention mechanism,which is expressed as,

$$X_{\text{dw}} = \text{depthwiseconv}(X_{\text{conv}}) \tag{2}$$

$$X_{\text{features}} = \text{ReLU}[\text{BatchNormalization}(W_{\text{dw}} \otimes X_{\text{dw}} + b_{\text{dw}})] \tag{3}$$

where $X_{\text{dw}}$is the depthwise convolution of the convolution output,X_{features} is the output features after the second standard convolution operation, W_{dw} and b_{dw} are the weights and biases of the depthwise stage. To dynamically recalibrate channel-wise

feature responses, a squeeze-and-excitation network is embedded in each block. The layer flattens each feature map into a single average scalar value. The output shape changes from $512 \times 512 \times 4$to a 1D vector of size 4. This step extracts the global contextual summary of each channel, stripping away spatial locations to focus purely on channel importance. These fully connected layers act as a bottleneck gating mechanism. The first dense layer acts as a compression bottleneck. The second dense layer expands the vector back out to a size of 4. Together, they learn the non-linear interdependencies between your channels to determine which features are most important. The 1D attention vector of size 4 is reshaped into a 4D tensor of shape $1 \times 1 \times 4$. This does not alter the data but changes its structure so it can interact mathematically with the original high-resolution feature maps. An element-wise multiplication is performed between the original feature maps from ReLU ($512 \times 512 \times 4$) and the attention weights from reshape ($1 \times 1 \times 4$), denoted as.

$$X_{\text{ref}} = F_{\text{scale}}\left[X_{\text{features}}, \sigma\left\{W_2 \times \text{ReLU}\left(W_1 \times F_{\text{sq}}(X_{\text{features}})\right)\right\}\right] \tag{4}$$

where X_{ref} is the output of the CFDU block, and F_{scale} represents the scaling operation used to compute the final refined feature maps by scaling the original features with the attention weights. The terms W_2 and W_1 denote the weights of two sequential dense layers, σ is the activation function maps weights to a scale factor, and F_{sq} represents the operation where spatial information is squeezed into a channel descriptor vector. Channels that are deemed highly relevant are scaled up, while less informative channels are scaled down. A neighborhood filter scans the refined feature maps and keeps only the maximum value in each window. This halves the spatial resolution from 512×512 to 256×256 . This reduces computational load and expands the receptive field, so successive layers can view larger regions of the image,which is expressed as,

$$X_{\text{down}} = \text{MaxPooling}(X_{\text{ref}}) \tag{5}$$

where X_{down} is the max. pooled output of the CFDU encoder block. The proposed multistage encoder pipeline architecture processes the input tensor through a sequence of steps, systematically increasing the number of feature channels while decreasing the spatial resolution. The X_{down} is fed to both the CAB and the CFDU decoder block.

3.2.2 CAB Block

In order to connect pathways of the encoder and decoder, a CAB block is used. This block uses a self-attention mechanism to learn long-range contextual relationships among the tokens across width of image. The entire latent feature map is located here. The bottleneck pipeline is a series of special stages that starts with pre attention layer normalization. Layer normalization is a way of normalizing the distributions of features along the channel axis. To stabilize downstream matrix computations for each pixel coordinate the normalized feature maps are projected through learnable weight matrices into separate query, key and value semantic spaces. This establishes the scaling up dot-product self-attention stage. The query and key is a matrix multiplication that results in an affinity map of the whole image landscape. This map estimates the

correlation between each pixel position and all others in the latent space by avoiding the limitation of a local convolution. Once scaled and the softmax applied row wise in these stages, the value tensor is multiplied by these weights to pool global contexts, which is normalization does. In order to maintain the integrity of architecture,the first residual addition layer combines this attention refined global context vector directly back to original input.The identity tensor is used to prevent vanishing gradients in the network. Then, the features are combined and then pass through a pre feed-forward network(FFN). One more step of layer normalization to normalize the variance range before feeding into the FFN. This network has two consecutive fully connected layers is the first layer projects features into a non-linear hidden layer.ReLU activation function is used to separate abstract semantic cues in space. In-contrast the second restores channels to their original depth dimensions. Finally, a second residual addition adds the refined FFN outputs back to the attention tensor, yielding a high-fidelity, globally synchronized latent map that is passed to the decoder for precise spatial reconstruction. The entire procedure is expressed as,

$$X_{\text{attn}} = \text{softmax}\left(\frac{X_{\text{down}} \times W_{q,k,v}}{\sqrt{d}}\right) \times V \tag{6}$$

where X_{attn} represents the CAB block output, $W_{q,k,v}$ are weights to project input embeddings into queries, keys, and values, $d$is the embedding dimensions,V is the actual information carried by each token. The CAB block output is fed to the CFDU decoder block.

3.2.3 CFDU Decoder Block

To match the encoder structure symmetrically, the decoder mirrors the spatial recovery phase using up-sampling and convolutions combined with skip connections from the encoder. The CAB block final output is up-sampled to double the spatial resolution of the feature map. The up-sampled output and CFDU encoder block output are concatenated and expressed as,

$$X_{\text{up}} = \text{Upsampling}(X_{\text{attn}}) \tag{7}$$

$$X_{\text{conc}} = C(X_{\text{down}}, X_{\text{up}}) \tag{8}$$

where X_{conc} merges low-level spatial details with high-level semantic data.Cdenotes the concatenation operation. The fused tensor X_{conc} is processed by reversing the channel expansion to smooth checkerboard artifacts and refine segmentations is expressed as,

$$X_{\text{dec}} = \text{ReLU}[\text{BatchNormalization}(W_{\text{dec}} \otimes X_{\text{conc}} + b_{\text{dec}})] \tag{9}$$

where $X_{dec}$is the standard convolution in the CFDU decoder block, W_{dec} and b_{dec} are the weights and biases of the first decoder stage. After this, a depthwise separable convolution operation is performed on the X_{dec} followed by a convolution operation denoted as,

$$X_{\text{depth}} = \text{depthwiseconv}(X_{\text{dec}}) \tag{10}$$

$$X_{\text{decd}} = \text{ReLU}[\text{BatchNormalization}(W_{\text{decd}} \otimes X_{\text{depth}} + b_{\text{decd}})] \tag{11}$$

where X_{decd} is the final output of the CFDU decoder block,W_{decd} and b_{decd} are the weights and biases of the final decoder stage. At the final output layer, a 1×1 convolution maps the deep feature channels back to the target number of classes, followed by a sigmoid activation function to generate pixel-wise probability maps, expressed as,

$$Y_{\text{pred}} = \sigma[W_{\text{final}} \otimes X_{\text{decd}} + b_{\text{final}}] \tag{12}$$

where $Y_{pred}$is the output of the proposed model,W_{final} and b_{final} are the weights and biases of the final proposed model. The proposed approach refines patch embedding with channel and spatial attention, enabling the model to focus on informative regions.

$$Y_{\infty} = \left(\frac{m}{\text{GeV}}\right)^{-3} \left[1 + \frac{3\ln(m/\text{GeV})}{15} + \frac{\ln(c_2/5)}{15}\right] \tag{13}$$

4 Experiment and Results

The study conducted experiments utilizing datasets including DRIVE, STARE, HRF, FIVES, and CHASE. Initially, preprocessing techniques were outlined, followed by the use of diverse evaluation measures for quantifying segmentation results. Detailed descriptions of experimental configuration and neural network training ensued. Subsequent presentation of ablation experiment outcomes aimed to validate the efficacy of the proposed approach. Ultimately, a comparative analysis with existing neural network methodologies documented in the literature was conducted.

4.1 Dataset

The experiments are conducted on five public datasets to thoroughly evaluate the proposed approach against the state of the art. Table 1 provides an overview of retinal vessel segmentation across various datasets, which consist of fundus images at different resolutions. These datasets were preprocessed by the CARP block, which

Table 1 Summary of Retinal Fundus Datasets

S.No.	Dataset	Image Type	No. of Images	Resolution
1	DRIVE[43]	Retinal fundus	20	768×584
2	STARE[44]	RGB fundus	20	700×600
3	HRF[45]	RGB fundus	20	1280×960
4	CHASEDB1[46]	RGB fundus	20	512×512
5	FIVES[47]	RGB fundus	20	2048×2048

improves image quality. After applying CARP for dataset diversity, the images are fed to various augmentation techniques, including vertical and horizontal flips, elastic transformations, optical distortions, and grid distortions. We consider 20 images from each dataset, and after augmentation, the count increases to 120. We split the 120 images into training, testing, and validation sets at 80:10:10.

4.2 Evaluation Metrics

Table 2 lists the mathematical expressions for the recall, accuracy, Jaccard, F1 score, and precision [48].

Table 2 Evaluation metrics and mathematical formulas.

Metric	Method
Accuracy	$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}}$
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$
Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$
Jaccard	$\text{Jaccard} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}$
F1 Score	$\text{F1 Score} = 2 \times \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$

4.3 Implementation Details

All experiments were performed on a system equipped with an NVIDIA GeForce RTX 3070 GPU with 8 GB of dedicated GPU memory and a 12th-generation Intel Core CPU with 64 GB of RAM. Training was performed using the TensorFlow library with mixed precision enabled. To mitigate the data-hungry nature of Transformers, our hybrid architecture restricts self-attention to the high-level bottleneck to minimize learnable parameters, and an offline augmentation pipeline (Elastic Transforms/Grid Distortions) to enforce structural regularization and ensure robust feature learning from limited samples. The retinal fundus images are resized to 512×512 pixels for processing. To ensure a fair comparison, all baseline models were re-implemented and trained under the same experimental conditions and hardware environment as the proposed model. This eliminates performance discrepancies caused by external factors such as varying libraries and hardware acceleration. Each model was trained for 100 epochs, with a learning rate of 0.001, a batch size of 1, and the root mean square propagation optimizer and cosine annealing schedule throughout.

4.4 Results and Discussion

To confirm both the effectiveness and generalizability of the proposed multi-block architecture, additional intensive experiments were conducted on a contextual dataset, each representing a unique type of imaging modality and diagnostic goal. The purpose of the experiments was to determine the segmentation accuracy, robustness to domain shifts, and the ability to track diseases over time. Table 3 summarizes that the model delivers solid performance overall. Figure 3 depicts the qualitative results across all evaluated networks. The proposed SRB model consistently delivers exceptional computational efficiency across all five benchmark datasets while maintaining highly competitive segmentation accuracy. To evaluate the generalizability of the proposed

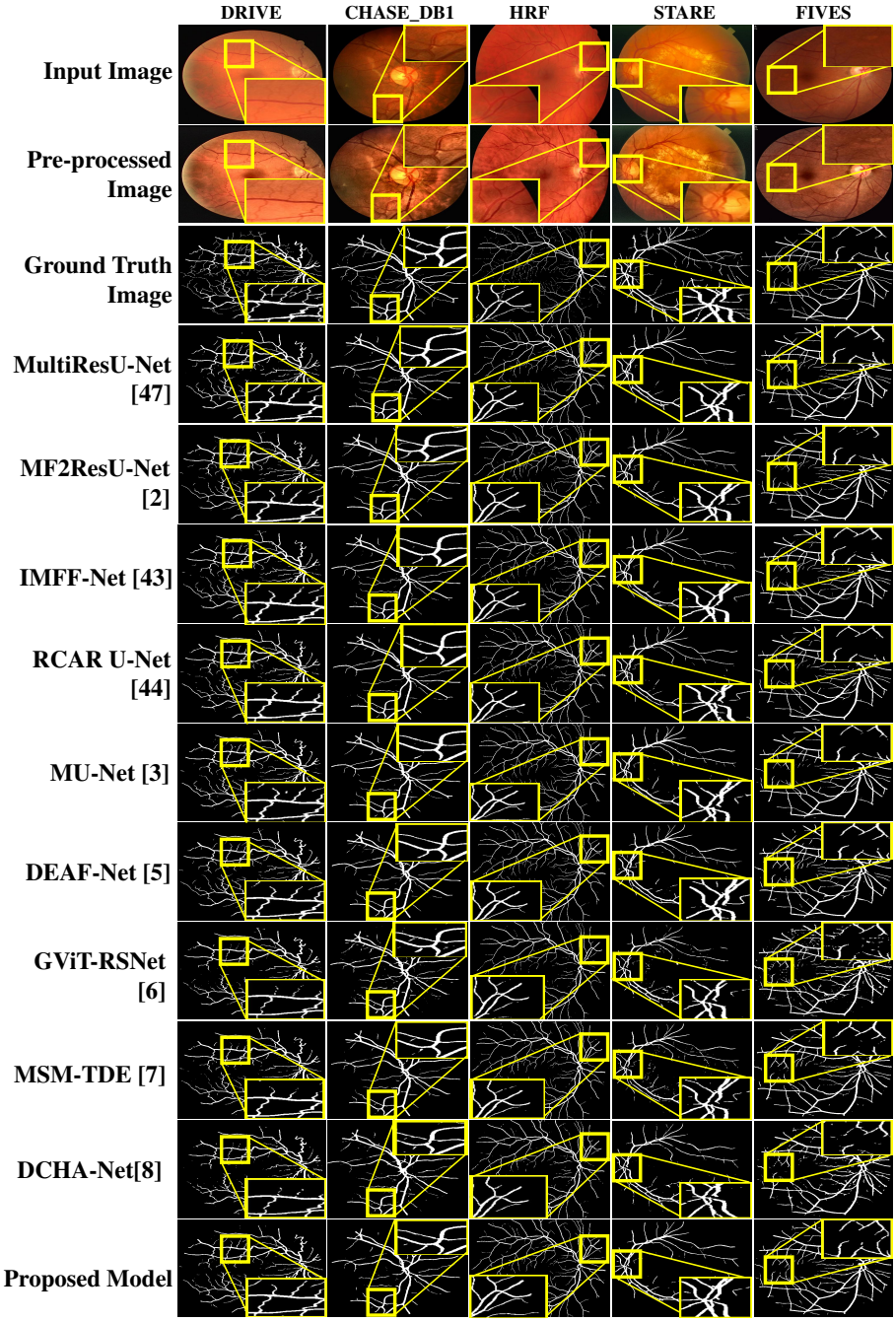


Fig. 3 Comparison of qualitative results of proposed method with existing methods using various datasets.

Dataset	Year	Model	Accuracy	F1 score	Jaccard	Recall	Precision
DRIVE	2020	MultiResUNet [47]	0.891	0.977	0.955	0.976	0.978
	2022	MF2ResU-Net[2]	0.897	0.979	0.958	0.981	0.976
	2024	IMFF-Net [43]	0.894	0.978	0.957	0.976	0.981
	2024	RCAR-UNet [44]	0.892	0.977	0.955	0.978	0.976
	2024	MU-Net [3]	0.889	0.977	0.956	0.976	0.979
	2025	DEAF-Net [5]	0.893	0.979	0.959	0.978	0.979
	2025	GViT-RSNet [6]	0.892	0.977	0.954	0.979	0.974
	2025	MSM-TDE [7]	0.897	0.978	0.957	0.982	0.974
	2026	DCHA-Net [8]	0.884	0.973	0.948	0.979	0.967
	2026	**Proposed Model**	**0.912**	**0.971**	**0.943**	**0.977**	**0.964**
STARE	2020	MultiResUNet [47]	0.939	0.985	0.971	0.994	0.976
	2022	MF2ResU-Net[2]	0.935	0.983	0.967	0.995	0.972
	2024	IMFF-Net [43]	0.939	0.985	0.969	0.995	0.975
	2024	RCAR-UNet [44]	0.935	0.984	0.968	0.995	0.973
	2024	MU-Net [3]	0.934	0.983	0.967	0.992	0.974
	2025	DEAF-Net [5]	0.927	0.979	0.961	0.989	0.971
	2025	GViT-RSNet [6]	0.931	0.979	0.959	0.989	0.968
	2025	MSM-TDE [7]	0.936	0.984	0.969	0.993	0.976
	2026	DCHA-Net [8]	0.933	0.984	0.968	0.993	0.976
	2026	**Proposed Model**	**0.931**	**0.981**	**0.963**	**0.991**	**0.973**
HRF	2020	MultiResUNet [47]	0.899	0.977	0.956	0.979	0.976
	2022	MF2ResU-Net[2]	0.891	0.974	0.950	0.977	0.972
	2024	IMFF-Net [43]	0.897	0.977	0.954	0.979	0.975
	2024	RCAR-UNet [44]	0.898	0.977	0.954	0.978	0.976
	2024	MU-Net [3]	0.895	0.976	0.954	0.977	0.976
	2025	DEAF-Net [5]	0.895	0.977	0.955	0.977	0.976
	2025	GViT-RSNet [6]	0.902	0.977	0.955	0.982	0.972
	2025	MSM-TDE [7]	0.891	0.975	0.952	0.976	0.975
	2026	DCHA-Net [8]	0.895	0.975	0.951	0.979	0.971
	2026	**Proposed Model**	**0.894**	**0.974**	**0.951**	**0.975**	**0.974**
CHASE	2020	MultiResUNet [47]	0.927	0.985	0.971	0.991	0.981
	2022	MF2ResU-Net[2]	0.929	0.987	0.973	0.991	0.982
	2024	IMFF-Net [43]	0.931	0.987	0.974	0.991	0.983
	2024	RCAR-UNet [44]	0.929	0.985	0.971	0.991	0.979
	2024	MU-Net [3]	0.923	0.985	0.971	0.989	0.979
	2025	DEAF-Net [5]	0.922	0.984	0.969	0.987	0.982
	2025	GViT-RSNet [6]	0.923	0.983	0.966	0.983	0.982
	2025	MSM-TDE [7]	0.924	0.984	0.969	0.989	0.978
	2026	DCHA-Net [8]	0.922	0.983	0.967	0.986	0.979
	2026	**Proposed Model**	**0.919**	**0.982**	**0.965**	**0.985**	**0.981**
FIVES	2020	MultiResUNet [47]	0.913	0.988	0.975	0.988	0.987
	2022	MF2ResU-Net[2]	0.909	0.987	0.975	0.988	0.987
	2024	IMFF-Net [43]	0.914	0.988	0.977	0.986	0.989
	2024	RCAR-UNet [44]	0.909	0.987	0.974	0.987	0.987
	2024	MU-Net [3]	0.915	0.988	0.977	0.988	0.989
	2025	DEAF-Net [5]	0.904	0.984	0.968	0.981	0.987
	2025	GViT-RSNet [6]	0.899	0.976	0.954	0.969	0.983
	2025	MSM-TDE [7]	0.909	0.986	0.973	0.985	0.988
	2026	DCHA-Net [8]	0.907	0.983	0.966	0.979	0.985
	2026	**Proposed Model**	**0.911**	**0.983**	**0.966**	**0.989**	**0.977**

Table 3 Performance Comparison Across Different Datasets

framework, cross-dataset validation was conducted across five standard benchmarks: DRIVE, STARE, HRF, CHASE, and FIVES. Table 3 summarizes the quantitative performance metrics under uniform hyperparameter settings. Crucially, the proposed SRB model maintains a staticarchitectural footprint of only 57,995 parameters across all trials. This footprint represents a profound reduction of up to 99.7% compared to heavy networks such as MSM-TDE (19.3M parameters) and MU-Net (17.2M parameters). An analysis of execution speed demonstrates the operational superiority of the proposed design. In the DRIVE dataset, the proposed SRB completes processing in 2 minutes, drastically outperforming DEAF-Net in 102 minutes while retaining an accuracy of 0.881 and an F1 score of 0.971. This trend persists in the STARE dataset, where the SRB requires only 2 minutes to achieve a high accuracy of 0.931 and a Jaccard of 0.963, closely rivaling the top-performing MultiResUNet (0.939 accuracy) at a fraction of the temporal cost. Furthermore, evaluation on the high-resolution HRF dataset shows that the proposed SRB logs a processing time of just 2 minutes. It achieves a 0.894 accuracy and 0.951 Jaccard, which closely approaches the peak performance of GViT-RSNet (0.902 accuracy, 6 minutes). On the CHASE dataset, the proposed model achieves an operational duration of 3 minutes, with strong 0.919 accuracy and 0.965 Jaccard. Finally, validation on the FIVES dataset confirms the proposed model reliability, yielding a 0.911 accuracy and 0.966 Jaccard in 2 minutes, matching the performance profiles of networks with millions of parameters. Consequently, these multi-dataset findings demonstrate that the proposed SRB provides an optimal trade-off between segmentation quality and edge-hardware viability.

4.5 Computational complexity

A comprehensive evaluation of our proposed model against ten state-of-the-art baselines across multiple performance dimensions, including accuracy, latency, computational complexity (FLOPs), and memory footprint is presented in Table 4. Our proposed architecture strictly dominates these high-performing baselines in terms of resource efficiency, achieving drastically lower latency, minimized FLOPs, and a highly reduced memory footprint. These multi-criteria results demonstrate that our model offers an optimal performance-to-resource trade-off, making it exceptionally well-suited for practical deployment in resource-constrained environments.

4.6 Ablation Study

To evaluate the individual contributions of the each architectural components, an ablation study was performed across all five datasets. Table 5 depicts the performance transitions among four unique configuration combinations under identical hyperparameter environment. The traditional encoder-decoder framework alone establishes a heavy baseline with 7,708,739 parameters and up to 11 minutes of processing time. By combining the attention mechanism with the traditional framework, drastically lowers the parameter count to 78,109 while preserving competitive segmentation scores. Conversely, isolating the CFDU block yields an intermediate network scale of 1,312,940 parameters and a uniform 5-minute execution time. Combining the attention mechanism and the CFDU block successfully downscales the network to its minimum

Table 4 Comparison of different models based on Flops, Parameters, Latency, and Memory Footprint.

Model	Flops (in GB ↓)	Parameters (↓)	Avg. Latency (in secs ↓)	Memory Foot Print(MB ↓)
MultiResUNet [47]	125.43	7,269,294	0.23	75.91
MF2ResU-Net[2]	413.56	8,259,521	0.03	150.04
IMFF-Net [43]	188.02	5,594,889	0.01	175.73
RCAR-UNet [44]	44.36	1,091,517	0.21	60.79
MU-Net [3]	584.67	17,213,177	0.22	226.61
DEAF-Net [5]	1,254.78	2,399,127	0.23	51.82
GViT-RSNet [6]	153.09	1,086,785	0.02	55.67
MSM-TDE [7]	388.1	19,303,681	0.04	114.41
DCHA-Net [8]	28.12	1,010,240	0.03	182.18
Proposed Model	**1.47**	**57,995**	**0.06**	**5.69**

footprint of 57,995 parameters. However, this combination suffers from unstable processing latencies ranging from 3 to 7 minutes. Crucially, introducing pre-processing alongside the attention mechanism and the CFDU block delivers the optimized proposed framework. While the model size remains static at 57,995 parameters, the attention mechanism and CFDU block systematically compresses processing times to the absolute minimum of 2 to 3 minutes across all datasets. Minor metric trade-offs, such as a marginal accuracy adjustment in the DRIVE and CHASE datasets, are heavily outweighed by these substantial operational gains. Therefore, the full integration of the attention mechanism and the systematic pre-processing of the CFDU block represents the optimal design for high-speed, edge-hardware deployment. The K-fold cross-validation results prove outstanding model stability and robust generalizability across all five datasets. To assess the stability, Table 6 outlines K-fold cross-validation results. The narrow standard deviations demonstrate excellent model consistency. The proposed model effectively avoids overfitting across all benchmarks.The STARE dataset evaluation shows a high accuracy score of 0.935. The proposed model with FIVES dataset achieves a peak F1 score of 0.985 and Jaccard index reaches to 0.970 ± 0.0014 The model reliably handles the challenging DRIVE dataset. DRIVE results show a Jaccard index of 0.951. HRF validation maintains a tight Jaccard score of 0.950. CHASE results exhibit exceptional precision-recall balance. It logs a 0.980 precision score seamlessly.However, These low-variance metrics confirms that the proposed model has high clinical reproducibility.

To ensure a fair and unbiased comparison among the five different datasets, we initially performed a balanced evaluation subset by sampling exactly 20 images from each dataset, irrespective of their varying original size image numbers. Consistent sampling was strictly necessitated by the availability of ground truth annotations in DRIVE and STARE datasets were limited by 20 images per dataset. To evaluate the optimized generalization capability of our proposed method, we conducted an extended test using a large-scale mixed dataset comprising 905 diverse images, which is the combination of all datasets. Further more, FIVES dataset consists of 800 images originally are also evaluated by the proposed model. Table 7 Significantly, the model performance remained highly stable across high dimensional datasets and

Dataset	E	B	D	P	Accuracy	F1 Score	Jaccard	Recall	Precision	Parameters	P.T
DRIVE	✓	×	×	×	0.894	0.977	0.955	0.978	0.976	7,708,739	8 m
	✓	✓	×	×	0.892	0.976	0.953	0.976	0.977	78,109	3 m
	×	×	✓	×	0.891	0.977	0.955	0.979	0.975	1,312,940	5 m
	×	✓	✓	×	0.884	0.973	0.948	0.978	0.969	57,995	7 m
	×	✓	✓	✓	0.881	0.971	0.943	0.977	0.964	57,995	2 m
STARE	✓	×	×	×	0.932	0.984	0.968	0.989	0.979	7,708,739	7 m
	✓	✓	×	×	0.935	0.982	0.966	0.989	0.975	78,109	3 m
	×	×	✓	×	0.937	0.985	0.969	0.993	0.973	1,312,940	5 m
	×	✓	✓	×	0.931	0.979	0.959	0.989	0.969	57,995	3 m
	×	✓	✓	✓	0.931	0.981	0.963	0.991	0.973	57,995	2 m
HRF	✓	×	×	×	0.893	0.976	0.953	0.977	0.975	7,708,739	11 m
	✓	✓	×	×	0.893	0.975	0.951	0.976	0.974	78,109	3 m
	×	×	✓	×	0.895	0.976	0.953	0.976	0.976	1,312,940	5 m
	×	✓	✓	×	0.897	0.976	0.952	0.976	0.975	57,995	4 m
	×	✓	✓	✓	0.894	0.974	0.951	0.975	0.974	57,995	2 m
CHASE	✓	×	×	×	0.923	0.985	0.969	0.991	0.978	7,708,739	10 m
	✓	✓	×	×	0.928	0.984	0.969	0.989	0.980	78,109	3 m
	×	×	✓	×	0.929	0.986	0.973	0.989	0.982	1,312,940	5 m
	×	✓	✓	×	0.922	0.983	0.967	0.985	0.981	57,995	4 m
	×	✓	✓	✓	0.919	0.982	0.965	0.985	0.981	57,995	3 m
FIVES	✓	×	×	×	0.901	0.979	0.961	0.974	0.986	7,708,739	9 m
	✓	✓	×	×	0.912	0.984	0.969	0.988	0.981	78,109	3 m
	×	×	✓	×	0.909	0.984	0.969	0.987	0.981	1,312,940	5 m
	×	✓	✓	×	0.912	0.985	0.969	0.986	0.984	57,995	4 m
	×	✓	✓	✓	0.911	0.983	0.966	0.989	0.977	57,995	2 m

Table 5 Ablation Study Across Different Datasets

Dataset	Accuracy	F1 score	Jaccard	Recall	Precision
DRIVE [45]	0.891 ± 0.0031	0.975 ± 0.0006	0.951 ± 0.0011	0.979 ± 0.0034	0.971 ± 0.0034
STARE [46]	0.935 ± 0.0011	0.983 ± 0.0004	0.966 ± 0.0007	0.991 ± 0.0011	0.974 ± 0.0011
HRF [47]	0.894 ± 0.0009	0.974 ± 0.0003	0.950 ± 0.0006	0.976 ± 0.0010	0.973 ± 0.0006
CHASE [48]	0.919 ± 0.0013	0.982 ± 0.0006	0.964 ± 0.0012	0.984 ± 0.0014	0.980 ± 0.0009
FIVES [49]	0.913 ± 0.0012	0.985 ± 0.0007	0.970 ± 0.0014	0.985 ± 0.0013	0.985 ± 0.0016

Table 6 Overall Performance Metrics with Confidence Intervals Across Different Datasets

more heterogeneous data distribution matches the metrics achieved on the balanced subset. This systematic regularity verifies that the strong generalization power and environmental adaptability of our approach across multiple data dimensions.

5 Conclusion

This paper demonstrates an ultra-lightweight SRD U-Net framework for retinal vessel segmentation. The Model solves ultimate computational complexity. Comprehensively the proposed architecture has evaluated with parameter count of 57,995 using FIVE bench mark datasets. This results in an novel reduction of up to 99.7% in parameters with respect to the best networks existed. Moreover, the model cuts execution time down to 2-3 minutes globally for DRIVE,STARE,CHASE,HRF and FIVES datasets evaluated, while maintaining a highly competitive accuracy, F1-scores, and Jaccard

Table 7 Performance Analysis Across Different Image Counts

S.No.	Dataset	Accuracy	F1 Score	Jaccard	Recall	Precision
1	FIVES with 800 images	0.922	0.988	0.976	0.992	0.984
2	Mixed	0.918	0.993	0.986	0.992	0.994

indices. The systematic ablation study with four structural configurations illustrates that the proposed model provides optimizing latency without compromising the segmentation accuracy at the pixel level. Moreover, exacting K-fold study states stability of the model and better generalization by the cross validation.Visual comparison of experiments also depicts that the SRD U-Net network model proposed in this paper has advanced than base line models. As a result, this systematic of the framework confirms that the proposed model is indeed real-life viable, directly fulfilling our introductory target of achieving a very efficient model for deployment on the clinical edge in real time.Future work will aim to improve the accuracy of network further in the retinal vessel segmentation task and related medical image tasks.

Declarations

We hereby declare that the submitted manuscript is our original work.

- Funding: Not applicable
- Conflict of interest/Competing interests : The authors declare that they have known competing personal relationships that could have appeared to influence the work reported in this paper.
- Ethics approval and consent to participate:Not applicable
- Consent for publication:Not applicable
- Data availability : All data generated during this study are included in this article
- Materials availability:All commercial reagents and materials are detailed in the Methods section
- Code availability:The custom software code generated during the current study is available from the corresponding author on reasonable request.
- Author contribution: Conceptualization and design: Rangu, Yamanakkanavar, A V, Data collection: Rangu. Data analysis: Rangu, Yamanakkanavar, A V. Drafting the manuscript: Rangu,Yamanakkanavar, A V. Critical revision and final approval: All authors.

References

[1] Steinmetz, J.D., Bourne, R.R., Briant, P.S., Flaxman, S.R., Taylor, H.R., Jonas, J.B., Abdoli, A.A., Abrha, W.A., Abualhasan, A., Abu-Gharbieh, E.G., *et al.*: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an

analysis for the global burden of disease study. The Lancet global health **9**(2), 144–160 (2021)

[2] Zhenchao, C., Shujie, S., Jing, Q.: Mf2resu-net: a multi-feature fusion deep learning architecture for retinal blood vessel segmentation. Digital Chinese Medicine **5**(4), 406–418 (2022)

[3] He, X., Wang, T., Yang, W.: Research on retinal vessel segmentation algorithm based on a modified u-shaped network. Applied Sciences **14**(1), 465 (2024)

[4] Li, Y., Yan, B., Hou, J., Bai, B., Huang, X., Xu, C., Fang, L.: Unet based on dynamic convolution decomposition and triplet attention. Scientific Reports **14**(1), 271 (2024)

[5] Cai, P., Li, B., Sun, G., Yang, B., Wang, X., Lv, C., Yan, J.: Deaf-net: Detail-enhanced attention feature fusion network for retinal vessel segmentation. Journal of Imaging Informatics in Medicine **38**(1), 496–519 (2025)

[6] Li, J., Cheng, Q., Wu, C.: Gvit-rsnet: A retinal vessel segmentation network using graph convolutional attention and multi-scale vision transformer. Pattern Recognition Letters **189**, 182–187 (2025)

[7] Zhang, H., Zhang, J., Zhong, X., Feng, Y., Li, G., Li, X., Lv, J., Ji, D.: Msm-tde: multi-scale semantics mining and tiny details enhancement network for retinal vessel segmentation. Complex & Intelligent Systems **11**(1), 114 (2025)

[8] Jiang, Y., Liu, W., Chen, J., Guo, H., Lu, M., Su, J.: Dcha-net: Dynamic convolution and hybrid attention network for retinal vessel segmentation. Intelligent Data Analysis **30**(1), 145–161 (2026)

[9] Khawaja, A., Khan, T.M., Naveed, K., Naqvi, S.S., Rehman, N.U., Nawaz, S.J.: An improved retinal vessel segmentation framework using frangi filter coupled with the probabilistic patch based denoiser. IEEE Access **7**, 164344–164361 (2019)

[10] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)

[11] Mu, N., Lyu, Z., Rezaeitaleshmahalleh, M., Tang, J., Jiang, J.: An attention residual u-net with differential preprocessing and geometric postprocessing: Learning how to segment vasculature including intracranial aneurysms. Medical image analysis **84**, 102697 (2023)

[12] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[13] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[14] Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R.: Dunet: A deformable network for retinal vessel segmentation. Knowledge-based systems **178**, 149–162 (2019)

[15] Poynton, C.: Digital Video and HD: Algorithms and Interfaces. Elsevier, ??? (2012)

[16] Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Graphics Gems IV, pp. 474–485. Academic Press, ??? (1994)

[17] Sapakova, S., Yesmukhamedov, N., Sapakov, A., Yemberdiyeva, A., Kozhamkulova, Z.: Methods for pre-processing and analysis of fund images for detection of diabetic retinopathy. Procedia Computer Science **272**, 496–501 (2025)

[18] Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual u-net for medical image segmentation. Journal of medical imaging **6**(1), 014006–014006 (2019)

[19] Ma, Y., Zhu, Z., Dong, Z., Shen, T., Sun, M., Kong, W.: Multichannel retinal blood vessel segmentation based on the combination of matched filter and u-net network. BioMed research international **2021**(1), 5561125 (2021)

[20] Tu, Z., Zhou, Q., Zou, H., Zhang, X.: A multi-task dense network with self-supervised learning for retinal vessel segmentation. Electronics **11**(21), 3538 (2022)

[21] He, Z., Li, X., Lv, N., Chen, Y., Cai, Y.: Retinal vascular segmentation network based on multi-scale adaptive feature fusion and dual-path upsampling. IEEE Access **12**, 48057–48067 (2024)

[22] Huy, V.T.Q., Lin, C.-M.: D2cbdamattunet: dual-decoder convolution block dual attention unet for accurate retinal vessel segmentation from fundus images. IEEE Access **13**, 19635–19649 (2025)

[23] Wen, B., Heinke, A., Agnihotri, A., Bartsch, D.-U., Freeman, W., Nguyen, T., An, C.: Universal vessel segmentation for multi-modality retinal images. IEEE Transactions on Image Processing (2025)

[24] Varma, A., Agrawal, M.: Thin vessel segmentation in fundus images using attention unet and modified frangi filtering. Biomedical Signal Processing and Control **99**, 106842 (2025)

[25] Li, Y., Zhang, Q., Zhou, H., An, Y., Li, J., Li, X., Zhou, H., Li, A.: Deep leaning-based interactive segmentation of three-dimensional blood vessel images. Biomedical Signal Processing and Control **104**, 107507 (2025)

[26] Ni, J., Sun, H., Xu, J., Liu, J., Chen, Z.: A feature aggregation and feature fusion network for retinal vessel segmentation. Biomedical Signal Processing and Control **85**, 104829 (2023)

[27] Zhang, X., Chen, Z., Wu, Q.J., Cai, L., Lu, D., Li, X.: Fast semantic segmentation for scene perception. IEEE Transactions on Industrial Informatics **15**(2), 1183–1192 (2018)

[28] Han, J., Wang, Y., Gong, H.: Fundus retinal vessels image segmentation method based on improved u-net. IRBM **43**(6), 628–639 (2022)

[29] Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging **38**(10), 2281–2292 (2019)

[30] Jiang, Y., Tan, N., Peng, T., Zhang, H.: Retinal vessels segmentation based on dilated multi-scale convolutional neural network. Ieee Access **7**, 76342–76352 (2019)

[31] Feng, S., Zhuo, Z., Pan, D., Tian, Q.: Ccnet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features. Neurocomputing **392**, 268–276 (2020)

[32] Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Sanders, K.M., Baker, S.A.: Rv-gan: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 34–44 (2021). Springer

[33] Wang, B., Wang, S., Qiu, S., Wei, W., Wang, H., He, H.: Csu-net: A context spatial u-net for accurate blood vessel segmentation in fundus images. IEEE Journal of Biomedical and Health Informatics **25**(4), 1128–1138 (2020)

[34] Zhang, Y., He, M., Chen, Z., Hu, K., Li, X., Gao, X.: Bridge-net: Context-involved u-net with patch-based loss weight mapping for retinal blood vessel segmentation. Expert Systems with Applications **195**, 116526 (2022)

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[36] Mineault, P.: Is attention all you need? In: From Human Attention to Computational Attention: A Multidisciplinary Approach, pp. 297–314. Springer, ???

(2025)

[37] Tong, L., Li, T., Zhang, Q., Zhang, Q., Zhu, R., Du, W., Hu, P.: Livit-net: A u-net-like, lightweight transformer network for retinal vessel segmentation. Computational and Structural Biotechnology Journal **24**, 213–224 (2024)

[38] Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D.: Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE Transactions on Instrumentation and Measurement **71**, 1–15 (2022)

[39] Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K.: Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(1), 55–68 (2023)

[40] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer

[41] Liu, M., Wang, Y., Wang, L., Hu, S., Wang, X., Ge, Q.: Imff-net: An integrated multi-scale feature fusion network for accurate retinal vessel segmentation from fundus images. Biomedical Signal Processing and Control **91**, 105980 (2024)

[42] Ding, W., Sun, Y., Huang, J., Ju, H., Zhang, C., Yang, G., Lin, C.-T.: Rcar-unet: Retinal vessel segmentation network algorithm via novel rough attention mechanism. Information Sciences **657**, 120007 (2024)

[43] Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4), 501–509 (2004)

[44] Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000)

[45] Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013**(1), 154860 (2013)

[46] Owen, C.G., Rudnicka, A.R., Mullen, R., Barman, S.A., Monekosso, D., Whincup, P.H., Ng, J., Paterson, C.: Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. Investigative ophthalmology & visual science **50**(5), 2004–2010 (2009)

[47] Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific data **9**(1), 475 (2022)

[48] Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)