Translight: Transformer Lightweight Model for Retinal Vascular Segmentation Using Fundus Images

Srinivas Rangu and Nagaraj Yamanakkanavar *

Department of Electronics and Communication Engineering, Central University of Karnataka, Kalaburagi, India
Email: srigana.57@gmail.com (S.R.); nagraj.p.y@gmail.com (N.Y.)
*Corresponding author

***Abstract*—Retina vascular segmentation exerts major part in the identification of age-related macular disease and diabetic retinopathy. Accurate retinal vessel segmentation depends on a range of enhancement techniques and deep learning models. Fundus images are primarily used for diagnosing eye diseases. High-quality visual information supports highly accurate clinical judgment. Conventional deep learning models commonly impair accuracy due to intricacy. However, the quality of the image frequently affects various factors such as noise. To overcome these limitations, we introduced a Stacked Efficient Depth (SED) encoder-decoder framework for retinal vascular segmentation. In this framework, we used four blocks, such as Gaussian Convolutional Residual (GCR), SED encoder, Transformer, and SED decoder. The GCR block introduces stochastic noise into the input during training, and the model develops increased robustness against various input disturbances. The SED encoder applies spatial filtering independently to each input channel. Moreover, a Vision Transformer block is used in the bottleneck to fuse the interest region detection of SED with global reasoning. The SED decoder block uses depthwise convolution to minimize the number of parameters. The proposed model's effectiveness is validated through both objective and subjective evaluations on publicly available datasets. Our proposed method demonstrates superior performance in evaluation metrics to existing approaches.**

***Keywords*—transformer, retinal vessel enhancement, stacked efficient depth, Gaussian convolutional residual**

I. Introduction

A key moment in diagnosing vision problems is the precise demarcation of retinal blood vessels. The unique vascular system, as depicted in images, which is often significantly affected by pathological conditions, remains a crucial challenge [1]. There is an imperative need for a significantly accurate, programmed, and broad diagnostic solution as a result of the surge in chronic illness, encompassing diabetic retinal disease. However, one of the main causes of visual impairment and blindness is Diabetic Retinopathy (DR) [2]. The segmentation of structurally complex diseases, such as retinal arteries, requires the ability to trace long-range spatial relationships [3]. However, Convolutional Neural Networks (CNNs) are particularly adept at extracting local features. Handcrafted features, such as quadrature filters [4], matching filters [5], and Gabor filters [6], were the cornerstone of early research on vessel segmentation. These techniques captured vessels with various widths and alignment by using definite kernel banks. The architecture lacks sequential progression modeling sets, despite showing a path forward for spatially aware learning [7]. A breakthrough in computer vision occurred with AlexNet, which won the ILSVRC 2012 competition and demonstrated the potential of CNNs on large-scale datasets. Since then, deep learning has advanced rapidly across diverse domains [8–12]. Ronneberger *et al*. [13] developed a model with two main components: one that shrinks the image to understand the context, and another that expands it to accurately capture fine details.

Oktay *et al*. [14] employed a medical imaging model that can automatically focus on important body structures, regardless of their size or shape. Ibtehaz and Rahman [15] introduced the Multires U-Net, which improves on the original U-Net design. These upgrades do not always show clearly in standard evaluation scores. Seo *et al.* [16] proposed a full-scale guided network, in which the information is refined by a guided convolution block. Radha and Karuna [17] developed a model that can efficiently recognize vessels with varying widths and orientations. Li *et al*. [18] proposed the combined approach GViT-RSNet, which combines graph convolution and vision transformers to meet local connectivity. Martinez-Pérez *et al*. [19] developed classical techniques, including region growing, piecewise thresholding, and concavity measurements. In addition to these, graph-theoretic approaches, such as the shortest-path tracking and fast marching algorithms, were introduced to trace vascular structures. Tan *et al*. [20] introduced a model that utilizes weighted fusion to connect feature maps from multiple levels, both top-down and bottom-up, thereby enhancing detection accuracy for objects of different sizes by adaptively combining information. Abtahi *et al*. [21] proposed a model that

Manuscript received January 23, 2026; revised February 24, 2026; accepted April 15, 2026; published September 11, 2026.

enhances the accuracy of segmentation and feature representation for retinal vasculature, which typically employs attention processes. Xiang *et al*. [22] developed a model that uses multiple decoding paths in conjunction with attention modules. The architecture aims to efficiently capture fine vessel structures, aiding in the early detection and management of retinal disorders.

Luo *et al*. [23] introduced a concept that uses multi-scale feature extraction and attention mechanisms to improve the accuracy of segmenting retinal blood vessels. Ahmed and Liatsis [24] proposed a novel retinal vascular segmentation model to capture complex vessel patterns and cross-channel interactions while reducing model complexity for edge deployment, called lightweight hyper complex U-net with vascular thickness-guided dice loss (LHU-VT). Dasgupta and Sonam [25] developed a model that combines convolutional feature extraction with skip connections and up-sampling to integrate semantic and spatial information. Subsequent modifications to the U-Net have aimed to improve segmentation accuracy. Alom *et al.* [26] incorporated recursive training strategies and residual connections, which improve performance but increase computational costs. Zhuang [27] introduced a multi-path network by serially combining two U-Net models, achieving richer feature learning at the expense of higher training time. Yun and Tan [28] extended the U-Net by incorporating conditional generative adversarial networks to enhance the balance and accuracy of segmentation. Jin *et al.* [29] proposed a model to enhance vascular feature representation by deformable CNNs. Although these approaches achieved performance gains, they substantially increased model complexity. To address these limitations, we introduce the Translight approach, which significantly improves upon existing frameworks. Our main contributions are summarized as follows:

- We proposed a Translight architecture with the lightweight convolution technique, which can remove irrelevant artifacts and enhance image contrast while preserving fine details.
- The transformer block at the bottleneck model gives global contextual relationships across the entire feature map, improving tasks like retinal image segmentation.
- Convolutional Block Attention Module (CBAM) block refines encoder features before concatenation with decoder up-sampled maps to emphasize relevant channels and spatial regions while suppressing noise.
- The LC decoder block suppresses noise twice spatially using CBAM, channel-wise through the LC encoder, recovering subtle retinal details lost in up-sampling.

The structure of this paper is as follows: Section II presents the related work, Section III covers the proposed methodology. To prove the effectiveness of our method, we discuss the experiments and results in Section IV, and Section V concludes the study with final observations.

II. Related Work

A. Transformer Architectures

Transformer-based networks are pioneering deep learning architectures originally created for natural language processing, but are now widely used in computer vision tasks such as medical image analysis. The self-attention mechanism is the primary innovation, enabling the model to dynamically weigh the significance of different elements within the input. Hatamizadeh *et al*. [30] proposed a transformer encoder to capture spatial and semantic features from 3D medical image patches.

Chen *et al*. [31] proposed a model that enables simultaneous local and global feature extraction for retinal vascular segmentation. Liang *et al*. [32] developed a hybrid CNN-transformer network that preserves micro vessels and enhances contextual feature fusion by utilizing spatial reconstruction and feature interaction transformer modules. Lv *et al*. [33] proposed a balanced local feature extraction and global context modeling for accurate segmentation using a dual-path U-Net with transformer blocks combined with convolutional decoding.

Jiang *et al*. [34] introduced transformers in parallel routes are used by a dual-branch network to segment retinal arteries while paying attention to both fine and coarse features. Mehmood *et al*. [35] used an effective extraction of retinal features, such as vessels and optic discs, a lightweight multitask CNN with integrated transformer modules is appropriate for real-time applications. Lin *et al*. [36] developed a model that enhances the identification of tiny vascular structures by including adaptive transformer attention blocks that are led by retinal stimulus information. Ke *et al*. [37] proposed a model that combines transformer blocks and convolutional layers in a U-Net architecture to take advantage of complementing local and global data for vessel segmentation.

Jiang *et al*. [38] introduced a hybrid model that performs vessel delineation by integrating a CNN for local characteristics and a vision transformer for global spatial dependencies. Tong *et al*. [39] developed MobileViT blocks that are embedded in a U-Net-style architecture to provide lightweight, precise vessel segmentation with less computational complexity. Yan *et al*. [40] introduced a striped pyramid pooling module that efficiently captures contextual information by aligning long, thin pooling kernels with the anisotropic, tree-like topology of retinal capillaries. Transformer networks use a primary self-attention mechanism to dynamically weigh the relevance of various input items, allowing for effective capture of long-range dependencies and global context. The design is centrally made up of arranged multi-head self-attention layers and feed-forward neural networks, with residual connections and layer normalization included to intensify training stability and flawless execution. The design allows transformers to succeed at challenging vision tasks like retinal vascular segmentation by replicating intricate global interconnectedness and spatial relationships.

B. Attention Mechanisms

Attention mechanisms, such as spatial and Channel Attention Module (CAM), allow the network to focus better on vessel features. Particularly fine and tortuous vessels that are difficult to segment precisely. Attention

modules distinguish simple features from complex backdrops by highlighting high-importance regions and reducing the impact of noise and artifacts. Self-attention processes are able to comprehend the global vascular structure and continuity, made possible by recording long-range spatial relationships within the retinal images. By dynamically weighing features at numerous scales, attention-based models can improve the segmentation of vessels of all sizes, from small capillaries to massive arteries.

Chen *et al*. [31] introduced a network that includes a feature generation attention module and enriched skip connections to further enhance segmentation accuracy. Ma and Li [41] proposed a model for thin vessels that utilizes a supervised attention mechanism integrated into a U-Net. It improves network focus on vessel areas by directing feature extraction with attention maps.

Chu *et al*. [42] used a Spatial Attention Module (SAM) at the decoder stage and a feature CAM inside residual spatial unit blocks to highlight important retinal vascular features while reducing noise and irrelevant areas. Jiang *et al*. [43] developed a group attention improvement module to reduce information loss through skip connections and facilitate detailed feature transmission by guiding low-level features with high-level vessel information. Jian *et al*. [44] introduced cross-fusion attention to describe interactions between the encoder and decoder. Local and global attention to highlight vessel-related traits. Javed *et al.* [45] proposed region-guided inverse addition attention blocks that focus on foreground vessel regions. Refining feature extraction and improving segmentation in regions of interest while maintaining model efficiency for deployment. Hernandez-Gutierrez *et al*. [46] proposed a model, reverse SAM, which recovers underlying feature information lost during upsampling between the encoder and decoder. A reverse CAM that records lost edge features and residual details during the encoding step. While maintaining a lightweight architecture suitable for real-time clinical and mobile healthcare applications, our reverse attention technique enhances the network capacity to precisely define vessel borders. Attention mechanisms in deep networks enable models to selectively focus on the most significant sections of input data by assigning different weights to various parts.

C. Light Weight Architectures

The lightweight models significantly lessen computational complexity and model size. It enables rapid reasoning in real-time applications and facilitates implementation on resource-constrained devices. The models typically employ parameter-efficient techniques, such as depth-wise channel attention, separable convolutions, and lightweight transformer blocks. Mehmood *et al.* [35] introduced a lightweight convolutional neural network utilizing a transformer architecture with fewer parameters and computational costs, while maintaining good segmentation accuracy. The hybrid solution quickly equilibrates the capability of CNN feature extraction, contrary to transformer-based global context modeling for retinal characteristics.

Tong *et al.* [39] proposed compact representations of attention modules into a compressed U-Net design. Daniel *et al.* [46] developed a model that recovers concealed vessel details and performs well in real-time clinical settings. Liu *et al.* [47] proposed a lightweight attention-based architecture and a hybrid CNN aimed at semi-supervised learning scenarios with fewer labels. Guo *et al.* [48] introduced a compact U-Net by inserting residual channel attention mechanisms that recalibrate channel-wise features. Zeng *et al.* [49] proposed an ultra-lightweight architecture with fewer parameters, optimized for embedded and mobile devices. Despite its remarkable compactness, it maintains competitive accuracy through the use of efficient convolutional modules and architectural enhancements. Although lightweight models are designed for resource and speed efficiency, limited capacity can render them more vulnerable to artifacts and image noise.

III. Proposed Methodology

The proposed model boosts performance and generalization in resource-constrained retinal vessel segmentation. Prior to the proposed methodology, preprocessing formalizes retinal fundus images by calibrating enlightenment, boosting vessel contrast, and reducing noise. The steps enhance the visibility of thin vessels in cluttered backgrounds, providing cleaner inputs that improve the extraction of deep learning features. To improve image quality, preprocessing steps such as gamma correction, grayscale conversion, and Contrast-Limited Adaptive Histogram Equalization (CLAHE) are performed. Data augmentation techniques are empirical to the preprocessed images, including flips, elastic transformations and distortions to further enhance dataset diversity. After performing augmentation, the processed images are provided to the proposed model. Fig. 1 illustrates the overall proposed framework. The proposed model is illustrated in four stages: such as (i) LC encoder block, (ii) Transformer block, (iii) CBAM block, and (iv) LC decoder block. The LC encoder block is used to extract rich hierarchical features using a few parameters. The transformer block applies its self-attention and feed-forward mechanisms to the flattened spatial tokens. The CBAM block enhances feature representation in CNNs by sequentially applying CAM and SAM. The LC decoder block reconstructs the spatial resolution lost during down-sampling. The detailed explanation of each stage is discussed as follows.

A. LC Encoder Block

The proposed LC encoder block model is inspired by the squeeze-and-excitation block. The LC encoder block is discussed in five phases to limit model complexity and aid generalization. The phases include a convolutional feature extraction, Global Average Pooling (GAP), channel importance learning, weight reshaping, and channel-wise reweighting. The combined LC encoder block is integrated into three encoder layers of the network to enhance feature recalibration and representation quality. The preprocessed image is represented as $A \in R^{H \times W}$, where A is the variable

representing the input, H is the height of the input, W is the width of the input, R is the set of real numbers, indicating each entry of A real-valued number. The input A is applied to the convolutional feature extraction of the LC encoder block. A convolution operation is performed on the input A to compute an output feature map $U \in R^{H \times W}$ and it is given as Eq. (1):

$$U_D = K_D \times A = \sum_{j=1}^{D'} k_D^j \times A_j + b_D \tag{1}$$

where U_D represents the output of the convolution layer in the LC encoder block, b_D is the bias term, A_j represents jth element of the input vector, $\times$ denotes the convolutional operator, and K_D is the weights for the Dth output channel. The resulting feature maps can be expressed as $U = [u_1, u_2, \ldots, u_D]$, k_D^j represents spatial filter for jth element of the input vector. The U_D passes through a GAP layer that squeezes spatial dimensions, producing a channel descriptor size that summarizes each channel. The descriptor proceeds to the excitation stage, where a channel importance layer reduces the channels, and another channel importance layer expands back to the original channels. Finally, in the scaling step weights are broadcast and multiplied elementwise with the original input feature map. Reweighting of each channel according to its learned importance and producing the output tensor, which retains the same shape as the input. The convolution layers are feed-forward networks that connect the output of the $(l-1)^{th}$ layer as input to the l^{th} layer. The convolution layer output is as expressed as Eq. (2):

$$\tilde{U}_D = \left\{ \sigma\left(E_2 \times \delta\left(E_1 \times \frac{1}{H \times W} \sum_{p,q=1}^{H,W} U_D(p,q) \right) \right) \right\} \left\{ \sum_{j=1}^{D'} k_D^j \times A_j + b_D \right\} \tag{2}$$

where $\tilde{U}_D$ is the output of the LC encoder block, σ is the Sigmoid activation, δ is the ReLU activation, E_1 first dense layer weight and E_2 second dense layer weight in the excitation step, $\frac{1}{H \times W} \sum_{p,q=1}^{H,W} U_D(p,q)$ is the global average pooling of the convolution output, p is the indexes of row kernel, q is the indexes of column kernel. After performing maxpooling, the output is expressed as Eq. (3):

$$\tilde{U}_{Dp} = maxpool(\tilde{U}_D) \tag{3}$$

where $\tilde{U}_{Dp}$ is the output of the LC encoder after the down-sampling. The $\tilde{U}_{Dp}$ is the input for both transformer block as well as CBAM block.

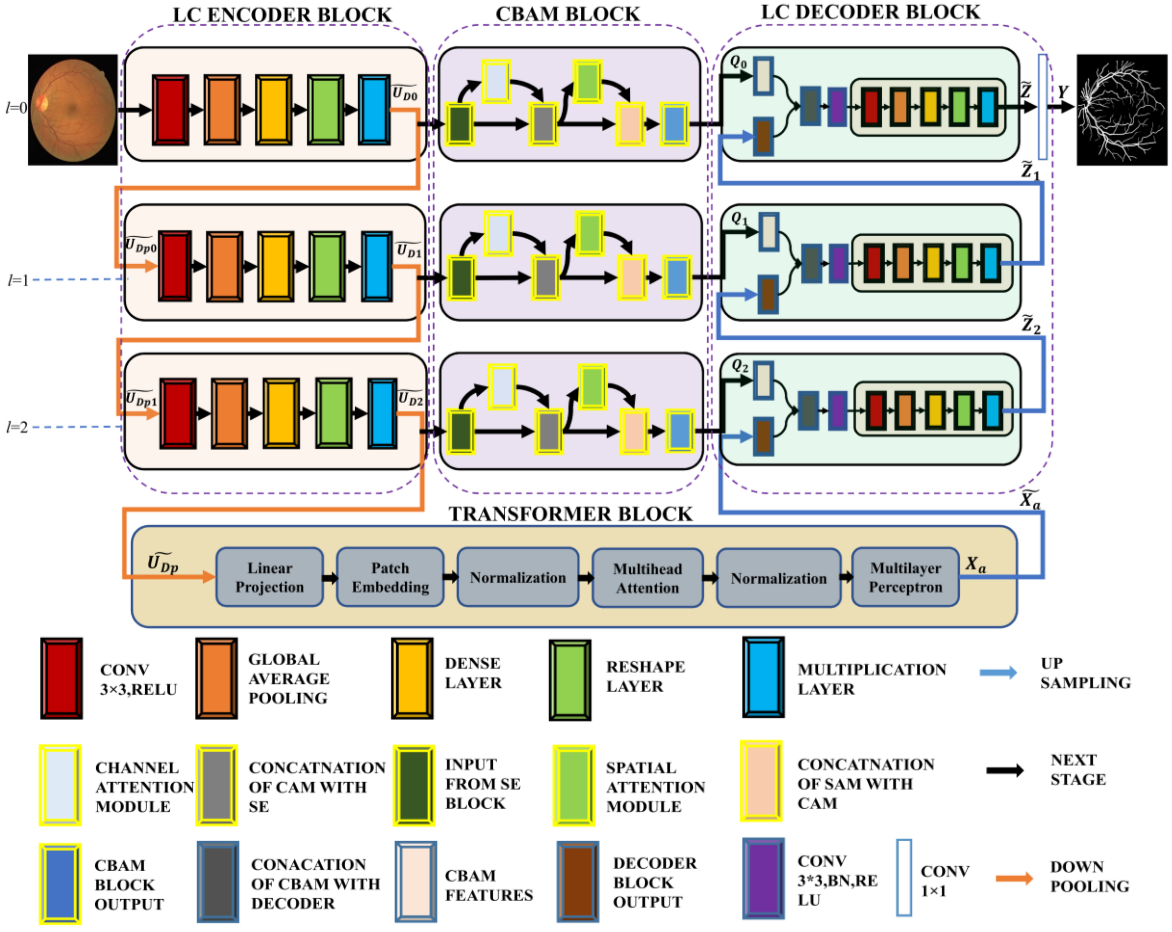

Fig. 1. Schematic portrayal of the proposed method.

B. Transformer Block

The transformer block comprises various stages, including linear projection, patch embedding, normalization, multi-head attention, and multilayer perception. The linear projection, where image patches are flattened and projected into a fixed-dimensional embedding space, converts spatial data into a format suitable for inputting to a transformer. Next, the patch embedding systematizes these projected vectors into a sequence, often affixing positional encodings to retain spatial relationships lost while flattening. The embedded sequence then undergoes normalization, typically layer normalization, which balances training and secures consistent feature scaling. This normalized input is applied into the multihead attention module. Multiple attention heads concurrently learn to focus on various parts of the image, capturing diverse contextual relationships across patches. The output of the attention mechanism is again normalized, sustains training stability, and prepares the data for the final stage. Lastly, the Multi-Layer Perceptron (MLP) applies to a feed-forward neural network to refine and transform the attended features, facilitating the model to learn complex patterns and make predictions. These stages are given as Eq. (4):

$$X_a = softmax\left(\frac{\tilde{U}_{Dp} \times W_{q,k,v}}{\sqrt{d}}\right) \times V \tag{4}$$

where X_a represents the output of the transformer block, V is the actual content carried by each token, $W_{q,k,v}$ are weights to project input embedding into queries, keys, and values, d embedding dimensions.

C. CBAM Block

In our proposed architecture, we introduced the CBAM attention mechanism from the LC encoder to the LC decoder block to improve the reconstruction of predicted images and reduce computational complexity, without degrading performance metrics. Furthermore, the LC encoder to LC decoder block performs cross-channel recalibration in retinal vessels and explicitly preserves object boundaries better than existing methods. The CBAM block enhances feature representations by sequentially applying CAM and SAM. First, the input feature map $\widetilde{U}_{Dp}$ is processed through both avgpooling and max pooling operations along the spatial dimensions, producing two descriptors that capture global context. These pooled features are passed through a shared MLP consisting of two fully connected layers, which learn to emphasize informative channels. The outputs of the MLP are summed and activated by a sigmoid function to generate the channel attention map C_{attn}, which is then multiplied elementwise with the original feature map to refine it along the channel dimension. The expressions are represented as Eq. (5):

$$C_{attn} = \sigma\left(P_2\left(\delta\left(P_1\left(avgpool\left(\tilde{U}_{Dp}\right)\right)\right)\right) + \left(P_2\left(\delta\left(P_1\left(avgpool\left(\tilde{U}_{Dp}\right)\right)\right)\right)\right) \tag{5}$$

where C_{attn} is the activate channel attention, P_1 represents the first dense layer, P_2 represents the second dense layer, δ is the ReLU activation function, σ is the sigmoid activation function. Next, the channel-refined feature undergoes spatial attention processing such as avgpooling and max pooling. The avgpooling and max pooling are applied along the channel axis to produce two 2D maps that highlight spatial cues. The maps are concatenated and fed into a convolutional layer to produce the spatial attention map S_{attn}. The expressions are given as Eq. (7):

$$S_{cat} = avgpool\left(C_{attn} \otimes \tilde{U}_{Dp}, asis = T\right) + maxpool\left(C_{attn} \otimes \tilde{U}_{Dp}, asis = T\right) \tag{6}$$

$$S_{attn} = \sigma\left\{conv(S_{cat})\right\} \tag{7}$$

where S_{attn} is the spatial attention, S_{cat} is a concanated connection of spatial avgpooling and spatial maxpooling, $\otimes$ denotes element-wise multiplication. Finally, this map is multiplied with the channel refined feature to yield the output of CBAM, which is now enriched with both channel wise and spatially aware attention, improving the networks ability to focus on relevant features represented as Eq. (8):

$$Q = \left\{S_{attn}\left(C_{attn} \otimes \tilde{U}_{Dp}, asis = T\right)\right\} \otimes \left\{C_{attn} \otimes \tilde{U}_{Dp}, asis = T\right\} \tag{8}$$

where, Q is the refined feature map output of the CBAM block.

D. LC Decoder Block

The LC decoder block is an enhanced, lighter network that can achieve the same accuracy as a much deeper network. The LC decoder block consists of three stages such as attention-guided feature fusion, deep feature extraction, and the LC encoder block. The attention-guided feature fusion fuses up sampled decoder feature outputs with CBAM refined features to produce context-aware, detail-preserving representations. At the deepest layer of the transformer block, specifically, the final output from multilayer perceptron block, we apply a transposed convolution to double the spatial resolution of the feature map. The upsampled features are then processed through a series of 3 × 3 convolutional layers. This upsampling and refinement procedure is repeated across subsequent stages until the resolution matches that of the original input. The feature map X_a which is the output of the transformer block, is up sampled, and it represented as Eq. (9):

$$\tilde{X}_a = \text{up sampling}(X_a) \tag{9}$$

where $\tilde{X}_a$ represents up sampled output of the transformer block. The up sampling output and CBAM block output are concanated. The combined output is given as Eq. (10):

$$Z = \tilde{X}_a + Q \tag{10}$$

where Z is the combined output of up sampling and CBAM block. The Z is applied to the max pooling layer. The output of the LC decoder block is represented as Eq. (11):

$$\tilde{Z} = \tilde{U}_D \left\{ maxpool(z) \right\} \tag{11}$$

where $\tilde{Z}$ is the final feature representation of the LC decoder block, it is passed through a convolutional layer to produce pixel-level semantic segmentation given as Eq. (12):

$$Y = \sigma\left(conv\left(\tilde{Z} \right) \right) \tag{12}$$

where Y is the output of the proposed model. The output has 1 channel, due to the filter size of 1 for binary segmentation. The proposed approach refines patch embedding with channel and spatial attention, enabling the model to focus on informative regions. The proposed model achieves more accurate and interpretable representations. The proposed algorithm is explained as follows.

Algorithm for the Proposed Methodology

Input: Retinal image $A \in R^{3\times w\times h}$; parameters α, β, k, σ_1, σ_2, maximum iterations N_{max}.
Output: Predicted binary vessel mask $Y(A)$.

1. Construct Mapping Functions
   Construct two mapping functions, δ_L and δ_U, parameterized by hyper parameters, according to rules or equations similar to Eqs. (10) and (11) in the reference.
2. Construct Mapping Functions
   Construct two mapping functions, δ_L and δ_U, parameterized by hyper parameters, σ_1, σ_2 according to rules or equations similar to Eqs. (10) and (11) in the reference.
3. Generate Pseudo Masks
   Generate lower and upper pseudo vessel masks:
   $M_L = \delta_L(A)$, $M_U = \delta_U(A)$
4. Set iteration counter $n = 1$.
5. Iterative Model Training
   While $n \leq N_{max}$:
6. Construct Mapping Functions
   Construct two mapping functions, δ_L and δ_U, parameterized by hyper parameters, σ_1, σ_2 according to rules or equations similar to Eqs. (10) and (11) in the reference.
7. Generate Pseudo Masks
   Generate lower and upper pseudo vessel masks: $M_L = \delta_L(A)$, $M_U = \delta_U(A)$
8. Set iteration counter $n = 1$.
9. Iterative Model Training
   While $n \leq N_{max}$:
   a. Update Model: Preprocess A using grayscale conversion, gamma correction, CLAHE, resizing, and augmentation:
   $A_{aug} = X(CLAHE(Gamma(Gray(A))))$
   Apply SE encoder, transformer/CBAM encoder as described in previous steps:
   $F_{enc} = Encoder(A_{aug}, \Theta n)$
   Use M_L and M_U to update the deep model parameters Θn via a relevant optimization (e.g., loss defined by mask bounds):
   $\Theta n + 1 = Optimize(F_{enc}, M_L, M_U, \alpha, \beta, k)$.
   b. Increment: $n = n+1$
10. Obtain Pixel-Level:
    Features After training, extract pixel-level features: $x = E(X; \Theta final)$
    where E is the final trained encoder-decoder function.
11. Calculate Foreground Mask:
    Use a game theoretic filter or a suitable rule to sharpen the foreground vessel mask:
    $Y(A) = G(a)$
12. Return Output:
    Return the optimized vessel mask $Y(A)$ as the final prediction.

IV. Experimental Results and Analysis

A. Materials

The experiments are carried out on four public datasets and one mixed dataset in order to conduct a thorough evaluation of the proposed approach in contrast to the state of the art. Table I illustrates an overview of retinal vessel segmentation in various datasets, which consist of fundus images with different numbers and resolutions. As part of the validation process for retinal vessel segmentation algorithms, datasets serve as crucial benchmarks.

1) DRIVE

These images were collected from diabetic retinopathy patients in the Netherlands in the Digital Retinal Images for Vessel Extraction (DRIVE) dataset [50]. There are 400 images of diabetic patients between the ages of 25 and 90 in the dataset. Among 40 randomly selected images, 7 showed mild symptoms of diabetic retinopathy. In the proposed model, we utilized a training DRIVE dataset comprising 20 images, each accompanied by corresponding ground truth data.

2) STARE

Structured analysis of the retina (STARE) was developed by Michael Goldbaum, M.D., at the University of California, San Diego [15]. In the STARE collection, 400 excellent RGB fundus images with annotations are available for the study of retinal diseases. Since each image is labeled for vasculature, optic disc, and other anatomical features, it is useful for glaucoma, AMD, and DR research. Although the complete dataset contains 400 images, we utilized the subset for which Adam Hoover's annotation provided 20 ground truth images.

3) CHASE_DB1

The Child Heart and Health Study in England (CHASE) dataset [51] covers 200 primary schools from London, Birmingham, and Leicester. CHASE_DB1 consists of 28 excellent RGB fundus images illustrating a variety of vascular and pathological features. It is often used in deep learning studies to segment retinal vessels. From a total set of 28 RGB images, a subset of 20 images was selected for training purposes.

4) HRF

A High-Resolution Fundus (HRF) dataset [52] was released by the GE-CZ research group in 2013. A high-resolution fundus camera captured 45 high-resolution retinal fundus images for the HRF collection. The images depict both healthy and sick retinas, illustrating diseases such as macular degeneration and diabetic retinopathy. We considered all images within the dataset for implementation in our proposed model.

5) MIXED

The MIXED dataset was obtained from the 4 datasets. A total of 100 retinal fundus images is included, those from people with various ailments, such as diabetic retinopathy. The dataset contains both normal and pathological cases and is used for research in vascular segmentation, disease analysis, and DR detection.

TABLE I. SUMMARY OF THE RETINAL VESSEL SEGMENTATION DATASETS

No.	Dataset	Image Type	# of Images	Resolution	Description
1	DRIVE [50]	Retinal fundus	20	768 × 584	7 images with mild diabetic retinopathy
2	STARE [15]	RGB fundus	20	~700 × 605 approx.	Used for glaucoma, AMD, and DR research
3	CHASE_DB1 [51]	RGB fundus	20	1280 × 960	Known for vascular and pathological variation
4	HRF [52]	High-resolution fundus	45	High resolution	Covers healthy and diseased retinas
5	MIXED	RGB fundus	100	512 × 512	Samples from each dataset

B. Experimental Settings

All experiments were performed on a system equipped with an NVIDIA GeForce RTX 3070 GPU and a 12th-generation Intel Core CPU. Training was accomplished by using the TensorFlow library, with mixed precision enabled. To ensure the integrity and reproducibility of our results, the dataset was partitioned into a stratified 80% training, 10% validation, and 10% independent test split, with the test set held out entirely until final evaluation. To mitigate the data-hungry nature of Transformers, our hybrid architecture restricts self-attention to the high-level bottleneck to minimize learnable parameters, while integrating CBAM-gated skip connections 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 1×10^{-4}, a batch size of 2, and the root mean square propagation optimizer and cosine annealing schedule throughout.

C. Evaluation Metrics

In image segmentation tasks, each pixel in the output can be categorized into one of four categories: True Positives (TP), which are accurately recognized target pixels; True Negatives (TN), which are accurately recognized non-target pixels; False Positives (FP), which are non-target pixels incorrectly labeled as target; and False Negatives (FN), which are target pixels that the model misses to detect [53]. Simultaneously, the outcomes form the entire prediction data used to assess model performance. Common metrics derived from this full set include accuracy (overall correctness), sensitivity or recall (ability to detect target pixels), specificity (ability to avoid false alarms), Dice score (measuring overlap between predicted and actual target regions), and the F1-Score (balancing precision and recall). These metrics are crucial for evaluating the effectiveness of vascular segmentation algorithms in identifying and isolating relevant anatomical structures. Table II illustrates the mathematical expressions for the recall, accuracy, Intersection over Union (IoU), F1-Score, and precision [54].

D. 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 tracking ability of diseases over time. Table III manifests that the model delivers solid performance overall. The U-Net architecture lacks sufficient capacity to capture global context, particularly in deeper layers. Attention U-Net added complexity might not justify the marginal improvement. Multires U-Net may require additional tooling for attention. PA-Net has a poor backbone feature map.

Overall, the proposed model consistently outperforms baseline methods across the dataset. The proposed model is a hybrid architectural design that enables an efficient approach to dataset-specific issues such as domain variance in fundus images and vascular topology, for the DRIVE datasets. Table IV shows that the STARE dataset, the proposed model achieved the highest accuracy of

0.935, driven by its multi-scale attention and transformer fusion design, which effectively distinguishes vessels from background regions. It also obtained the best F1-Score of 0.979 by maintaining an excellent balance between precision and recall through hybrid attention mechanisms. The AVA Net performed best in terms of IoU, reaching 0.960, since its advanced vascular attention mechanism captures fine-grained vessel patterns with high spatial consistency. For recall, the proposed model again led with 0.989, as its transformer-based feature fusion enhances the identification of thin and low-contrast vessels. Simultaneously, Multires U-Net achieved the highest precision of 0.975 due to its skip connections and multi-resolution feature extraction, which reduces false positives while preserving vessel continuity. Considering all factors, the proposed model influenced most metrics by merging multi-scale fusion and attention, while Multires U-Net and AVA Net provided strengths in precision and structure overlap, respectively.

TABLE II. EVALUATION OF METRICS AND MATHEMATICAL FORMULAS

Metric	Description	Method
Accuracy	Measures how many pixels out of all the pixels were correctly identified.	$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$
Precision	Calculates the relevance rate of retrieved instances.	$Precision = \frac{TP}{TP + FP}$
Recall	Calculates the proportion of relevant instances recovered over all relevant instances.	$Recall = \frac{TP}{TP + FN}$
Intersection over Union (IoU)	By comparing the ground truth and predicted regions, it penalizes false negatives and false positives.	$IoU = \frac{TP}{TP + FP + FN}$
F1-Score	Combining Precision and Recall achieving a balance.	$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

TABLE III. COMPARATIVE RESULTS FOR DIFFERENT SEGMENTATION METHODS USING THE DRIVE DATASET

S.No.	Year	Model	Accuracy	F1-Score	IoU	Recall	Precision
1	2015	U-Net [13]	0.864	0.943	0.894	0.981	0.903
2	2018	Attention U-Net [14]	0.879	0.946	0.899	0.943	0.899
3	2019	Multires U-Net [15]	0.859	0.940	0.924	0.952	0.970
4	2020	BIFPN [20]	0.842	0.948	0.902	0.969	0.902
5	2021	U-NetR [30]	0.874	0.970	0.943	0.968	0.913
6	2022	AVA Net [21]	0.888	0.975	0.951	0.981	0.967
7	2023	AFFD Net [22]	0.870	0.974	0.951	0.981	0.950
8	2024	PA-Net [23]	0.880	0.952	0.908	0.932	0.908
9	2024	RetinaLiteNet [35]	0.841	0.941	0.889	0.932	0.951
10	2024	DA-TransUNet [55]	0.883	0.974	0.949	0.979	0.968
11	2025	LHUVT [24]	0.852	0.955	0.914	0.942	0.964
12	2025	LiViT Net [39]	0.855	0.957	0.919	0.947	0.968
13	2025	DBBCS [56]	0.882	0.974	0.949	0.977	0.971
14	2025	Proposed Model	0.890	0.976	0.953	0.983	0.977

TABLE IV. COMPARATIVE RESULTS FOR DIFFERENT SEGMENTATION METHODS USING THE STARE DATASET

S.No.	Year	Model	Accuracy	F1-Score	IoU	Recall	Precision
1	2015	U-Net [13]	0.823	0.900	0.823	0.882	0.926
2	2018	Attention U-Net [14]	0.932	0.965	0.932	0.899	0.932
3	2019	Multires U-Net [15]	0.895	0.966	0.936	0.959	0.975
4	2020	BIFPN [20]	0.934	0.966	0.935	0.983	0.935
5	2021	U-NetR [30]	0.890	0.962	0.928	0.954	0.972
6	2022	AVA Net [21]	0.919	0.979	0.96	0.983	0.976
7	2023	AFFD Net [22]	0.898	0.967	0.938	0.961	0.975
8	2024	PA-Net [23]	0.933	0.965	0.933	0.965	0.933
9	2024	RetinaLiteNet [35]	0.832	0.925	0.862	0.894	0.961
10	2024	DA-TransUNet [55]	0.921	0.979	0.959	0.984	0.974
11	2025	LHUVT [24]	0.885	0.955	0.914	0.947	0.964
12	2025	LiViT Net [39]	0.891	0.962	0.927	0.954	0.971
13	2025	DBBCS [56]	0.921	0.980	0.962	0.988	0.973
14	2025	Proposed Model	0.935	0.979	0.953	0.989	0.964

Table V illustrates that the CHASE_DB1 dataset, the Attention U-Net achieved the highest overall accuracy of 0.930, primarily due to its affinity dense connections that improve contextual learning and maintain vessel continuity with minimal errors. The Proposed Model led in both F1-Score and IoU, with values of 0.983 and 0.971, respectively, which can be attributed to its integration of multi scale attention and transformer fusion that optimizes both overlap and overall segmentation quality. For recall, the proposed model also performed best at 0.984, as its hybrid architecture is particularly effective for detecting fine and low-contrast vessels. In terms of precision, Multires U-Net stood out with a value of 0.984, owing to its multi-resolution feature blocks and skip connections that restrict false positives, ensuring highly accurate vessel predictions. Overall, the proposed model dominated

several core metrics with its advanced design, while AFFD Net and Multires U-Net excelled in accuracy and precision thanks to their specialized connectivity and feature extraction strategies.

Table VI demonstrates that the HRF dataset performance highlights the proposed model design strengths. The proposed model leads to an F1-Score of 0.983, IoU of 0.971, recall of 0.984, and precision of 0.974, manifesting its valuable combination of transformer fusion and multi-scale attention that enhances boundary precision and vessel detection. Overall, the intense metrics across all models demonstrate that the HRF dataset enables consistently accurate retinal vessel segmentation, with newer models offering peripheral yet meaningful gains in obtaining fine structures and minimizing errors.

TABLE V. COMPARATIVE RESULTS FOR DIFFERENT SEGMENTATION METHODS USING THE CHASE_DB1 DATASET

S.No.	Year	Model	Accuracy	F1-Score	IoU	Recall	Precision
1	2015	U-Net [13]	0.927	0.962	0.927	0.96	0.927
2	2018	Attention U-Net [14]	0.930	0.964	0.930	0.927	0.930
3	2019	Multires U-Net [15]	0.858	0.952	0.908	0.922	0.984
4	2020	BIFPN [20]	0.929	0.963	0.929	0.919	0.929
5	2021	U-NetR [30]	0.872	0.959	0.922	0.939	0.981
6	2022	AVA Net [21]	0.917	0.982	0.965	0.984	0.979
7	2023	AFFD Net [22]	0.913	0.982	0.965	0.982	0.982
8	2024	PA-Net [23]	0.927	0.962	0.927	0.962	0.927
9	2024	RetinaLiteNet [35]	0.861	0.941	0.891	0.924	0.961
10	2024	DA-TransUNet [55]	0.902	0.974	0.949	0.975	0.972
11	2025	LHUVT [24]	0.881	0.960	0.924	0.951	0.969
12	2025	LiViT Net [39]	0.878	0.959	0.923	0.949	0.971
13	2025	DBBCS [56]	0.910	0.981	0.963	0.984	0.978
14	2025	Proposed Model	0.819	0.983	0.971	0.984	0.984

TABLE VI. COMPARATIVE RESULTS FOR DIFFERENT SEGMENTATION METHODS USING THE HRF DATASET

S.No.	Year	Model	Accuracy	F1-Score	IoU	Recall	Precision
1	2015	U-Net [13]	0.927	0.962	0.927	0.96	0.927
2	2018	Attention U-Net [14]	0.930	0.964	0.930	0.927	0.930
3	2019	Multires U-Net [15]	0.858	0.952	0.908	0.922	0.984
4	2020	BIFPN [20]	0.929	0.963	0.929	0.919	0.929
5	2021	U-NetR [30]	0.872	0.959	0.922	0.939	0.981
6	2022	AVA Net [21]	0.917	0.982	0.965	0.984	0.979
7	2023	AFFD Net [22]	0.913	0.982	0.965	0.982	0.982
8	2024	PA-Net [23]	0.927	0.962	0.927	0.962	0.927
9	2024	RetinaLiteNet [35]	0.861	0.941	0.891	0.924	0.961
10	2024	DA-TransUNet [55]	0.902	0.974	0.949	0.975	0.972
11	2025	LHUVT [24]	0.881	0.960	0.924	0.951	0.969
12	2025	LiViT Net [39]	0.878	0.959	0.923	0.949	0.971
13	2025	DBBCS [56]	0.910	0.981	0.963	0.984	0.978
14	2025	Proposed Model	0.819	0.983	0.971	0.984	0.984

TABLE VII. COMPARATIVE RESULTS FOR DIFFERENT SEGMENTATION METHODS USING THE MIXED DATASET

S.No.	Year	Model	Accuracy	F1-Score	IoU	Recall	Precision
1	2015	U-Net [13]	0.865	0.942	0.901	0.960	0.926
2	2018	Attention U-Net [14]	0.861	0.930	0.879	0.956	0.907
3	2019	Multires U-Net [15]	0.867	0.941	0.899	0.961	0.923
4	2020	BIFPN [20]	0.891	0.932	0.883	0.986	0.883
5	2021	U-NetR [30]	0.868	0.940	0.898	0.963	0.921
6	2022	AVA Net [21]	0.862	0.941	0.899	0.957	0.927
7	2023	AFFD Net [22]	0.861	0.94	0.899	0.955	0.928
8	2024	PA-Net [23]	0.867	0.946	0.906	0.961	0.933
9	2024	RetinaLiteNet [35]	0.842	0.926	0.874	0.936	0.918
10	2024	DA-TransUNet [55]	0.819	0.932	0.883	0.979	0.883
11	2025	LHUVT [24]	0.858	0.931	0.882	0.954	0.912
12	2025	LiViT Net [39]	0.865	0.938	0.895	0.961	0.919
13	2025	DBBCS [56]	0.867	0.943	0.904	0.962	0.927
14	2025	Proposed Model	0.863	0.949	0.916	0.988	0.943

Table VII shows that the proposed model accomplished the highest recall of 0.988 and precision of 0.943, indicating its potent capability to detect vessels accurately across assorted image sources while reducing false positives. It also led in F1-Score 0.949 and an IoU of 0.916, focusing on its balanced performance in both spatial overlap and sensitivity due to its multi-scale attention and transformer-based fusion. BIFPN exhibited the highest accuracy of 0.891, which can be connected to its effective bidirectional feature pyramid that merges multi-scale features well, however its lower precision of 0.883 suggests certain false-positive segmentation. PA-Net performed vigorously with a high F1-Score of 0.946 and precision of 0.933, assisting with position attention modules that improve spatial focus. Other models, such as Multires U-Net and U-NetR, preserved consistent but

gently lower performance metrics, exhibiting reliable vessel segmentation but with less prominence on reducing false positives correlated to the proposed model. Overall, the proposed model shows evident advantages in dealing with the complexity of the mixed dataset with its sophisticated attention driven architecture.

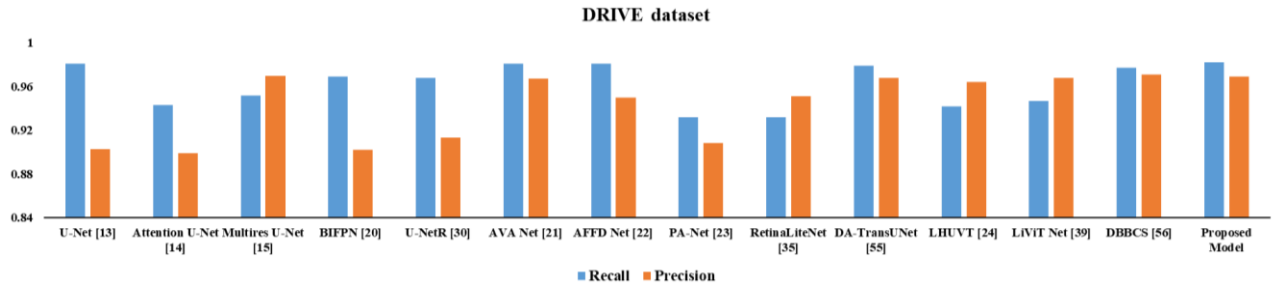

Fig. 2. Training performance compared with validation performance across epochs on the DRIVE dataset.

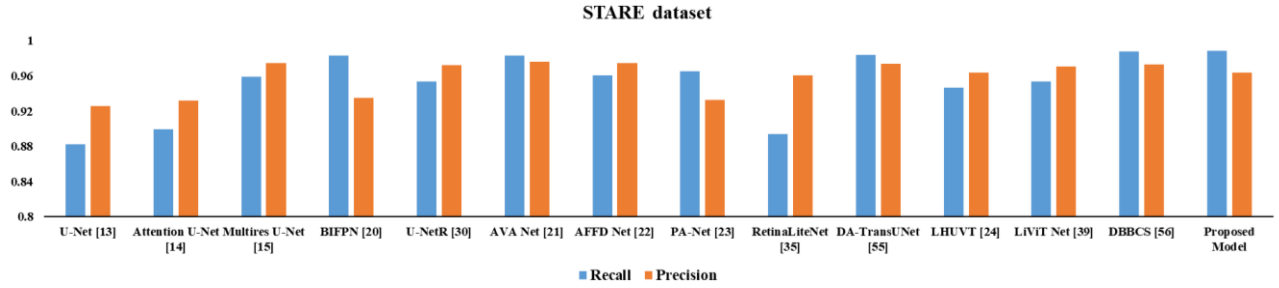

Fig. 3. Training performance compared with validation performance across epochs on the STARE dataset.

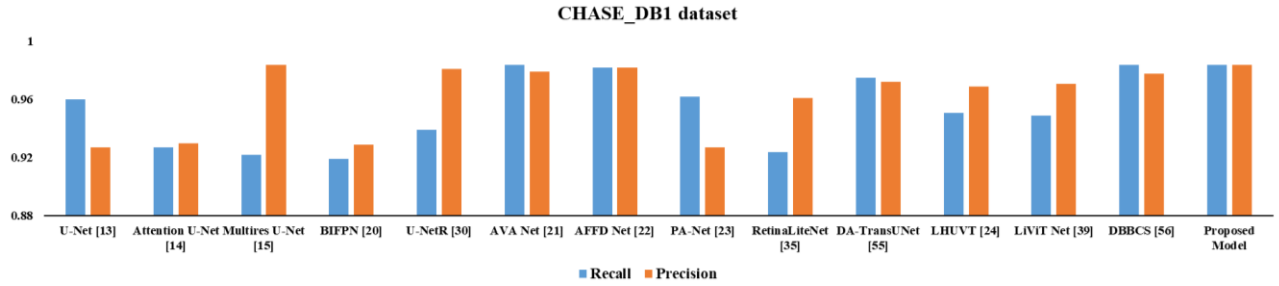

Fig. 4. Training performance compared with validation performance across epochs on the CHASE_DB1 dataset.

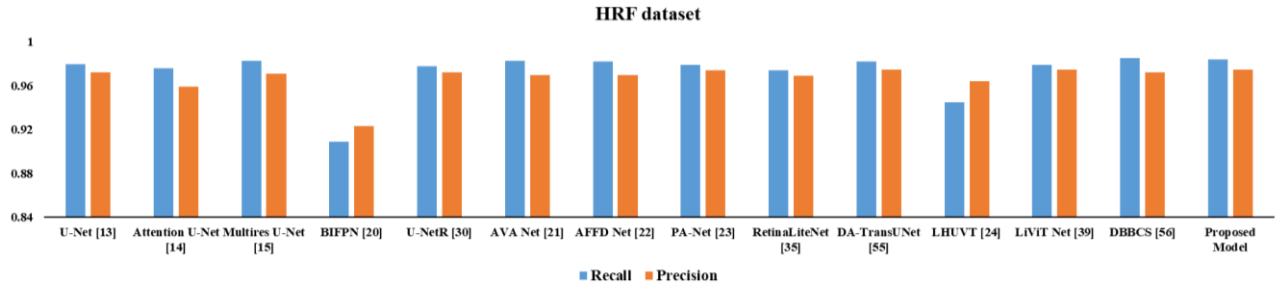

Fig. 5. Training performance compared with validation performance across epochs on the HRF dataset.

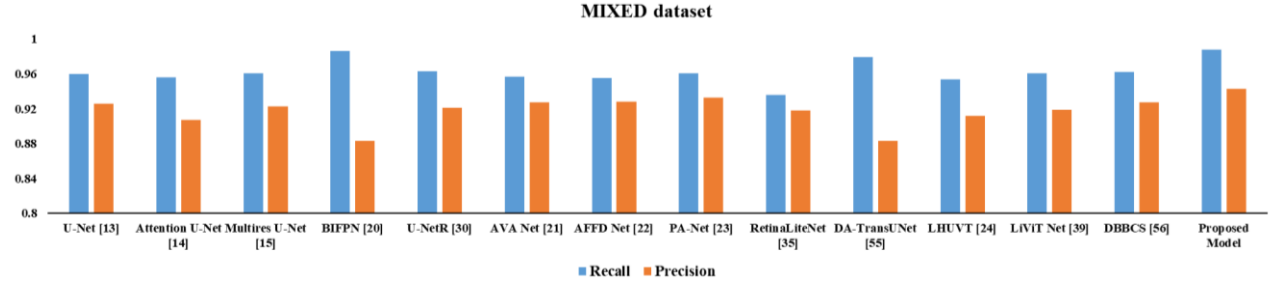

Fig. 6. Training performance compared with validation performance across epochs on the MIXED dataset.

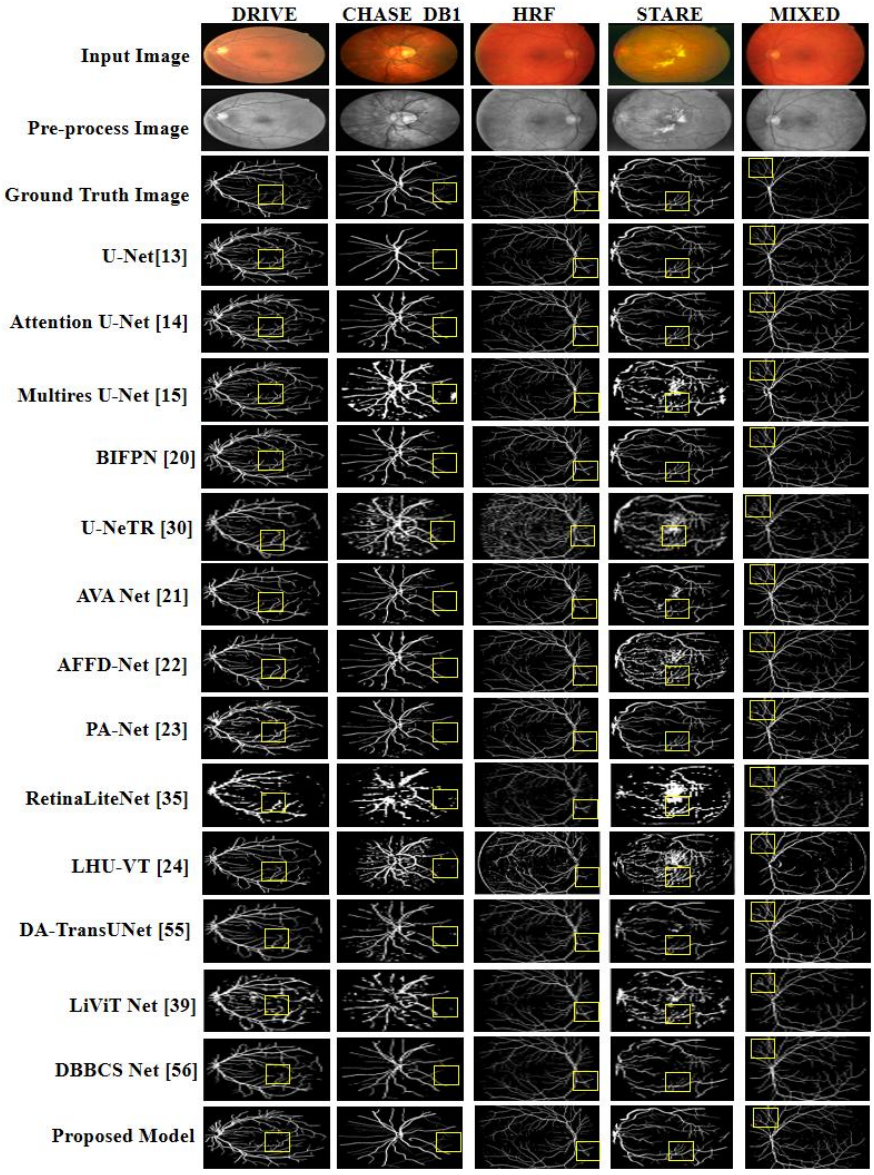

Fig. 7. Typical visual results for different methods and the proposed model with the DRIVE, CHASE_DB1, HRF, STARE, and MIXED datasets, original, ground truth, and predicted images, respectively.

TABLE VIII. ANALYSIS OF LIGHTWEIGHT EFFICIENCY METRICS OF THE RETINAL VESSEL SEGMENTATION ON VARIOUS MODELS COMPARED WITH THE PROPOSED MODEL

S.No.	Model	Flops (G)	Parameters	Avg. Lat. (s)	Memory footprint (MB)
1	U-Net [13]	385.53	3,10,55,297	0.019	518.27
2	Attention U-Net [14]	365.21	81,43,169	0.015	589.36
3	Multires U-Net [15]	125.43	72,69,294	0.080	535.86
4	BIFPN [20]	285.06	1,59,74,145	0.027	504.109
5	U-NetR [30]	38.75	39,93,69,201	0.219	1557.96
6	AVA Net [21]	242.05	2,39,46,001	0.218	434.89
7	AFFD Net [22]	146.76	8,75,909	0.460	15.53
8	PA-Net [23]	60.64	7,16,673	0.012	362.35
9	RetinaLiteNet [35]	01.96	30,337	0.006	181.03
10	DA-TransUNet [55]	213.58	17,06,309	0.049	358.791
11	LHUVT [24]	217.58	11,447,553	0.111	61.63
12	LiViT Net [39]	17.35	92,04,143	0.155	73.89
13	DBBCS [56]	446.83	3,14,04,300	0.194	258.16
14	Proposed Model	93.25	63,94,636	0.387	29.43

The comparison of the training and validation performance of various models with the proposed model is illustrated in Figs. 2–7. The output images, which include training, ground truth, and predicted images, are illustrated in Fig. 7. The image represents a comparative analysis of retinal blood vessel segmentation of various methods. Every method converts the retinal image into a binary map featuring vascular structures in white against a

black backdrop. This visual comparison is crucial for evaluating how well every approach captures the complexities of vessel patterns, which is key to progressing medical imaging and diagnosis. It manifests clearer vessel continuity, superior branch detection, and minimized noise contrasted to the other approaches. As it builds on transformer principles, the proposed method probably benefits from enhancing context awareness and, long-range feature modeling beyond local pixel neighborhoods. While existing architectures leverage complex attention mechanisms to capture global context, their hybrid Transformer-CNN architectures often incur significant computational overhead and high GPU memory usage. The existing models, despite their specialized design, frequently introduce structural redundancy through multi-path connections or dense blocks, leading to slower inference times and larger parameter counts. In contrast, our proposed model adopts a streamlined architectural approach that avoids these resource-intensive bottlenecks. Table VIII presents an analysis of lightweight efficiency metrics, including Floating-Point Operations (FLOPs), average latency, and peak memory footprint, for segmentation of retinal vessels. We observe in Table VIII that the lightweight efficiency metrics of the proposed model have fewer parameters and lower computational efficiency, making it suitable for resource-constrained environments compared to existing methods.

E. Ablation Study

To validate the effectiveness of the different configurations of the models can be interpreted as ablation variants evaluated on various datasets using standard segmentation metrics such as accuracy, F1-Score, IoU, recall, and precision. The strongest configurations in the tables would be those whose metrics highlight optimal values in bold and demonstrate a clear gain in sensitivity when additional modules are attached to the decoding stage or combined with other components, reflecting yellow regions in qualitative visualizations. Table IX shows that the three compact conditions (about 6.3 M parameters) achieve the best results overall the proposed model (LC Encoder + Transformer + CBAM+ LC Decoder) reach accuracies above 0.92, F1-Scores around 0.98, and IoU near 0.95 with very high recall close to 0.99, while keeping inference between 6 and 9 minutes, so they offer the most favorable balance between accuracy, overlap quality, parameter count, and processing time.

TABLE IX. ABLATION STUDY OF THE RETINAL VESSEL SEGMENTATION ON VARIOUS DATASETS

Dataset	Proposed Model Blocks	Accuracy	F1- Score	IoU	Recall	Precision	Parameters	Processing Time (min)
DRIVE	Conventional U-Net	0.894	0.943	0.894	0.988	0.903	31,055,297	10
	Transformer Block	0.874	0.970	0.943	0.968	0.973	399,369,201	9
	LC Encoder & Decoder Block	0.883	0.973	0.949	0.980	0.967	6,372,331	7
	Transformer + CBAM	0.884	0.975	0.951	0.978	0.971	6,380,482	10
	Proposed Model	0.890	0.976	0.953	0.983	0.977	6,394,636	10
CHASE _DB1	Conventional U-Net	0.927	0.962	0.927	0.960	0.927	31,055,297	10
	Transformer Block	0.872	0.959	0.922	0.939	0.981	399,369,201	6
	LC Encoder & Decoder Block	0.913	0.979	0.960	0.982	0.977	6,372,331	7
	Transformer + CBAM	0.910	0.979	0.959	0.983	0.975	6,380,482	12
	Proposed Model	0.914	0.983	0.971	0.984	0.984	6,394,636	10
HRF	Conventional U-Net	0.908	0.979	0.959	0.986	0.972	31,055,297	13
	Transformer Block	0.900	0.975	0.951	0.978	0.972	399,369,201	19
	LC Encoder & Decoder Block	0.904	0.977	0.956	0.981	0.974	6,372,331	14
	Transformer + CBAM	0.904	0.977	0.956	0.981	0.974	6,380,482	17
	Proposed Model	0.909	0.979	0.959	0.987	0.975	6,394,636	22
STARE	Conventional U-Net	0.823	0.900	0.823	0.882	0.926	31,055,297	8
	Transformer Block	0.890	0.962	0.928	0.954	0.972	399,369,201	8
	LC Encoder & Decoder Block	0.922	0.976	0.954	0.987	0.966	6,372,331	6
	Transformer + CBAM	0.901	0.968	0.939	0.965	0.972	6,380,482	10
	Proposed Model	0.925	0.976	0.953	0.989	0.964	6,394,636	9
MIXED	Conventional U-Net	0.865	0.942	0.901	0.960	0.926	31,055,297	15
	Transformer Block	0.868	0.940	0.898	0.963	0.921	399,369,201	25
	LC Encoder & Decoder Block	0.874	0.944	0.906	0.970	0.922	6,372,331	16
	Transformer + CBAM	0.876	0.946	0.908	0.972	0.924	6,380,482	36
	Proposed Model	0.863	0.939	0.896	0.958	0.923	6,394,636	50

TABLE X. PERFORMANCE METRICS OF THE PROPOSED MODEL USING 5-FOLD CROSS-VALIDATION WITH THE MIXED DATASET

Fold No.	Accuracy	F1-Score	IoU	Recall	Precision
Fold 1	0.92	0.90	0.87	0.92	0.90
Fold 2	0.87	0.85	0.82	0.87	0.87
Fold 3	0.92	0.90	0.87	0.92	0.90
Fold 4	0.86	0.88	0.84	0.87	0.86
Fold 5	0.92	0.90	0.87	0.92	0.90

Fig. 8 shows that both the training and validation Dice coefficients show a steady increase, starting from approximately 0.18 and reaching final values of 0.80 and 0.73, respectively. The close alignment between the two curves suggests that the model is learning effectively and generalizing well to unseen data without significant

overfitting. Regarding loss graphs steady decrease in both training and validation loss, with the epoch increasing consistently. The two curves track closely together, indicating that the model is generalizing well to the MIXED dataset and is not simply memorizing the training data. To ensure statistical reliability with a limited dataset, a 5-fold cross-validation strategy was employed with the MIXED dataset. Table X shows that the data were partitioned into five subsets, with four folds used for training and one for independent validation in each iteration. The proposed model achieves the best performance by increasing micro vessel segmentation while reducing miss segmentation, by utilizing the attention mechanism and LC blocks which further demonstrate rationality and effectiveness.

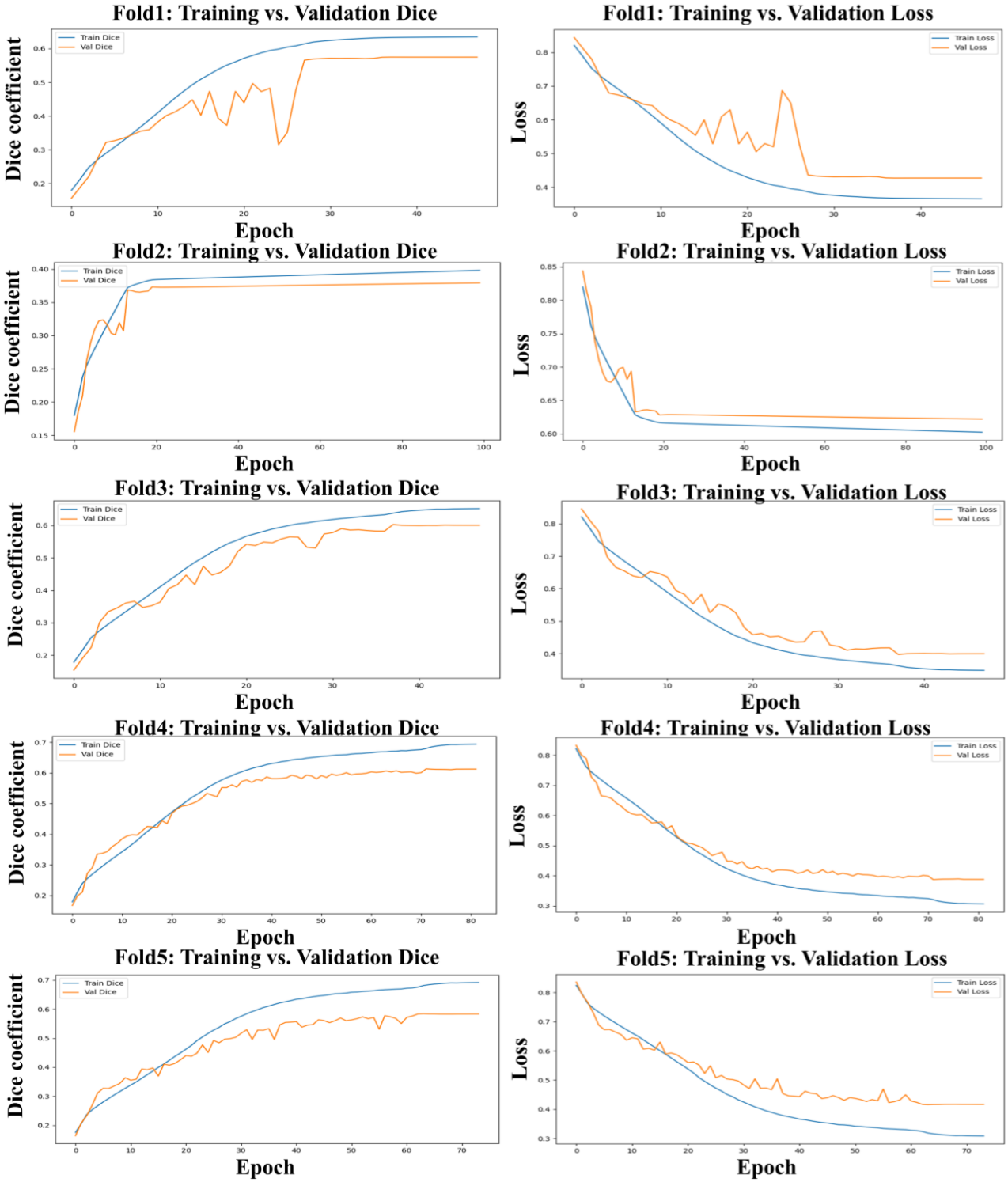

Fig. 8. Epoch-wise training and validation performance on the MIXED dataset.

V. Conclusion

This paper presents a unified, systematically driven proposed framework covering novel architectures transformers used as a bottleneck, an LC encoder block, a CBAM attention block and an model by exploring refined optimization techniques.

LC decoder block. The model efficiently captures intricate image structures without excessive parameters, merging the benefits of CNNs and Transformers. Through meticulous analysis across various scales and dynamic allocation of attention to misclassified pixels, the proposed framework preserves vessel structure continuity. The proposed architecture is likely to overcome significant current limitations of state-of-the-art methods with respect to global context modeling, resilience to imaging artifacts, and time-related disease tracking. Quantitative measures of the proposed approach, against benchmark datasets, demonstrate potential to enhance medical diagnosis through precise retinal vessel segmentation. In future endeavors, we intend to enhance the precision of the proposed model by exploring refined optimization techniques. We aim to incorporate advanced regularization methods and structures, reducing computational costs without compromising performance. We are also interested in adapting the model to various medical imaging fields to ensure flexibility and versatility. We are interested in introducing real-time data enhancement to improve the model's resilience.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

Conceptualization, Srinivas Rangu; methodology, Srinivas Rangu and Nagaraj Yamanakkanavar; formal analysis, Srinivas Rangu and Nagaraj Yamanakkanavar; investigation, Srinivas Rangu and Nagaraj Yamanakkanavar; writing—original draft preparation, Srinivas Rangu; writing—review and editing, Srinivas Rangu and Nagaraj Yamanakkanavar. All authors have read and agreed to the published version of the manuscript.

References

[1] J. Li *et al*., "MDAC-Net: A multi-scale dense attention cascade network for retinal vessel segmentation from fundus images", *IEEE Access*, vol. 13, 2025.

[2] V. T. Q. Huy and C. M. Lin, "D2CBDAMAttUnet: Dual-decoder convolution block dual attention Unet for accurate retinal vessel segmentation from fundus images," *IEEE Access*, vol. 13, 2025.

[3] A. Bhati, S. Jain, N. Gour, P. Khanna, A. Ojha, and N. Werghi, "Dynamic statistical attention-based lightweight model for retinal vessel segmentation: DyStA-RetNet," *Comput. Biol. Med.*, vol. 186, 2025.

[4] G. Läthén, J. Jonasson, and M. Borga, "Blood vessel segmentation using multi-scale quadrature filtering," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 762–767, 2010.

[5] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Trans. Med. Imag*., vol. 8, no. 3, pp. 263–269, 1989.

[6] F. Farokhian and H. Demirel, "Blood vessels detection and segmentation in retina using gabor filters," in *Proc. High Capacity Opt. Netw. Emerg./Enabling Technol*., 2013, pp. 104–108.

[7] Y. Hu *et al*., "Mendelian randomization highlights causal association between genetically increased C-reactive protein levels and reduced Alzheimer's disease risk," *Alzheimer's Dement.*, vol. 18, no. 10, pp. 2003–2006, 2022.

[8] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag*., vol. 19, no. 3, pp. 203–210, 2000.

[9] B. S. Y. Lam, Y. Gao, and A. W. C. Liew, "General retinal vessel segmentation using regularization-based multiconcavity modelling," *IEEE Trans. Med. Imag*., vol. 29, no. 7, pp. 1369–1381, 2010.

[10] H. Jiang, B. He, D. Fang, Z. Ma, B. Yang, and L. Zhang, "A region growing vessel segmentation algorithm based on spectrum information," *Comput. Math. Methods Med.*, vol. 2013, 123456, 2013.

[11] T. Mapayi, S. Viriri, and J. R. Tapamo, "Adaptive thresholding technique for retinal vessel segmentation based on GLCM-energy information," *Comput. Math. Methods Med*., vol. 2015, 123456, 2015.

[12] V. M. Saffarzadeh, A. Osareh, and B. Shadgar, "Vessel segmentation in retinal images using multi-scale line operator and K-means clustering," *J. Med. Signals Sens.*, vol. 4, no. 2, pp. 122–129, 2014.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[14] O. Oktay *et al*., "Attention U-Net: Learning where to look for the pancreas," arXiv Preprint, arXiv:1804.03999, 2018. https://doi.org/10.48550/arXiv.1804.03999

[15] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[16] S. Seo, H. Yoon, S. Kim, and J. Lee, "Full-scale representation guided network for retinal vessel segmentation," *BMC Medical Imaging*, vol. 25, no. 1, 484, 2025.

[17] K. Radha and Y. Karuna, "Gabor-modulated depth separable convolution for retinal vessel segmentation in fundus images," *Comput. Biol. Med.*, vol. 188, 2025.

[18] J. Li, Q. Cheng, and C. Wu, "GViT-RSNet: A retinal vessel segmentation network using graph convolutional attention and multi-scale vision transformer," *Pattern Recognit. Lett.*, vol. 189, pp. 182–187, 2025.

[19] M. E. Martínez-Pérez *et al*., "Retinal blood vessel segmentation by means of scale-space analysis and region growing," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 1999, pp. 90–97.

[20] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.

[21] M. Abtahi *et al*., "An open-source deep learning network AVA-Net for arterial-venous area segmentation in optical coherence tomography angiography," *Commun. Med.*, vol. 3, no. 1, 2023.

[22] Z. Xiang *et al*., "AFFD-Net: A dual-decoder network based on attention-enhancing and feature fusion for retinal vessel segmentation," *IEEE Access*, vol. 11, pp. 45871–45887, 2023.

[23] X. Luo, L. Peng, Z. Ke, J. Lin, and Z. Yu, "PA-Net: A hybrid architecture for retinal vessel segmentation," *Pattern Recognit.*, vol. 161, 2025.

[24] W. Ahmed and P. Liatsis, "LHU-VT: A lightweight hypercomplex U-Net with vessel thickness-guided dice loss for retinal vessel segmentation," *Comput. Biol. Med.*, vol. 185, 109470, 2025.

[25] A. Dasgupta and S. Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, 2017, pp. 248–251.

[26] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2019.

[27] J. Zhuang, "LadderNet: Multi-path networks based on U-Net for medical image segmentation," arXiv Preprint, arXiv:1810.07810, 2018.

[28] J. Yun and N. Tan, "Retinal vessel segmentation based on conditional deep convolutional generative adversarial networks," arXiv Preprint, arXiv:1805.04224, 2018.
[29] Q. Jin, Z. Meng, and G. Chen, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, 2019.
[30] A. Hatamizadeh *et al*., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 574–584.
[31] D. Chen, W. Yang, L. Wang, S. Tan, J. Lin, and W. Bu, "PCAT-UNet: UNet-like network fused convolution and transformer for retinal vessel segmentation," *PLoS ONE*, vol. 17, no. 1, e0262689, 2022.
[32] L. Liang, B. Lu, J. Wu, Y. Li, and X. Sheng, "SFIT-Net: Spatial reconstruction feature interaction transformer retinal vessel segmentation algorithm," *Biomed. Signal Process. Control*, vol. 106, 2025.
[33] N. Lv, L. Xu, Y. Chen, W. Sun, J. Tian, and S. Zhang, "TCDDU-Net: Combining transformer and convolutional dual-path decoding U-Net for retinal vessel segmentation," *Sci. Rep.*, vol. 14, no. 1, 2024.
[34] L. Jiang *et al*., "TransDualSegNet: Transformer dual segment network for retinal vasculature segmentation in OCT," *Quant. Imag. Med. Surg.*, vol. 15, no. 10, 2025.
[35] M. Mehmood, M. Alsharari, S. Iqbal, I. Spence, and M. Fahim, "RetinaliteNet: A lightweight transformer based CNN for retinal feature segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 2454–2463.
[36] J. Lin, X. Huang, H. Zhou, Y. Wang, and Q. Zhang, "Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images," *Med. Image Anal.*, vol. 89, 2023.
[37] A. Ke, J. Luo, and B. Cai, "UNet-like network fused swin transformer and CNN for semantic image synthesis," *Sci. Rep.*, vol. 14, no. 1, 16761, 2024.
[38] M. Jiang, Y. Zhu, and X. Zhang, "CoVi-Net: A hybrid convolutional and vision transformer neural network for retinal vessel segmentation," *Comput. Biol. Med.*, vol. 170, 108047, 2024.
[39] L. Tong *et al*., "LiViT-Net: A U-Net-like, lightweight transformer network for retinal vessel segmentation," *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 213–224, 2024.
[40] W. Yan, Y. Jiang, Z. Zhang, Y. Yan, and B. Liu, "SRNet: Striped pyramid pooling and relational transformer for retinal vessel segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2023, p. 347.
[41] Z. Ma and X. Li, "An improved supervised and attention mechanism-based U-Net algorithm for retinal vessel segmentation," *Comput. Biol. Med.*, vol. 168, 107770, 2024.
[42] B. Chu, J. Zhao, W. Zheng, and Z. Xu, "(DA-U)²Net: Double attention U²Net for retinal vessel segmentation," *BMC Ophthalmol.*, vol. 25, no. 1, 86, 2025.
[43] Y. Jiang, J. Chen, W. Yan, Z. Zhang, H. Qiao, and M. Wang, "MAG-Net: Multi-fusion network with grouped attention for retinal vessel segmentation," *Math. Biosci. Eng.*, vol. 21, no. 2, pp. 1938–1958, 2024.
[44] M. Jian, W. Xu, C. Nie, S. Li, S. Yang, and X. Li, "DAU-Net: A novel U-Net with dual attention for retinal vessel segmentation," *Biomed. Phys. Eng. Express*, vol. 11, no. 2, 2025.
[45] S. Javed, T. M. Khan, A. Qayyum, A. Sowmya, and I. Razzak, "Region guided attention network for retinal vessel segmentation," arXiv Preprint, arXiv:2407.18970, 2024.
[46] F. D. Hernandez-Gutierrez *et al*., "Retinal vessel segmentation based on a lightweight U-Net and reverse attention," *Mathematics*, vol. 13, no. 13, 2203, 2025.
[47] Y. Liu, G. Zhang, Y. Yang, W. Gong, and X. Liu, "HCM: A hybrid CNN-Mamba architecture for semi-supervised retinal vessel segmentation," in *Proc. Artif. Intell. Educ. Syst.*, 2025, pp. 406–412.
[48] C. Guo, M. Szemenyei, Y. Hu, W. Wang, W. Zhou, and Y. Yi, "Channel attention residual U-Net for retinal vessel segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 1185–1189.
[49] Z. Zeng, J. Liu, X. Huang, K. Luo, X. Yuan, and Y. Zhu, "Efficient retinal vessel segmentation with 78K parameters," *J. Imaging*, vol. 11, no. 9, 306, 2025.
[50] J. Staal, M. D. Abramoff, and M. Niemeijer, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, 2004.
[51] M. M. Fraz *et al*., "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, 2012.
[52] H. Chen and K. Kim, "Multi-convolutional channel residual spatial attention U-net for industrial and medical image segmentation," *IEEE Access*, vol. 12, pp. 76089–76101, 2024.
[53] H. Qin and W. Wang, "An energy-efficient lightweight framework for superior retinal vessel segmentation," *Peer-to-Peer Netw. Appl.*, vol. 17, pp. 3133–3145, 2024.
[54] N. Yamanakkanavar, J. Y. Choi, and B. Lee, "MRI segmentation and classification of human brain using deep learning for diagnosis of Alzheimer's disease: A survey," *Sensors*, vol. 20, no. 11, 3243, 2020.
[55] G. Sun *et al*., "DA-TransUNet: Integrating spatial and channel dual attention with transformer U-net for medical image segmentation," *Front. Bioeng. Biotechnol.*, vol. 12, 1398237, 2024.
[56] L. Chen, C. Yuan, H. Xu, Y. He, and J. Jiang, "Robust denoising of structure noise through dual-diffusion Brownian bridge modeling and coupled sampling," *Electronics*, vol. 14, no. 21, 4243, 2025.