Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Depthwise Residual Transformer for Retinal Vascular Enhancement using Fundus Images

Srinivas Rangu[1], Uday Patil[2], Nagaraj Yamanakkanavar[3]

[1,3]Department of Electronics and Communication Engineering, Central University of Karnataka, India.
[2]Department of Electrical Engineering, Central University of Karnataka, India.

E-mail: [1]srigana.57@gmail.com, [2]udaypatil@cuk.ac.in, [3]nagrajpy@cuk.ac.in

Abstract

The high-quality visual information in fundus images enables highly accurate clinical judgment for the diagnosis of eye diseases. The image quality subsequently suffers due to factors such as refractive media excretions and inconsistent patient cooperation. To avoid these conditions, we propose a depthwise residual (DWR) transformer-based enhancement network for distorted fundus images. The proposed framework contains four blocks: the DWR block, the encoder block, the transformer block, and the decoder block. The DWR block is designed to improve model generalization over noisy inputs by applying channel-wise filtering. The encoder block has convolutional filters to capture fine-grained local features from the input images. The transformer block effectively captures global dependencies from the encoder blocks. The decoder block works for spatial resolution reconstruction while concatenating contextual features from the encoder with fine-grained information for accurate localization. The proposed model performance is evaluated through both subjective and objective analysis on open--source datasets such as DRIVE, STARE, CHASE_DB1, HRF, and MESSIDOR. Our proposed method achieves an excellent SSIM of 0.901 with the DRIVE dataset and PSNR of 31.596 with the MESSIDOR dataset, surpassing state–of–the–art methods across all datasets.

Keywords: Fundus Image Enhancement, Transformer, Depthwise Convolution, And Deep Learning.

1. Introduction

Retinal vessel enhancement historically relies on intensity- and geometry-based image processing to increase vessel contrast and suppress background artifacts. Gaussian smoothing and the Laplacian of Gaussian are used for noise suppression and edge enhancement in preprocessing and vessel detection [1]. Multi-scale top-hat transforms and morphological reconstruction address the extraction of bright or dim features and the removal of center reflexes to enhance vessel detail [2,3]. The multi-scale techniques including matched filters and coherence filters, are designed to characterize linear vessel profiles at various resolutions [4,5]. Contrast-limited adaptive histogram equalization (CLAHE), adaptive histogram equalization (AHE), histogram fitting, and homomorphic filters stabilize local contrast and illumination prior to segmentation [6]. Fuzzy c-means has been used to improve low-quality images by reassigning intensity clusters and then applying AHE to achieve vessel-preserving contrast gain [7]. Apart from the preprocessing techniques, better enhancement correlates with improved

DOI: https://doi.org/10.36548/jiip.2026.2.003
Received: 15.02.2026, received in revised form: 16.03.2026, accepted: 30.03.2026, published: 09.04.2026

image-quality metrics, which are used to explain downstream convolutional neural network (CNN) performance [8]. Chollet et al. [9] developed depthwise separable convolutions as an efficient alternative to standard convolutions. Gaussian noise modeling in digital images uses the additive model, popularized in early digital image processing, as a standard assumption from signal processing applied to images [10]. Hatamizadehet al. [11] proposed a transformer encoder to capture spatial and semantic features from 3D medical image patches. Chen et al. [12] introduced a model that enables simultaneous local and global feature extraction for retinal vascular segmentation. Lianget al. [13] developed a hybrid CNN-transformer network that preserves micro vessels and enhances contextual feature fusion by utilizing spatial reconstruction and feature interaction transformer modules. Lvet al. [14] proposed balanced key point detection and global context modeling for accurate segmentation, using a dual-path U-Net with transformer blocks and convolutional decoders. Jiang et al. [15] introduced a dual-branch network based on parallel transformers to segment retinal arteries. Mehmood et al. [16] developed a lightweight multitask CNN with integrated transformer modules for impact capture of retinal vessels and optic discs. Lin et al. [17] proposed an adaptive transformer attention block that enhances small vascular structures. Ronnebergeret al. [18] introduced a model capable of producing precise segmentations with limited training data. Soni et al. [19] developed a DenseNet architecture that promotes the repeated utilization of feature maps utilization and strengthens information. Ramos et al. [20] proposed the residual dense U-Net (RDU-Net), a network for image denoising inspired by the dense architecture. Luo et al. [21] introduced transformer architecture avoids the limitations based on large-scale datasets and computational resources. Chen et al. [22] developed a model based on the Brownian bridge diffusion process, called dual-diffusion Brownian bridge coupled sampling (DBBCS) for image denoising. Furthermore, these approaches achieved performance gains, substantially but increased in model complexity. To avoid these limitations, we introduce a depthwise residual transformer approach that significantly improves upon existing frameworks. Due to architectural benefits from the addition of a depthwise residual (DWR) block with a transformer block, the computational complexity has been reduced. Our main contributions are summarized as follows:

- We introduced a depthwise residual transformer framework with the depthwise convolution technique, which can remove irrelevant artifacts and enhance image contrast while preserving fine details.
- The encoder block uses a series of convolutions to characterize fine-grained spatial dependencies and local textures inherent in the input images.
- The transformer block at the bottleneck model provides global contextual information across the entire feature map.
- The decoder block is used to retain fine-grained details, often using skip connections to combine high-resolution information directly from the encoder block.
- Finally,the proposed model leverages the capabilities of residual and transformer networks for the enhancement task to improve vessel visibility and diagnostic clarity. The results on synthetically degraded fundus images show that the proposed model outperforms existing methods.

The structure of this paper is as follows: Section 2 covers a detailed explanation of the proposed methodology. Section 3 presents the experiments and results, and Section 4 concludes the study with final observations.

2. Proposed Methodology

The proposed model enhances the constrained environment of retinal fundus image. Data augmentation techniques are employed on the images to increase dataset diversity. The data augmentation methods consist of horizontal flip, vertical flip, optical distortions, and elastic transformations. Figure 1 portrays the proposed model for retinal image enhancement.

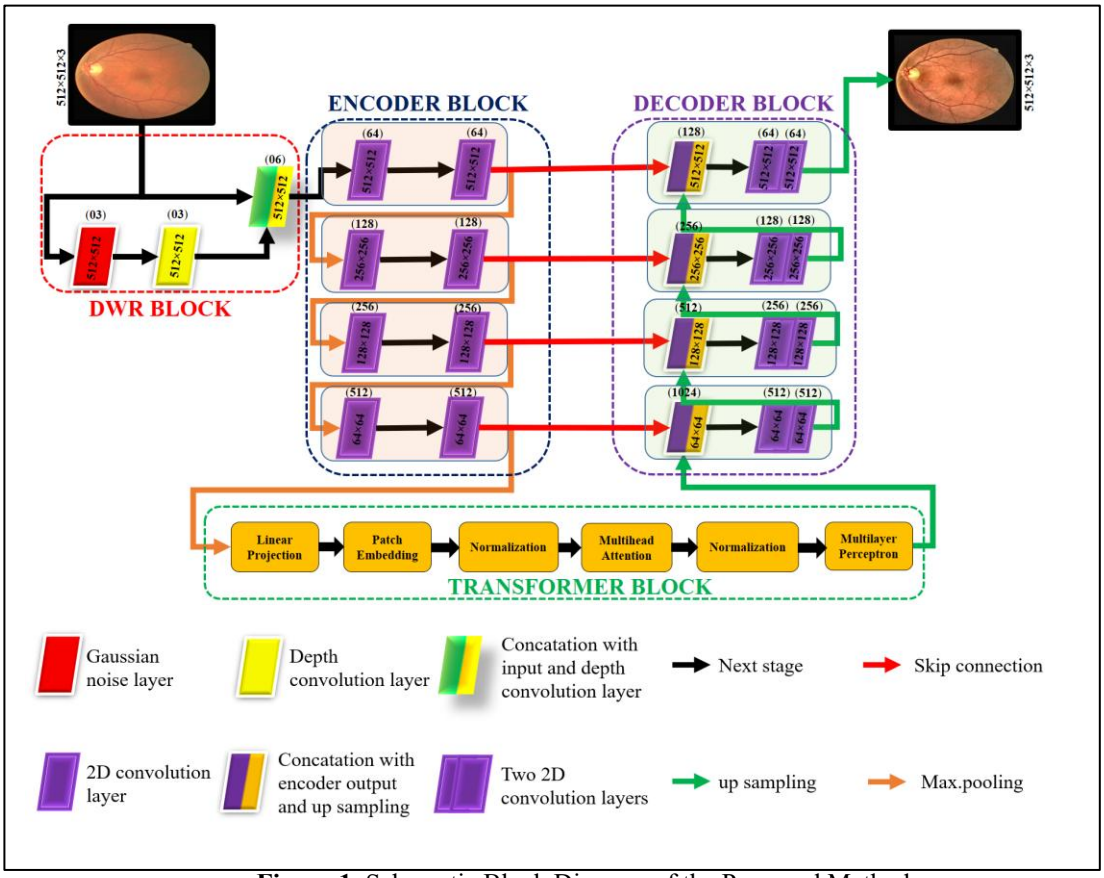

Figure 1. Schematic Block Diagram of the Proposed Method

The proposed model is explained in four stages: (i) DWR block, (ii) Encoder block, (iii) Transformer block, and (iv) Decoder block. Each stage's detailed explanation is discussed as follows.

2.1 DWR Block

The augmented image is represented as $A \in R^{H \times W}$, where A denotes the input, R is the set of real numbers, H is the input height, and W is the input width. The augmented images are fed to the two-branch DWR block. One branch conveys the raw input and the secondary branch transmits the noisy input. The secondary branch has a Gaussian noise layer and a depthwise convolution layer. The distortion model is defined as an additive stochastic process where the input image is corrupted by zero-mean Gaussian noise with a standard deviation of 0.01. This simulates the high-frequency electronic noise typically found in low-light fundus photography.

The Gaussian noise layer serves as a regularization mechanism by injecting stochastic perturbations into the input data. The input A is applied to Gaussian noise layer and the corresponding output is expressed as,

$$B = A + \epsilon, \qquad \{\ \epsilon \sim \mathrm{N}(0, \sigma^2)\} \tag{1}$$

where B represents the output of the Gaussian noise layer, σ is a noise standard deviation, ϵ denotes a random variable sampled from the Gaussian distribution. The output of the Gaussian noise layer is applied to the depthwise convolution layer. Depthwise convolution performs channel-wise filtering to retain channel-specific features. Depthwise convolution is performed on the B to compute an output feature map $D \in R^{H \times W \times C}$ and it is represented as,

$$D_{i,j,c} = \left(\sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} W_{m,n,c} \cdot B_{i+m,j+n,c}\right) x_3 + b \tag{2}$$

where $D_{i,j,c}$represents the output at spatial position(i,j) in channel c ,k_h is the height of the kernel,k_w is the width of the kernel, $W_{m,n,c}$ denotes the kernel weights,$B_{i+m,j+n,c}$represents input pixel at spatial offset(m,n), x_3 represents the 3 mode product, b is the bias term. The output of depthwise convolution layer is concatenated with the raw input image and it is denoted as,

$$E_{i,j,c} = (D_{i,j,c} \oplus A) \tag{3}$$

where $E_{i,j,c}$ represents the output of the DWR block, the symbol $\oplus$ represents the concatenation operation. The output of DWR block is propagated to encoder block.

2.2 Encoder Block

The next stage of the proposed framework is a four-layer encoder block. With four layers, the model is well balanced between efficiency and complexity. Each layer uses two successive convolution operations for the extraction of fine features. The first convolution operation extracts spatial features with a 3×3 kernel for a balanced trade-off between receptive field size and computational cost. Stacking multiple 3×3 layers increases the effective receptive field. The second convolution operation allows the encoder block to learn richer and more complex feature representations. A convolution operation is applied on the $E_{i,j,c}$ to get an output feature map $\boldsymbol{U}$ and it is expressed as,

$$U_O = K_O * E_{i,j,c} = \sum_{j=1}^{O'} K_O^j * E_j + b_O \tag{4}$$

where U_O represents the output of the convolution layer,$b_O$is the bias term,E_j stands for j^{th}element of the input vector,$*$ denotes the convolutional operator, K_O is the weights for the O^{th} output channel. After performing max. pooling, the output is represented as,

$$U_{Op} = maxpool(U_O) \tag{5}$$

where U_{Op} is output of the encoder block. The convolution layers are feed-forward networks that connect the output of the $(l\text{-}1)^{th}$ layer as input to the l^{th} layer. The U_{Op} is theinput for both the transformer block and the decoder block.

2.3 Transformer Block

The transformer block comprises several phases such as linear projection, patch embedding, normalization, multi-head attention, and a multilayer perceptron. The image patches are flattened and projected into a fixed-dimensional embedding space by the linear projection. Next, patch embedding organizes these projected vectors into a sequence. Positional encodings utilize fixed sinusoidal 2D encodings to retain spatial relationships that are lost during flattening. The embedded sequence then passes through normalization, which stabilizes training. The multihead attention module takes the normalized input. Multiple attention heads learn to focus on various parts of the image and capture diverse contextual information across patches. The output of the attention mechanism is again normalized, maintaining training stability, and preparing the data for the final stage. Lastly, the multilayer perceptron applies a feed-forward neural network to refine and transform the attended features. The multilayer perceptron facilitates the model's ability to learn difficult patterns and make accurate predictions. The transformer stages are denoted as,

$$X_a = softmax[\frac{(U_{Op} \cdot W_{q,k,v}) \cdot K^T}{\sqrt{d_k}}] \times V \tag{6}$$

where X_a is the output of the transformer block, V represents the actual content carried by each token, $W_{q,k,v}$ denotes weights to project input embeddings into queries (q), keys (k), and values (v), d_k denotes embedding dimensions, K^T is formally called as the transposed key matrix. The output of transformer block is given to the decoder block for image reconstruction.

2.4 Decoder Block

The decoder block is a key component of the decoder path in architectures such as U-Net, designed to retain high-resolution feature maps. The output of the transformer block $X_a$is upsampled by bilinear interpolation with a stride of 2 to increase spatial resolution. The output of the transformer block $X_a$upsampled and represented as,

$$\widetilde{X_a}=\text{upsampling}(X_a) \tag{7}$$

where $\widetilde{X_a}$represents upsampled output of the transformer block. The upsampled output is concatenated with the corresponding skip connection from the encoder. The decoder block fuses coarse information from deeper layers. The concatenation preserves all information by stacking feature maps along the channel dimension. The combined output is expressed as,

$$Z = \left(\widetilde{X_a} \oplus U_O\right) \tag{8}$$

where Z represents the input for the decoder block. The combined output features are processed by two successive 3×3 convolutional operations are represented as,

$$Z_O = K_O * Z = \sum_{j=1}^{O'} K_O^j * Z_j + b_O \tag{9}$$

where Z_O stands for the output of the decoder block. The upsampled output given to next layer is expressed as,

$$\tilde{Z} = U_{Op}\{\text{upsampling}(Z_O)\} \tag{10}$$

where $\tilde{Z}$ is the final feature representation of the decoder block. Finally, a convolution operation is used at the last layerto combine the decoder contextual features into meaningful semantic categories. This improves class discrimination by emphasizing relevant patterns while suppressing noise, and it is denoted as,

$$Y = \sigma_S(conv(\tilde{Z})) \tag{11}$$

where Y is the output of proposed model and σ_S represents the sigmoid activation. The model learns fine-grained refinements for tasks such as denoising, color correction, and low-light boosting by the sigmoid activation function.

3. Experimental Results and Analysis

3.1 Materials

The experiments are conducted on five public datasets to evaluate the proposed approach compared with the state-of-the-art methods. Table 1 provides an overview of retinal fundus images in various datasets with different numbers and resolutions. As part of the validation process for retinal vessel enhancement algorithms, these datasets serve as crucial benchmarks.

Table 1. Summary of the Retinal Fundus Images Datasets

S.No.	Dataset	Image Type	# of Images	Resolution	Description
1	DRIVE [23]	Retinal fundus	40	768 × 584	7 images with mild diabetic retinopathy (DR).
2	STARE [24]	RGB fundus	400	~700 × 605	Used for glaucoma and DR research.
3	CHASE_DB1 [25]	RGB fundus	28	1280 × 960	Known for vascular and pathological variation.
4	HRF [26]	RGB fundus	45	High resolution	Covers healthy and diseased retinas.
5	MESSIDOR [27]	RGB fundus	1748	512 × 512	Image-wise diagnostic labels for risk of macular edema.

- **DRIVE:** These images were collected from DR patients in the Netherlands in the Digital Retinal Images for Vessel Extraction (DRIVE) dataset [23]. The DRIVE dataset contains 40 images, seven of which showed mild symptoms of DR for retinal vessel extraction.
- **STARE:** The Structured Analysis of the Retina (STARE) [24] was developed by Michael Goldbaum at the University of California, San Diego. In the STARE collection, 400 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, age-related macular degeneration, and DR research.
- **CHASE_DB1:** The Child Heart and Health Study in England (CHASE) dataset [25] covers various primary schools from London, Birmingham, and Leicester. The CHASE_DB1 consists of 28 images illustrating a variety of vascular and pathological features. It is often used in deep learning studies to segment retinal vessels.

- **HRF:** A High-Resolution Fundus dataset (HRF) [26] was released by the GE-CZ research group in 2013. The high-resolution fundus camera captured 45 retinal fundus images for the HRF collection. The images depict both healthy and sick retinas, illustrating diseases such as macular degeneration and DR.
- **MESSIDOR:** The Methods to Evaluate Segmentation and Indexing Techniques in Retinal Ophthalmology (MESSIDOR) database was established in 2008 by the MESSIDOR project in France [27]. It was funded by the French Ministry of Research and Defense to provide a standard for computer-assisted diagnosis. Messidor-2 was created to expand the original set by including 1,748 images from several subjects. The dataset contains both normal and pathological cases and is used for research in vascular segmentation, disease analysis, and DR detection.

3.2 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 using the TensorFlow library with mixed precision enabled. The input images across datasets were resized to 512 × 512 pixels for processing. To increase the diversity of training data, various augmentation techniques were employed. The augmentation techniques included vertical flip, horizontal flip, grid distortion, elastic deformation, and optical distortion. Each dataset was split into training (80%), validation (10%), and testing (10%) subsets. Every model was trained for 100 epochs, with a learning rate of 0.001, a batch size of 2, a transformer block patch size of 1, and the Adam optimizer utilized for the entire process. An Early Stopping callback was employed to avoid overfitting. The robustness and reproducibility of results were achieved by setting the random seed to 42.

3.3 Evaluation Indicators

To confirm both the effectiveness and generalizability of the proposed multi-block architecture, additional intensive experiments are conducted on contextual datasets. Each dataset represents a unique type of imaging modality and diagnostic goal. The purpose of the experiments is to determine enhancement quality and robustness to domain shifts. Table 2 shows the performance metrics, such as peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and mean squared error (MSE), that are used to compare the proposed method with the state-of-the-art approach. Typically, the degradation of an image is calculated in correlation to a reference image [28]. In accordance with standard practice in the image restoration literature the arithmetic mean of PSNR and SSIM is used across the test sets. SSIM and PSNR provide a direct comparison with existing state-of-the-art methods, which typically do not provide confidence intervals [29].

Table 2. Evaluation Metrics and Their Mathematical Formulas

Metric	Description	Equation
Mean square error (MSE)	A reference-based metric, when a lower value indicates better performance.	$\frac{1}{mn}\sum_{n=0}^{m}\sum_{m=1}^{n}[\tilde{h}(n,m) - h(n,m)]^2$
Peak signal-to-noise ratio (PSNR)	The disparity between the maximum possible signal power and the power of the distorting noise.	$10\log_{10}[(peakvalue)^2 /(MSE)]$
Structural similarity index method (SSIM)	Image degradation is measured as the paradigm shift in structural particulars.	$\{[l(x,y)]^{\propto} \times [c(x,y)]^{\beta} \times [s(x,y)]^{\gamma}\}$

3.4 Results and Discussion

The output images consist of raw input images, and the corresponding predicted images are illustrated in Figure 2. This image represents a comparative analysis of retinal blood vessel enhancement using various methods. This visual comparison is crucial for evaluating how well each approach captures the complexities of vessel patterns. The proposed model shows clearer vessel continuity, superior branch detection, and reduced noise compared to other approaches. As the proposed model builds on transformer principles, it likely probable benefits from enhancing context awareness and long-range feature modeling beyond local pixel neighborhoods. The performance evaluation of the proposed method is carried out on synthetically distorted fundus images. The low-quality images are taken from published datasets. The experimental results of each dataset are explained as follows. Table 3 illustrates that the model delivers solid performance using the DRIVE dataset. 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 dataset. The proposed model outperforms in terms of evaluation metrics but parameter count, it lags behind the models [19-21].

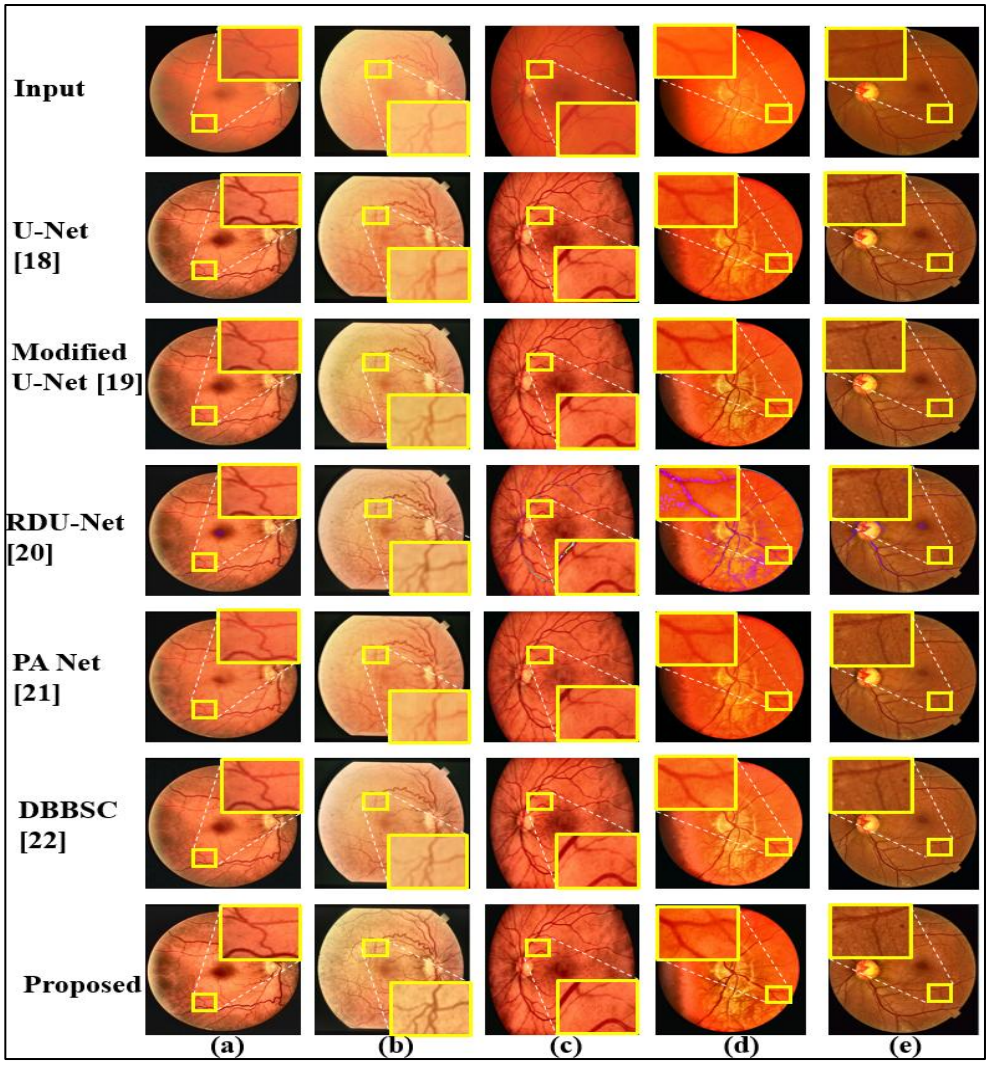

Figure 2. Visual Results for Various Methods and Proposed Model with the (a) DRIVE, (b) STARE, (c) HRF, (d) CHASE_DB1, and (e) MESSIDOR Datasets, Original and Enhanced Images, Respectively

The model achieved the highest SSIM of 0.901 and PSNR of 30.208 compared with all existing models. The DBBCS model achieved an SSIM of 0.885, but its learnable parameters are far from those of the proposed model. The Modified U-Net achieved a PSNR score of 29.738, placing it in second place. RDU-Net and PA Net excelled in processing time due to their specialized connectivity and feature extraction strategies.

Table 3. Comparative Results for Different Enhancement Methods Using the DRIVE Dataset

S.No.	Year	Model	SSIM	PSNR	Time	Parameters
1	2015	U-Net [18]	0.848	28.238	28 minutes	31,055,427
2	2020	Modified U-Net [19]	0.867	29.738	9 minutes	7,922,627
3	2021	RDU-Net [20]	0.881	28.794	7 minutes	5,304,739
4	2024	PA Net [21]	0.793	26.672	5 minutes	717,763
5	2025	DBBCS [22]	0.885	28.819	29 minutes	31,404,430
6	2026	Proposed	0.901	30.208	12 minutes	12,930,766

Table 4 shows that on the STARE dataset the proposed model attained the highest PSNR of 30.263 and an SSIM of 0.884. However, DBBCS also obtained a good SSIM of 0.841 through the hybrid attention mechanisms. The processing time for the proposed model is less than that of the PA Net model. However, in terms of SSIM and PSNR, the PA Net is far nehind the proposed model. Considering all factors, the proposed model demonstrates enhanced performance due to the combination of multi-scale fusion and attention.

Table 4. Comparative Results for Diverse Enhancement Methods Using the STARE Dataset

S.No.	Year	Model	SSIM	PSNR	Time	Parameters
1	2015	U-Net [18]	0.845	28.102	22 minutes	31,055,427
2	2020	Modified U-Net [19]	0.812	27.715	7 minutes	7,922,627
3	2021	RDU-Net [20]	0.814	22.157	9 minutes	5,304,739
4	2024	PA Net [21]	0.764	25.515	5 minutes	717,763
5	2025	DBBCS [22]	0.841	27.391	11 minutes	31,404,430
6	2026	Proposed	0.884	30.263	6 minutes	12,930,766

Table 5 demonstrates the proposed model's performance using the CHASE_DB1 dataset, achieving the highest SSIM of 0.864, primarily due to its dense connections that improve contextual learning and maintain vessel continuity with fewer errors. For SSIM, the U-Net model also acheived 0.856, as its hybrid architecture is particularly effective for detecting fine and low-contrast vessels. Overall, the proposed model dominated several core metrics with its advanced design.

Table 5. Comparative Results for Various Enhancement Methods Using the CHASE_DB1 Dataset

S.No.	Year	Model	SSIM	PSNR	Time	Parameters
1	2015	U-Net [18]	0.856	27.469	21 minutes	31,055,427
2	2020	Modified U-Net [19]	0.804	28.317	9 minutes	7,922,627
3	2021	RDU-Net [20]	0.795	20.061	7 minutes	5,304,739
4	2024	PA Net [21]	0.704	25.129	5 minutes	717,763
5	2025	DBBCS [22]	0.765	25.777	36 minutes	31,404,430
6	2026	Proposed	0.864	30.012	5 minutes	12,930,766

Table 6 shows that the HRF dataset performance highlights the proposed model design's strengths. The proposed model leads to an SSIM of 0.822 and a PSNR of 30.391. The intense metrics across all models demonstrate that the HRF dataset enables consistently accurate retinal vessel enhancement. However, the HRF dataset gives low evaluation metric values compared to all other datasets. The HRF dataset has high-resolution images, which increases computational complexity. The HRF dataset contains abnormal retinal structures such as microaneurysms and exudates, which have similar intensities to blood vessels. Because of this,

the deep learning models perform poorly in generalization compared with other datasets. The proposed model offers peripheral yet meaningful gains in obtaining fine structures and minimizing errors.

Table 6. Comparative Results for Various Enhancement Methods using the HRF Dataset

S.No.	Year	Model	SSIM	PSNR	Time	Parameters
1	2015	U-Net [18]	0.769	28.745	27 minutes	31,055,427
2	2020	Modified U-Net [19]	0.797	29.572	14 minutes	7,922,627
3	2021	RDU-Net [20]	0.703	27.961	6 minutes	5,304,739
4	2024	PA Net [21]	0.737	26.805	10 minutes	717,763
5	2025	DBBCS [22]	0.793	29.664	37 minutes	31,404,430
6	2026	Proposed	0.822	30.391	8 minutes	12,930,766

Table 7 illustrates that the proposed model accomplished the highest PSNR of 31.596. This indicates proposed model potent capability to detect vessels accurately across assorted image sources. Modified U-Net performed vigorously with a high PSNR of 31.242 assisted by position attention modules that improve spatial focus. Other models, such as PA Net, DBBCS, and U-Net, maintained consistent but slightly lower performance on the SSIM metric. Overall, the proposed model shows evident advantages in dealing with the complexity of the MESSIDOR dataset with its sophisticated attention-driven architecture.

Table 7. Comparative Results for Distinct Enhancement Methods Using the MESSIDOR Dataset

S.No.	Year	Model	SSIM	PSNR	Time	Parameters
1	2015	U-Net [18]	0.786	31.184	20 minutes	31,055,427
2	2020	Modified U-Net [19]	0.768	31.242	13 minutes	7,922,627
3	2021	RDU-Net [20]	0.811	27.164	10 minutes	5,304,739
4	2024	PA Net [21]	0.754	29.981	6 minutes	717,763
5	2025	DBBCS [22]	0.768	29.725	37 minutes	31,404,430
6	2026	Proposed	0.894	31.596	9 minutes	12,930,766

3.5 Ablation Study

To validate the effectiveness of the different configurations of the models, we performed an ablation study. Table 8 shows the results of the ablation study. In this table, we present the effects of the encoder block, the DWR block, and the transformer block. We utilized the decoder block in all configurations. We formed various configurations based on these blocks. The configurations are evaluated on the DRIVE, STARE, CHASE_DB1, HRF, and MESSIDOR datasets. The standard image quality metrics, such as SSIM and PSNR, are used as evaluation metrics. The proposed configuration achieves the best results among all other configurations. The proposed model required a maximum processing time of 12 minutes on the DRIVE dataset, while the minimum time observed was 5 minutes on the STARE and CHASE_DB1 datasets. The encoder block without the DWR block and transformer block gives the nominal results, with the number of parameters increasing significantly. The model trained with an encoder block and a DWR block has the highest parameter count among all other configurations.

Table 8. Ablation Study of the Retinal Vessel Enhancement on the Various Datasets

Dataset	Encoder Block	DWR Block	Transformer Block	SSIM	PSNR	Time	Parameters
DRIVE	✓	×	×	0.848	28.238	28 minutes	3,10,55,427
	✓	✓	×	0.881	30.799	29 minutes	3,14,04,430
	✓	×	✓	0.891	30.155	26 minutes	1,29,28,963
	×	✓	✓	0.733	26.241	8 minutes	2,54,222

	✓	✓	✓	0.901	31.208	12 minutes	1,29,30,766
STARE	✓	×	×	0.845	28.102	22 minutes	3,10,55,427
	✓	✓	×	0.812	26.684	12 minutes	3,14,04,430
	✓	×	✓	0.782	27.128	8 minutes	1,29,28,963
	×	✓	✓	0.716	23.608	5 minutes	2,54,222
	✓	✓	✓	0.864	30.102	5 minutes	1,29,30,766
HRF	✓	×	×	0.769	28.745	27 minutes	3,10,55,427
	✓	✓	×	0.789	29.304	36 minutes	3,14,04,430
	✓	×	✓	0.792	27.153	26 minutes	1,29,28,963
	×	✓	✓	0.736	24.712	6 minutes	2,54,222
	✓	✓	✓	0.822	30.391	8 minutes	1,29,30,766
CHASE_ DB1	✓	×	×	0.856	27.469	21 minutes	3,10,55,427
	✓	✓	×	0.783	25.873	12 minutes	3,14,04,430
	✓	×	✓	0.791	26.821	9 minutes	1,29,28,963
	×	✓	✓	0.667	24.688	6 minutes	2,54,222
	✓	✓	✓	0.864	30.012	5 minutes	1,29,30,766
MESSID OR	✓	×	×	0.786	31.184	20 minutes	3,10,55,427
	✓	✓	×	0.784	30.382	37 minutes	3,14,04,430
	✓	×	✓	0.721	29.247	19 minutes	1,29,28,963
	×	✓	✓	0.781	27.127	6 minutes	2,54,222
	✓	✓	✓	0.894	31.596	9 minutes	1,29,30,766

The model with the DWR block and transformer block has the advantage in terms of processing time and parameter count. From these observations, the proposed model offers the best balance among overlap quality, parameter count, and processing time. The proposed model achieves the best performance by enhancing images with minimal error. To complement PSNR and SSIM, we evaluated perceptual similarity using learned perceptual image patch similarity (LPIPS) across five benchmark datasets. Table 9 describes how the LPIPS values represent a strong balance between perceptual similarity and enhancement. The LPIPS values ranged from 0.26 (MESSIDOR) to 0.36 (HRF), with standard deviations consistently below 0.10. These results indicate moderate perceptual similarity between enhanced and original images, with stable performance across datasets. While LPIPS values are higher than those expected for pixel-perfect reconstruction, they are consistent with enhancement tasks where perceptual detail is introduced. Importantly, the enhancements also improved vessel segmentation accuracy, underscoring their clinical relevance.

Table 9. Summary of LPIPS Scores Across Datasets for the Proposed Model

Dataset	Mean ± Std.	Min.	Max.
DRIVE	0.309±0.027	0.256	0.364
CHASE_DB1	0.335±0.092	0.271	0.368
HRF	0.361±0.012	0.347	0.395
STARE	0.322±0.072	0.271	0.374
MESSIDOR	0.263±0.024	0.246	0.281

4. Conclusion

In this paper, we introduce a depthwise residual transformer framework that removes irrelevant artifacts while maintaining fine details and enhancing image contrast. In addition, the paper presents a unified, systematically driven proposed framework that includes a transformer architecture used as a bottleneck. The proposed model is efficient for intricate image structures by combining the advantages of CNNs and transformers. The proposed model demonstrated superior performance in retinal image analysis compared to existing methods. The model achieved an SSIM of 0.901 on the DRIVE dataset and a PSNR of 31.596 on the MESSIDOR

dataset, illustrating the potential to improve medical diagnosis through precise retinal vessel enhancement. Future work should incorporate clinically relevant evaluations such as vessel segmentation performance, and statistical significance testing to provide a more comprehensive assessment. In addition, we intend to incorporate well regularization methods and structures, minimizing computational costs without reducing performance.

Acknowledgement

NA.

References

[1] Alqahtani, Fasial Majed, Somaya Adwan, Mohd Yazed Ahmad, Salmah Binti Karman, Rasha Maged, and Lamya'A. Maged. "Impact of Retinal Enhancement Techniques on Blood Vessel Segmentation." In 2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE, 2024, 517-521.

[2] Bandara, A. M. R. R., and P. W. G. R. M. P. B. Giragama. "A Retinal Image Enhancement Technique for Blood Vessel Segmentation Algorithm." In 2017 IEEE international conference on industrial and information systems (ICIIS), IEEE, 2017, 1-5.

[3] Pal, Mahua Nandy, Minakshi Banerjee, and Shuvankar Roy. "Deep Learning Supported Evaluation of Retinal Vessel Enhancement Techniques." In Advances in Data Science and Computing Technology, Apple Academic Press, 2022, 3-18.

[4] Liao, Miao, Shao-Wei Zheng, and Yu-Qian Zhao. "A Novel Method for Retinal Vascular Image Enhancement." Journal of Optoelectronics Laser 23, no. 11 (2012): 2237-2242.

[5] Abdushkour, Hesham, Toufique A. Soomro, Ahmed Ali, Fayyaz Ali Jandan, Herbert Jelinek, Farida Memon, Faisal Althobiani, Saleh Mohammed Ghonaim, and Muhammad Irfan. "Enhancing Fine Retinal Vessel Segmentation: Morphological Reconstruction and Double Thresholds Filtering Strategy." PLoS One 18, no. 7 (2023): e0288792.

[6] Shabbir, Safia, Anam Tariq, and M. Usman Akram. "A Comparison and Evaluation of Computerized Methods for Blood Vessel Enhancement and Segmentation in Retinal Images." International Journal of Future Computer and Communication 2, no. 6 (2013): 600.

[7] Hien, Nguyen Mong. "Retinal Vessels Segmentation Based on Enhancing Multi-scale Line Detection." In International Conference on the Development of Biomedical Engineering in Vietnam, Cham: Springer Nature Switzerland, 2022, 519-528.

[8] Soomro, Toufique Ahmed, Ahmed J. Afifi, Ahmed Ali Shah, Shafiullah Soomro, Gulsher Ali Baloch, Lihong Zheng, Ming Yin, and Junbin Gao. "Impact of Image Enhancement Technique on CNN Model for Retinal Blood Vessels Segmentation." IEEE Access 7 (2019): 158183-158197.

[9] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 1251-1258.

[10] Gonzalez, Rafael C, “Digital Image Processing”, Pearson Education India, 2009.

[11] Hatamizadeh, Ali, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. "Unetr: Transformers for 3d Medical Image Segmentation." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, 574-584.

[12] Chen, Tianhua, and Mike Lucock. "The Mental Health of University Students During the COVID-19 Pandemic: An Online Survey in the UK." PloS one 17, no. 1 (2022): e0262562.

[13] Liang, Liming, Baohe Lu, Jian Wu, Yulin Li, and Xiaoqi Sheng. "Sfit-Net: Spatial Reconstruction Feature Interaction Transformer Retinal Vessel Segmentation Algorithm." Biomedical Signal Processing and Control 106 (2025): 107688.

[14] Lv, Nianzu, Li Xu, Yuling Chen, Wei Sun, Jiya Tian, and Shuping Zhang. "Tcddu-Net: Combining Transformer and Convolutional Dual-Path Decoding U-Net for Retinal Vessel Segmentation." Scientific Reports 14, no. 1 (2024): 25978.

[15] Jiang, Ligang, Jing Hu, Zhuoran Wang, Guohui Yuan, Chongjun Huang, Zhiming Xiong, Meizhen Zhang, Weihua Yang, and Yuhua Tong. "TransDualSegNet: Transformer Dual Segment Network for Retinal Vasculature Segmentation in OCT." Quantitative Imaging in Medicine and Surgery 15, no. 10 (2025): 9338.

[16] Mehmood, Mehwish, Majed Alsharari, Shahzaib Iqbal, Ivor Spence, and Muhammad Fahim. "Retinalitenet: A Lightweight Transformer Based CNN for Retinal Feature Segmentation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, 2454-2463.

[17] Lin, Ji, Xingru Huang, Huiyu Zhou, Yaqi Wang, and Qianni Zhang. "Stimulus-Guided Adaptive Transformer Network for Retinal Blood Vessel Segmentation in Fundus Images." Medical Image Analysis 89 (2023): 102929.

[18] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In International Conference on Medical image computing and computer-assisted intervention, Cham: Springer international publishing, 2015, 234-241.

[19] Soni, Ashish, Radhakanta Koner, and Vasanta Govind Kumar Villuri. "M-Unet: Modified U-Net Segmentation Framework with Satellite Imagery." In Proceedings of the Global AI congress 2019, Singapore: Springer Singapore, 2020, 47-59.

[20] Gurrola-Ramos, Javier, Oscar Dalmau, and Teresa E. Alarcón. "A Residual Dense U-Net Neural Network for Image Denoising." IEEE Access 9 (2021): 31742-31754.

[21] Luo, Xuebing, Lingxi Peng, Ziyan Ke, Jinhui Lin, and Zhiwen Yu. "Pa-Net: A Hybrid Architecture for Retinal Vessel Segmentation." Pattern Recognition 161 (2025): 111254.

[22] Chen, Long, Changan Yuan, Huafu Xu, Ye He, and Jianhui Jiang. "Robust Denoising of Structure Noise Through Dual-Diffusion Brownian Bridge Modeling and Coupled Sampling." Electronics 14, no. 21 (2025): 4243.

[23] Staal, Joes, Michael D. Abràmoff, Meindert Niemeijer, Max A. Viergever, and Bram Van Ginneken. "Ridge-Based Vessel Segmentation in Color Images of the Retina." IEEE transactions on medical imaging 23, no. 4 (2004): 501-509.

[24] Ibtehaz, Nabil, and M. Sohel Rahman. "MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation." Neural networks 121 (2020): 74-87.

[25] Fraz, Muhammad Moazam, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R. Rudnicka, Christopher G. Owen, and Sarah A. Barman. "An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation." IEEE transactions on biomedical engineering 59, no. 9 (2012): 2538-2548.

[26] Chen, Haoyu, and Kyungbaek Kim. "Multi-Convolutional Channel Residual Spatial Attention U-Net for Industrial and Medical Image Segmentation." IEEE Access 12 (2024): 76089-76101.

[27] Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain et al. "Feedback on a Publicly Distributed Image Database: the Messidor Database." Image Analysis & Stereology (2014): 231-234.

[28] Liang, Haoyi, and Daniel S. Weller. "Comparison-Based Image Quality Assessment for Selecting Image Restoration Parameters." IEEE transactions on image processing 25, no. 11 (2016): 5118-5130.

[29] Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising." IEEE transactions on image processing 26, no. 7 (2017): 3142-3155.