JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 3 (2025) 387-402
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Robust OCTA Vessel Segmentation for Early Detection of Neurodegenerative Disorders Using Multi-Scale CNN and Transformer Networks

Madhu C K[1,2]*, Raghunadhan K R[1], Uma B[2], Tejonidhi M R[1,2], Vinod A M[1,2]

[1] *Computer Science and Engineering,*
Nitte (Deemed to be University), NMAM Institute of Technology, Nitte, 574110, Karnataka, INDIA
[2] *Computer Science and Engineering,*
Malnad College of Engineering, Hassan, 573202, Karnataka, INDIA

*Corresponding Author: ckm@mcehassan.ac.in
DOI: https://doi.org/10.30880/jscdm. 2025.06.03.024

Article Info

Received: 3 July 2025
Accepted: 19 November 2025
Available online: 30 December 2025

Keywords

Retinal vessel segmentation, Optical Coherence Tomography Angiography (OCTA), Shallow Feature Extraction Module (SFEM), Convolutional Block Attention Modules (CBAM), Deep learning, U-Net.

Abstract

Early detection of vascular neurodegenerative diseases relies on precise segmentation of retinal vessels in Optical Coherence Tomography Angiography (OCTA) images. However, complex vascular structures, low-contrast micro-vessels, and overlapping vessel types pose significant challenges to conventional segmentation methods. This study proposes a deep learning architecture named CGOctaNet that combines convolutional neural networks (CNNs) with a transformer-based global context modeling framework for robust OCTA vessel segmentation. The model employs a U-Net-like encoder–decoder structure integrated with three modules: (i) a Shallow Feature Extraction Module (SFEM) to preserve vessel boundaries; (ii) a multi-scale convolution encoder for local geometric feature learning; and (iii) a Transformer bottleneck that captures long-range dependencies and inter-vessel relationships for enhanced structural consistency. The Transformer bottleneck enables awareness of global context by learning the spatial relationships between remote areas of the vessels, thereby complementing the small receptive field of CNNs and enhancing segmentation continuity in the presence of noise. This hybrid design achieves better generalization and fine-grained segmentation accuracy than CNN-only models in the past. Experiments on OCTA-500 and ROSE benchmark datasets demonstrate that CGOctaNet outperforms state-of-the-art methods, achieving Dice scores of 93.50%, 90.25%, and 89.50% on OCTA-3M, OCTA-6M, and ROSE datasets, respectively. The improvements are attributed to effective integration of local and global contextual cues, adaptive attention refinement, and balanced optimization through hybrid Dice-Cross Entropy loss.

1. Introduction

Dementia is a rapidly growing global health challenge, affecting over 55 million people worldwide and adding around 10 million new cases each year [1, 2]. The condition progressively deteriorates cognitive abilities, leading to loss of independence and imposing high economic costs, which are projected to exceed USD 2.8 trillion by 2030 [2, 3]. Although several diagnostic modalities exist, such as cerebrospinal fluid (CSF) analysis, magnetic resonance

This is an open access article under the CC BY-NC-SA 4.0 license.

imaging (MRI), blood-based biomarkers, and genetic testing [4], their invasiveness, high cost, and time requirements limit their use for early and large-scale screening [5]. This has created a demand for non-invasive, rapid, and scalable biomarkers capable of detecting preclinical changes in neurodegenerative disorders like Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI).

Because of its shared embryological origin with the brain, the retina is increasingly regarded as a window to the brain [6], [7]. Several studies have demonstrated the accumulation of amyloid-beta and tau proteins in the retinal layers of AD patients, mirroring the neuropathological features of the brain [8]. These findings establish retinal imaging as a promising biomarker for early detection of AD.

Recent advances in Optical Coherence Tomography (OCT) and Optical Coherence Tomography Angiography (OCTA) have enabled detailed visualization of the retinal microvasculature, providing insight into vascular alterations associated with neurodegenerative disorders. OCT offers high-resolution, depth-resolved imaging of retinal layers [9], while OCTA provides volumetric angiographic information without contrast dyes. It allows precise mapping of the Superficial Vascular Plexus (SVP) and Deep Vascular Plexus (DVP), facilitating quantitative analysis of microvascular impairment linked to AD [10, 11]. Accurate segmentation of OCTA images is thus essential for isolating anatomical and pathological regions, including the Foveal Avascular Zone (FAZ), which serve as biomarkers for early-stage cognitive decline [12, 13].

Traditional rule-based methods, such as thresholding, edge detection, and active contour models, have shown limited success, performing well only on high-contrast images while failing under noise, motion artefacts, and inter-patient variability [14]. These methods also struggle to generalize across imaging devices and require extensive manual parameter tuning. To overcome these challenges, deep learning–based approaches using Convolutional Neural Networks (CNNs) have been widely adopted for OCTA segmentation tasks [15]. Architectures like U-Net, U-Net++, and Attention U-Net have achieved promising results by learning hierarchical representations directly from raw data. However, their reliance on local receptive fields restricts their ability to capture long-range dependencies and maintain topological vessel continuity - key factors in accurate microvascular analysis.

In order to improve contextual learning at a global scale, newly Transformer-based networks, as well as CNN-Transformer hybrids, have been considered. These models are better at focusing on clinically relevant areas, but they can be computationally intensive and lack scalability to small datasets typical of medical imaging. Moreover, the problem of an imbalanced dataset, scanner variability, and minor variability in vascular differences among the patients continues to impede the applicability of models.

Thus, the present research proposes a new hybrid deep learning model, referred to as CGOctaNet, which is a multi-scale convolutional encoder with a Transformer bottleneck that successfully encodes local textures and global dependencies. The suggested approach enhances the integrity of vessel boundaries, the thin capillary connectivity, and the uniformity of performance for segmentation in low-contrast and noisy settings. The solution offers a powerful and generalizable methodology for early detection of neurodegenerative disorders using OCTA imaging.

2. Literature Review

Optical Coherence Tomography Angiography (OCTA) has emerged as a powerful non-invasive imaging modality for visualizing the retinal microvasculature and diagnosing vascular and neurodegenerative diseases such as diabetic retinopathy, glaucoma, and Alzheimer's Disease (AD). However, accurate vessel segmentation in OCTA images remains challenging due to inherent noise, varying image quality, and the complex multi-scale nature of retinal vasculature.

Recent developments in deep learning have shifted segmentation research from traditional rule-based techniques to data-driven architectures capable of learning hierarchical and contextual representations. Wang et al. [16] proposed SAM-OCTA, a fine-tuned version of the Segment Anything Model (SAM) using Low-Rank Adaptation (LoRA), achieving high segmentation accuracy and highlighting the adaptability of foundation models for biomedical imaging. Sedighin et al. [17] employed low-rank tensor ring decomposition to improve FAZ segmentation and morphological consistency. Zou et al. [18] developed OCTAMamba, a U-shaped neural architecture integrating multi-scale dilated convolutions and feature recalibration modules for efficient vessel modeling.

To mitigate the scarcity of labeled OCTA datasets, Wittmann et al. [19] introduced a simulation-based segmentation framework generating realistic 3D synthetic data with projection artefacts and signal loss, enabling annotation-free training. Shen et al. [20] presented HAIC-Net, a semi-supervised method combining self-supervised classification with dual consistency training, which preserved vascular topology while reducing annotation cost. Similarly, Zhang et al. [21] designed LA-Net, applying layer and boundary attention for enhanced vessel delineation in volumetric OCTA data. Jiang et al. [22] improved the SegNet backbone using deformable convolutions and Convolutional Block Attention Modules (CBAM) to ensure vessel continuity and smoother

boundaries. Li et al. [23] proposed a direction-guided network that leveraged vessel orientation information to maintain topological consistency in fine vascular structures.

Even with these developments, there are a number of serious concerns. The majority of current models are scanner-specific and cannot easily be applied to new scanners and pathological settings. The transformer-based architectures, being efficient, are usually resource-intensive and require large amounts of annotated data, which constrains their scalability. By contrast, CNN-based networks retrieve primarily local spatial data, which results in discontinuous vessel prediction in high-capillary-density or low-contrast areas. In addition, manual annotation remains manual, limiting data diversity and affecting reproducibility. Such limitations underscore the need for a lightweight, context-aware hybrid model that can effectively combine local detail preservation and global dependency learning.

To fill these gaps, this paper introduces CGOctaNet, a hybrid deep learning model that integrates both multi-scale convolutional encoding of local spatial detail and a Transformer bottleneck for reasoning over global information. This model allows the local texture of vessels and long-range inter-vessel relations to be represented in the design, thereby achieving topological consistency and better segmentation accuracy. Also, the decoder is equipped with Convolutional Block Attention Modules (CBAM), which enable it to refine features and suppress background noise.

The proposed CGOctaNet is more balanced in performance than the latest models like SAM-OCTA, OCTAMamba, and HAIC-Net in providing segmentation accuracy and computational efficiency. It has a multi-scale, attention-guided design that provides strong generalization across datasets and imaging conditions, making CGOctaNet a scientifically supported and effective system for reliable OCTA vessel segmentation and early neurodegenerative disease detection.

3. Proposed Model

This section presents a detailed discussion of the proposed CGOctaNet, a deep learning framework for high-precision segmentation of OCTA images. The CGOctaNet model focuses on capturing both local and global contextual features by integrating convolution encoders. Moreover, it incorporates the Convolutional Block Attention Module (CBAM) and the Transformer Block in the bottleneck to enhance these details. UNet inspires the overall architecture, and it follows a symmetric encoder-decoder topology with skip connections, enabling effective fusion of hierarchical features.

3.1 Architectural Overview

The proposed CGOctaNet model adopts a Unet-like framework with three main components, including (a) an encoder module for capturing the multi-scale local feature, (b) a Transformer bottleneck for long-range dependency modelling, and (c) a decoder module that is equipped with the CBAM attention modules to reconstruct the precise segmentation mask. The encoder module consists of a double convolution block and a MaxPooling layer. The transformer bottleneck module helps capture long-range dependencies and encode global contextual relations across different regions of the feature map. Similarly, the decoder module mirrors the encoder with upsampling layers. At each decoding stage, features from the corresponding encoder level are concatenated via skip connections, enabling the recovery of spatial details.

In OCTA images, blood vessels may span large spatial regions, and capturing these connections helps understand vessel continuity and shape at a broader level. Similarly, the contextual information refers to the semantic information between parts of the image. The contextual information helps to differentiate between similar-looking structures based on their appearance. Figure 1 illustrates the complete CGOctaNet architecture, depicting the flow of information between the encoder, Transformer bottleneck, and decoder with CBAM refinement.

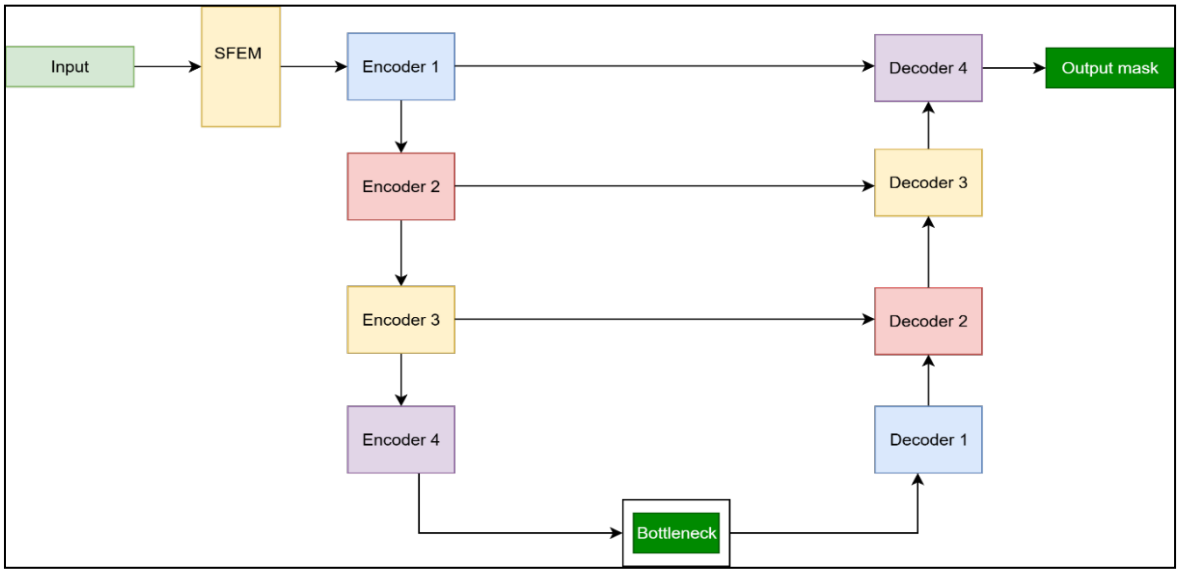

Fig. 1 *Complete architecture of the CGOctaNet model*

The Transformer bottleneck is an important element that helps capture non-local interactions between distant vessel regions, which CNNs alone cannot effectively model due to their limited receptive fields. This design enables CGOctaNet to maintain global vascular continuity and reduce segmentation discontinuities prevalent in microvascular regions. According to this process, shallow feature extraction is the initial stage of the proposed model, in which different features are extracted. The encoder module uses 4 encoders to encode the data. These encoded features are processed through the bottleneck module, which helps to obtain the long-range dependencies. Furthermore, the decoder module performs reconstruction by processing the data through 4 sophisticated decoder modules. The outcome of the final decoder block produces the segmentation mask.

3.2 Problem Statement

Let us consider that the retinal OCTA image domain is denoted as $\Omega \in \mathbb{R}^2$ and each image is represented by an intensity function as given in eq. (1)

$$I : \Omega \to \mathbb{R} \tag{1}$$

The goal of segmentation is to learn the mapping $\mathcal{F}_\theta$ which assigns each pixel $x \in \Omega$ a corresponding label $y \in \{0,1,2, \dots, C-1\}$ where is the number of vessel classes (e.g., artery, vein, capillary, background, etc.). Below given eq. (2) describes this expression as:

$$\mathcal{F}_\theta : I(x) \to \hat{y}(x) \tag{2}$$

The segmentation task is modeled as a pixel-wise classification problem, as expressed in eq. (3):

$$\hat{y}(x) = \arg\max_c \hat{Y}_c(x) \,, \forall_{x \in \Omega} \tag{3}$$

During the training process, the main aim of this model is to minimize the loss function $\mathcal{L}$ which is a combination of cross-entropy and dice loss.

3.3 CGOctaNet Architecture

A. Shallow feature extraction

To enhance low-level vessel boundary preservation, we introduce a Shallow Feature Extraction Module (SFEM) at the input stage of the proposed CGOctaNet architecture. This module consists of a sequence of 2D convolutional layers with small receptive fields (3x3 kernels) followed by Batch normalization and ReLU activation. This block helps capture edge features, vessel contours, and intensity gradients, which are crucial for enhancing microvascular detail in retinal OCTA images. Below given eq. (4) presents the expression of the SFEM block as :

$$F_0 = \phi\big(BN(W_2 * \phi(BN(W_1 * I + b_1)) + b_2)\big) \tag{4}$$

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 387-402

where I is the input OCTA image, $W_1, W_2 \in \mathbb{R}^{C\times1\times3\times3}$ are the convolution weight tensors, b_1, b_2 are the learnable biases, $\phi(.)$ is the ReLU non-linearity, and BN is the batch normalization. Figure 2 shows the feature extraction block.

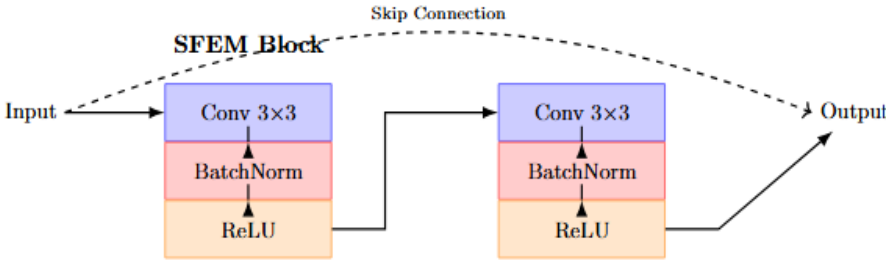

Fig. 2 *Feature extraction block*

The obtained output feature map $F_0 \in \mathbb{R}^{C\times H\times W}$ is then forwarded to the encoder stage. The SFEM ensures the enhancement of thin vessel structure and location contrast variation before proceeding to high-level abstracts in the encoder phase. Table 1 shows the shallow feature-extraction block for OCTA images.

Table 1 *Shallow feature extraction block for OCTA images*

Layer Type	Parameters	Output Shape	Purpose
Input	–	1 × H ×W	Raw grayscale OCTA image
Conv2D-1	3×3, stride=1, pad=1, out_channels=C	C × H × W	Capture vessel edges and gradients
BatchNorm2D-1	–	C × H × W	Normalize activation
ReLU-1	–	C × H × W	Introduce non-linearity
Conv2D-2	3×3, stride=1, pad=1, out_channels=C	C × H × W	Extract local texture and microvascular details
BatchNorm2D-2	–	C × H ×W	Improve feature stability
ReLU-2	–	C × H ×W	Final shallow feature representation

B. Encoder Module

This section presents a detailed description of the proposed encoder module. As discussed before, each encoder block consists of two sequential convolutional layers with a kernel size of 3×3, followed by batch normalization and ReLU activation. A max-pooling layer with a stride of 2 is applied after each block to downsample the feature maps. These blocks effectively capture edge-level, texture-level, and shape-level information from OCTA images. The output of the shallow feature block is presented as eq. (5):

$$F_o \in \mathbb{R}^{C_0\times H\times W} \tag{5}$$

where C_0 is the number of channels, and H×W is the spatial resolution of the input OCTA image.

The encoder module consists of several encoder stages, where each encoder stage E_i comprises a Double Convolution Block, which is a stack of two consecutive 2D convolutional layers with ReLU activation and Batch Normalization. It can be represented as eq. (6):

$$F_i = ConvBlock(F_{i-1}) = \phi(BN(W_{i2} * (\phi(BN(W_{i1} * F_{i-1} + b_{i1})) + b_{i2})) \tag{6}$$

where $W_{i1}, W_{i2} \in \mathbb{R}^{C_i\times C_{i-1}3\times3}$ represents the convolution weights, b_{i1}, b_{i2} represents the corresponding biases. this block helps to enhance the spatial locality while enabling the deeper feature extraction. Thus, it becomes important to distinguish capillaries from background noise.

Further, the module incorporates Max Pooling after the double convolution block, where a 2×2 Max Pooling layer down-samples the spatial dimensions by a factor of 2. This operation is expressed as eq. (7):

$$F_i^{\downarrow} = MaxPool(F_i), where\ F_i^{\downarrow} \in \mathbb{R}^{C_i\times\frac{H}{2}\times\frac{W}{2}} \tag{7}$$

This progressively reduces the resolution while increasing the receptive field, which is essential for identifying large vessel structures and establishing spatial continuity in wide-field-of-view OCTA scans. The encoder path consists of N such stages (e.g., N=4). The output feature map of this stage is presented as eq. (8) where different features are combined at each level as follows:

$$\{F_1, F_2, \dots, F_N\} \tag{8}$$

Where F_1 corresponds to shallow features, and F_N is the deepest and most abstract representation fed into the Transformer bottleneck. Each encoder output F_i is also skip-connected to the corresponding decoder layer D_i to facilitate feature fusion and spatial detail preservation during reconstruction. The following Figure 3 depicts the overall architecture of the encoder module.

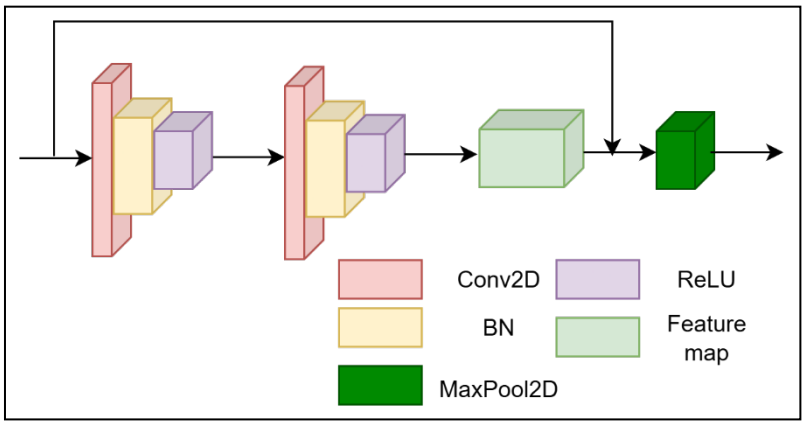

Fig. 3 *Encoder module (single Block)*

C. Transformer Bottleneck Module

This section presents a detailed discussion of the proposed transformer-based bottleneck module. This transformer bottleneck design is based on the vision transformer paradigm (ViT), in which multi-head attention helps the model focus on both global and local parts of vessels simultaneously. This context aggregation through attention enhances the differentiation of overlapping vessel structures, thin capillaries, and low-intensity capillary segments of microvessels, which traditional CNNs fail to detect. The proposed bottleneck is positioned between the proposed encoder and decoder modules. Unlike traditional bottlenecks, which rely solely on convolutional modules, the proposed module focuses on capturing long-range dependencies and contextual information, which are crucial for identifying fine vessel branches and capillaries amid overlapping vascular artifacts and background noise.

The proposed bottleneck module consists of flattening with positional encoding, transformer encoder layers, and feed-forward ML blocks. Let the $F_N \in \mathbb{R}^{C\times H\times W}$ represents the output feature map obtained from the final encoder stage. This feature map is flattened to form a sequence of tokens, making it suitable for transformer processing. The flattening process is expressed in eq. (9):

$$X_0 = Flatten(F_N) \in \mathbb{R}^{(H.W)\times C} \tag{9}$$

In order to retain the spatial structure that is lost during the flattening process, we incorporated a 2D sinusoidal positional encoding $P \in \mathbb{R}^{(H.W)\times C}$ are added to input tokens, which is expressed in eq. (10).

$$Z_0 = X_0 + P \tag{10}$$

This process ensures spatial correspondence that is important for reconstructing vessel continuity across wide-field angiographic frames. Further, we stack L transformer encoder blocks, which comprise Multi-Head Self-Attention (MHSA) and Feedforward Network (FFN) layers. This is further augmented by layer normalization and residual connections. Eq. (11) and (12) can be referred to obtain the multi-head self-attention of l^{th} block as:

$$Q = Z_{l-1}W_Q, K = Z_{l-1}W_K, V = Z_{l-1}W_V \tag{11}$$

$$MHSA(Z_{l-1}) = Concat(head_1, \dots, head_h)W_0 \tag{12}$$

Where each head computes $head_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$ Later, Add and Normalization are performed on it as follows:

$$Z_l^{'} = LayerNorm\left(Z_{l-1} + MHSA(Z_{l-1})\right) \tag{13}$$

The feed-forward MLP block can be represented with the help of eq. (14) and eq. (15):

$$FFN(Z_l') = \sigma(Z_l' W_1 + b_1) W_2 + b_2 \tag{14}$$

$$Z_l = LayerNorm(Z_l^{'} + FFN(Z_l^{'}) \tag{15}$$

Where $W_1 \in \mathbb{R}^{C \times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff} \times C}$ and σ represents the activation function. This module helps spatial tokens to consider all other tokens, and thus it models the non-local interaction across distant relevant regions in the retinal scan. After the final Transformer block, we reshape the output tokens back into the original 2D feature format as given in eq. (16):

$$F_T = Reshape(Z_L) \in \mathbb{R}^{C \times H \times W} \tag{16}$$

This transformation helps to integrate both local and global context. The local contextual information is obtained with the help of convolutions, and global information is obtained by a self-attention mechanism. The final outcome of this block is obtained by fusing the deepest encoder attribute through residual merging as eq (17):

$$F_T^{'} = \lambda . F_t + (1 - \lambda) . F_n \tag{17}$$

Where $\lambda \in [0,1]$.

The following Figure 4 depicts the overall architecture of this block

Fig. 4 *Bottleneck module*

D. Decoder module

This section presents a detailed overview of the proposed decoder module for OCTA image segmentation. It is used to reconstruct the high-resolution OCTA segmentation mask from attributes derived from compressed transformer-enhanced features. During this process of reconstruction, it performs progressive upsampling and integrates encoder skip features via concatenation. Each stage uses CBAM to improve the feature representation, followed by a convolutional refinement block. The main aim of this block is to recover the spatial structure and preserve boundary and edge information to improve retinal vessel segmentation. CBAM inclusion enhances the network's attentiveness to areas of interest and avoids background noise, particularly towards images with unbalanced brightness or motion artefacts. This process improves the structural preservation and the accuracy in segmentation of thin vessel boundaries.

Let $F_4, F_3, F_2, F_1 \in \mathbb{R}^{C_i \times H \times W_i}$ represents the hierarchical encoder outputs obtained from the deepest to shallowest levels and $T \in \mathbb{R}^{C_4 \times H_4 \times W_4}$ represents the output processed by the transformer bottleneck. The decoder comprises four upsampling blocks, where each block performs the following operations:

1. **Upsampling:** this operation is performed using bilinear interpolation followed by a 1×1 convolution to increase spatial resolution and reduce channel dimension.
2. **Skip Connection Fusion**: Concatenating the up-sampled feature map with the corresponding encoder feature map F_i.
3. **CBAM Attention**: Applying CBAM [1] to refine the concatenated feature map via channel and spatial attention.
4. **Double Convolution**: Using two consecutive 3×3 convolutions with ReLU and batch normalization to enhance semantic richness.

Each decoder block operation can be expressed using the eq. (18-21):

$$\tilde{T}_i = Upsample(T_{i+1}) \in \mathbb{R}^{C_i \times H_i \times W_i} \tag{18}$$

$$C_i = Concat(\tilde{T}_i, F_i) \in \mathbb{R}^{2C_i \times H_i \times W_i} \tag{19}$$

$$A_i = CBAM(C_i) \tag{20}$$

$$U_i = Conv_{3\times3}\left(BN\left(ReLU\left(Conv_{3\times3}(A_i)\right)\right)\right) \tag{21}$$

Where CBAM comprises channel and spatial attention, Conv, ReLU, and BN refer to convolution, activation, and batch normalization, respectively. The channel attention and spatial attention mechanism are expressed by using eq. (22) and eq. (23):

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{22}$$

$$M_s(F) = \sigma(Conv_{7\times7}([AvgPool(F); MaxPool(F)])) \tag{23}$$

With the help of these two relations, the refined feature can be expressed as eq. (24):

$$F' = M_c(F).F, F'' = Ms(F').F' \tag{24}$$

The output of the last decoder module $U_1 \in \mathbb{R}^{C_1 \times H \times W}$ is passed through a 1×1 convolution to reduce the channel dimension to the number of classes, followed by a sigmoid activation function. This can be expressed as eq. (25):

$$Y = \sigma(Conv_{1\times1}(U_1)) \in \mathbb{R}^{K \times H \times W} \tag{25}$$

This provides the final outcome as a segmentation map $\hat{Y}$, with each pixel assigned to a probability for each vessel class, as shown in Figure 5.

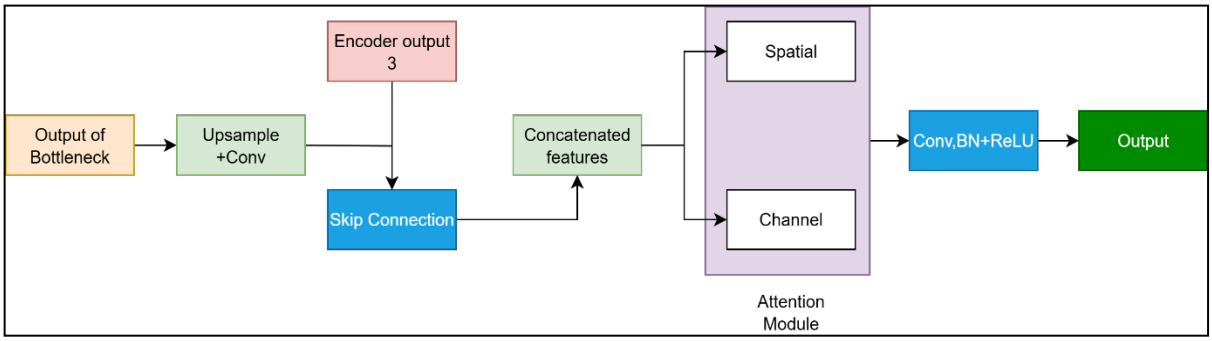

Fig. 5 *Decoder block module (single unit)*

Despite CGOctaNet's better performance compared to the state-of-the-art, several challenges remain. The model can be mildly sensitive to extreme artefacts of illumination or severe motion noise when acquiring OCTA. Moreover, the Transformer bottleneck itself has moderate memory needs of the graphics card, which may restrict its use on low-resource devices. Its efficiency and adaptability for large-scale clinical implementations can be further optimized in the future through lightweight transformer variants and domain adaptation methods.

4. Results and Discussion

This section presents detailed results for the CGOctaNet approach and compares its performance with that of state-of-the-art segmentation methods. First, we describe the experimental setup, then present the dataset details and performance measurement parameters. In the next subsection, the outcome of the proposed CGOctaNet model is presented, including qualitative and quantitative analyses. The obtained performance is then compared with the existing methods.

4.1 Experimental Setup

The proposed experiments are carried out on a workstation with Windows 10 operating system with the following specifications: Intel(R) Core (TM) i7 processor, 16 GB RAM, and NVIDIA GeForce RTX 2030 GPU with 8 GB of memory. This workstation was used to develop and evaluate the proposed OCTA vessel segmentation model. The complete programming is done in Python 3.8, using frameworks such as PyTorch and CUDA acceleration to train and run the model efficiently.

The model structure is based on a U-Net-like encoder-decoder, followed by a Transformer bottleneck module and Convolutional Block Attention Modules (CBAM) in the decoder branch. In the encoder block, features at four levels are extracted in a multi-scale fashion and processed through long-range dependencies in the transformer module to refine and progressively up-sample the features in the decoder block.

To conduct experiments, the OCTA-500 and ROSE datasets have been employed, with images and their corresponding ground truth provided. Input images and masks were standardised in size to 512x512 pixels before processing through the proposed architecture. To improve generalization, image normalization and augmentation methods, including horizontal/vertical flips, brightness adjustments, and rotation, were employed using the Albumentations library. Adam optimizer with a starting learning rate of 10^{-4} was used to train the model, and a combined Binary Cross-Entropy (BCE) and Dice Loss objective was employed to mitigate class imbalance and improve segmentation accuracy. The training was implemented with a batch size of 4 and a learning rate decayed using a cosine annealing scheduler. The model was trained for up to 100 epochs, and early stopping based on the validation Dice coefficient was used to prevent overfitting.

4.2 Dataset Details

The OCTA-500 dataset provides a comprehensive, publicly accessible collection of 500 volumetric retinal optical coherence tomography (OCTA) images, encompassing both healthy and diseased subsets. By including manually annotated vessel segmentation masks alongside en-face projections, this dataset establishes a foundation for validating vessel extraction methods and investigating microvascular characteristics. Consequently, it enables the comparison of vascular structures, precise boundary definition, and the derivation of quantitative biomarkers associated with retinal pathologies. In addition, the ROSE dataset (Retinal OCTA Segmentation) serves as a widely adopted benchmark, featuring a diverse range of OCTA images captured at various fields of view, including 2×2 mm, 3×3 mm, and 6×6 mm. This dataset is accompanied by expert-annotated binary vessel masks and Foveal Avascular Zone (FAZ) identifiers, facilitating the development and evaluation of segmentation algorithms. Notably, the ROSE datasets are particularly valuable for assessing the generalizability of segmentation performance across different image dimensions and disease states, thereby supporting the creation of more robust and adaptable analytical tools.

4.3 Performance Measurement Parameters

The CGOctaNet architecture performs segmentation on retinal OCTA images. Thus, the performance of this work is measured in terms of the Dice score and the IoU score.

Dice score is a widely used parameter to assess the performance of segmentation models. It measures the similarity between the predicted segmentation mask and the ground truth mask. The Dice score can be expressed as:

$$Dice = 2 * \frac{|X \cap Y|}{|X| + |Y|} \tag{26}$$

IoU, or Intersection over Union, also known as the Jaccard index or Jaccard similarity coefficient, is a widely used metric in computer vision for evaluating the performance of tasks such as object detection and image segmentation. It quantifies the degree of overlap between a predicted region and the ground truth region. It can be computed as:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} \tag{27}$$

Similarly, we incorporate specificity, precision, and F1-score parameters to demonstrate the pixel-level accuracy of the proposed model. With the help of eq. (28-30), these parameters can be computed as:

$$Spe = \frac{TN}{TN + FP} \tag{28}$$

$$Precision = \frac{TP}{TP + FP} \tag{29}$$

$$F1Score = \frac{2 * Precision.Recall}{Precision + Recall} \tag{30}$$

The specificity shows the model's accuracy to identify the background pixel correctly, precision refers to how many predicted vessel pixels are actually correct, and the F1 score provides a harmonic mean of precision and recall

4.4 Outcome of Proposed Approach

This segment presents the outcome of the CGOctaNet approach and its comparative analysis with state-of-the-art models. According to the CGOctaNet architecture, the first block is a shallow feature-extraction block that uses a combination of convolutional layers to extract the initial feature set. The obtained feature map using this block is depicted in Figure 6. For simplicity of feature representation, we have shown 8 different feature maps from the SFEM block.

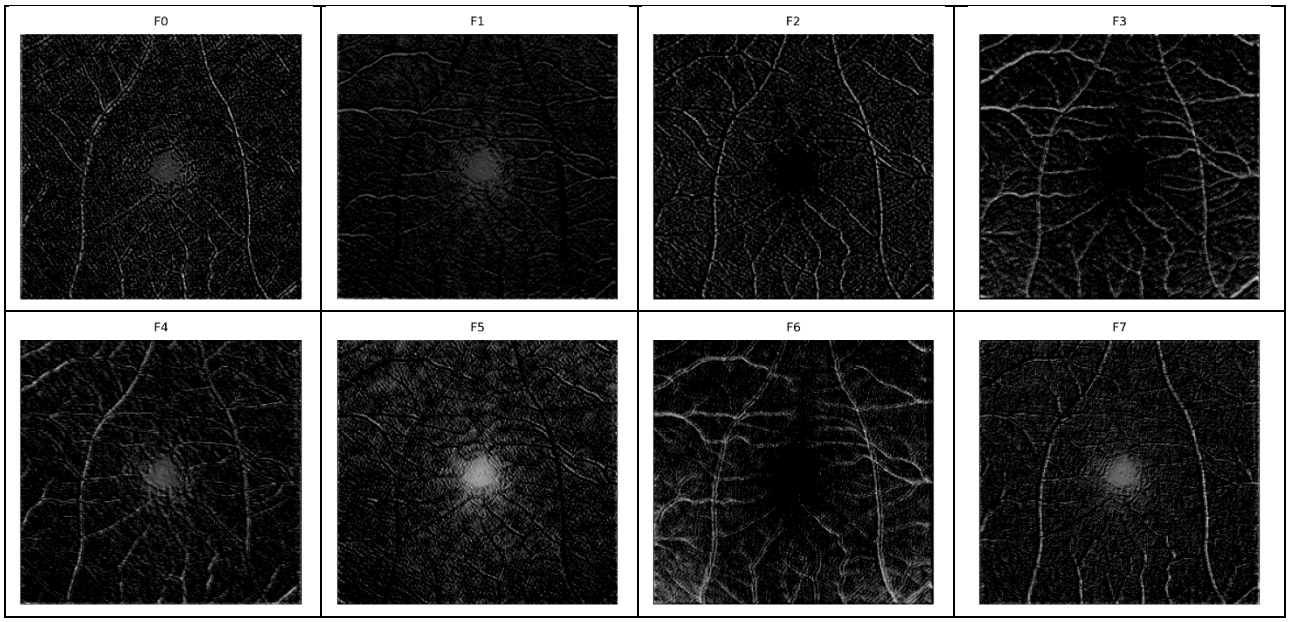

Fig. 6 *Outcome of shallow feature extraction block*

In the next stage, the encoder module is applied, using a different layer combination to extract deep features from the input image. The following figure depicts the output of each stage of the encoder. These encoder blocks progressively reduce the resolution and extract deep features, as shown in Figure 7.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 387-402

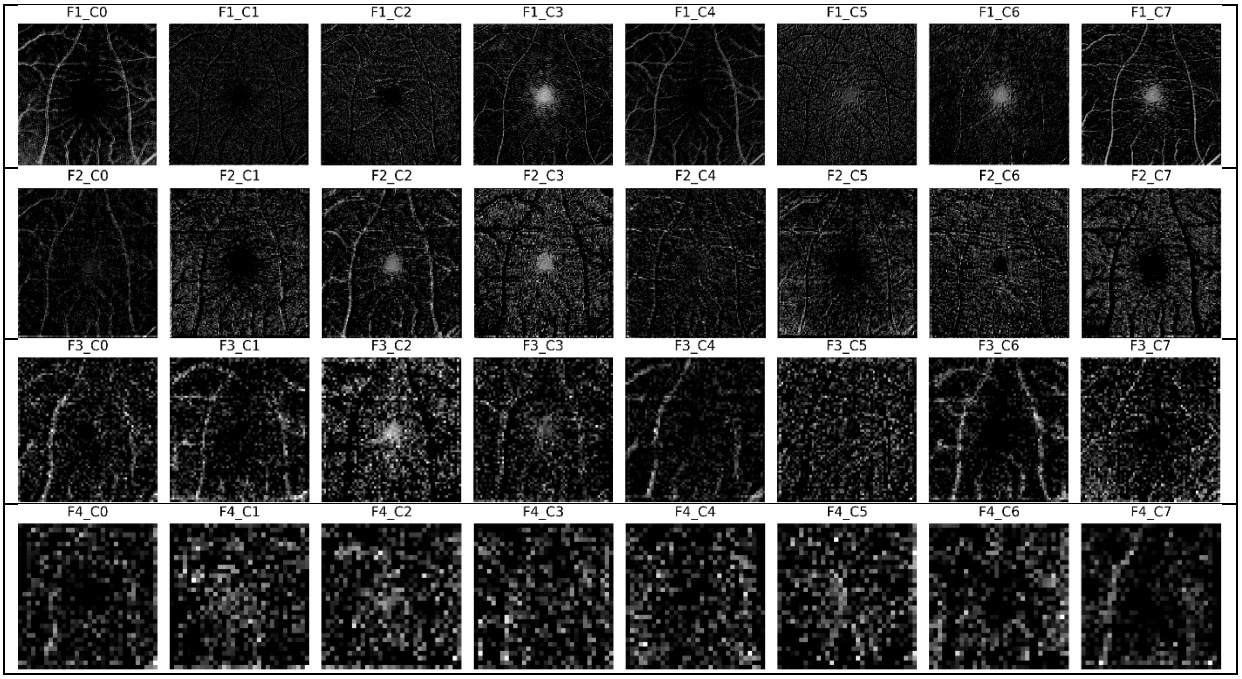

Fig. 7 *Outcome of encoder stages*

The final features of the encoder block before processing through the bottleneck are depicted in the figure below, which is later used by the bottleneck modules. Figure 8 shows the final feature of the encoder block before the bottleneck.

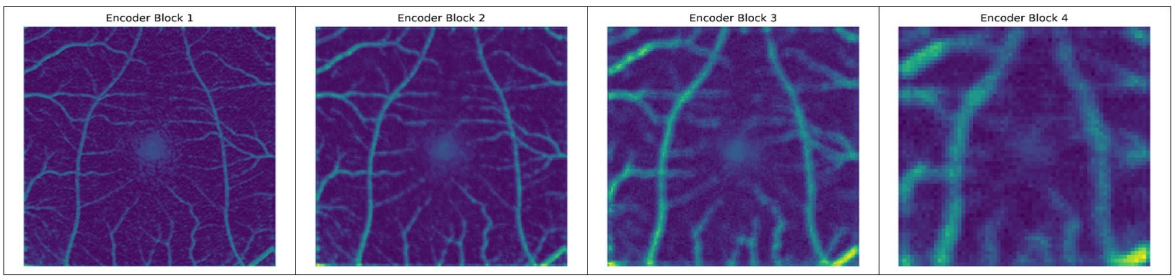

Fig. 8 *Final feature of the encoder block before the bottleneck*

Furthermore, a Transformer bottleneck layer is employed, consisting of an MHSA and an FFN layer. This stage helps to enhance the global feature representation. Moreover, the proposed module focuses on capturing long-range dependencies and contextual information, which helps improve overall segmentation performance. Figure 9 below depicts the outcome of the proposed transformer bottleneck layer.

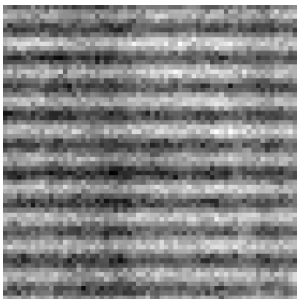

Fig. 9 *Bottleneck output*

Finally, the decoder block focuses on reconstructing the segmentation mask by performing progressive Up-sampling and concatenating features from the encoder block. Figure 10 below depicts the decoder's output for the vessel and capillary image segmentation tasks.

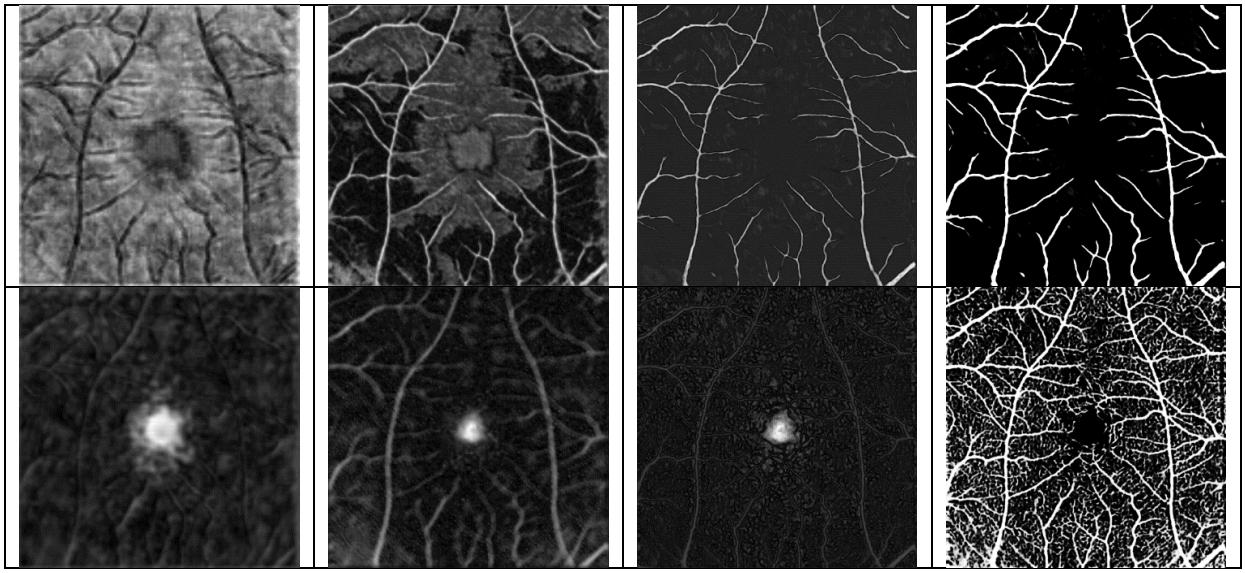

Fig. 10 *Decoder stage output*

Based on the outcome of CGOctaNet, the final segmentation of different classes is obtained and depicted in Figure 11 below, where the first row of the figure demonstrates the segmented mask of the large vessel class, the second row depicts the FAZ segmentation, and the third row shows the output of the capillary segmentation class.

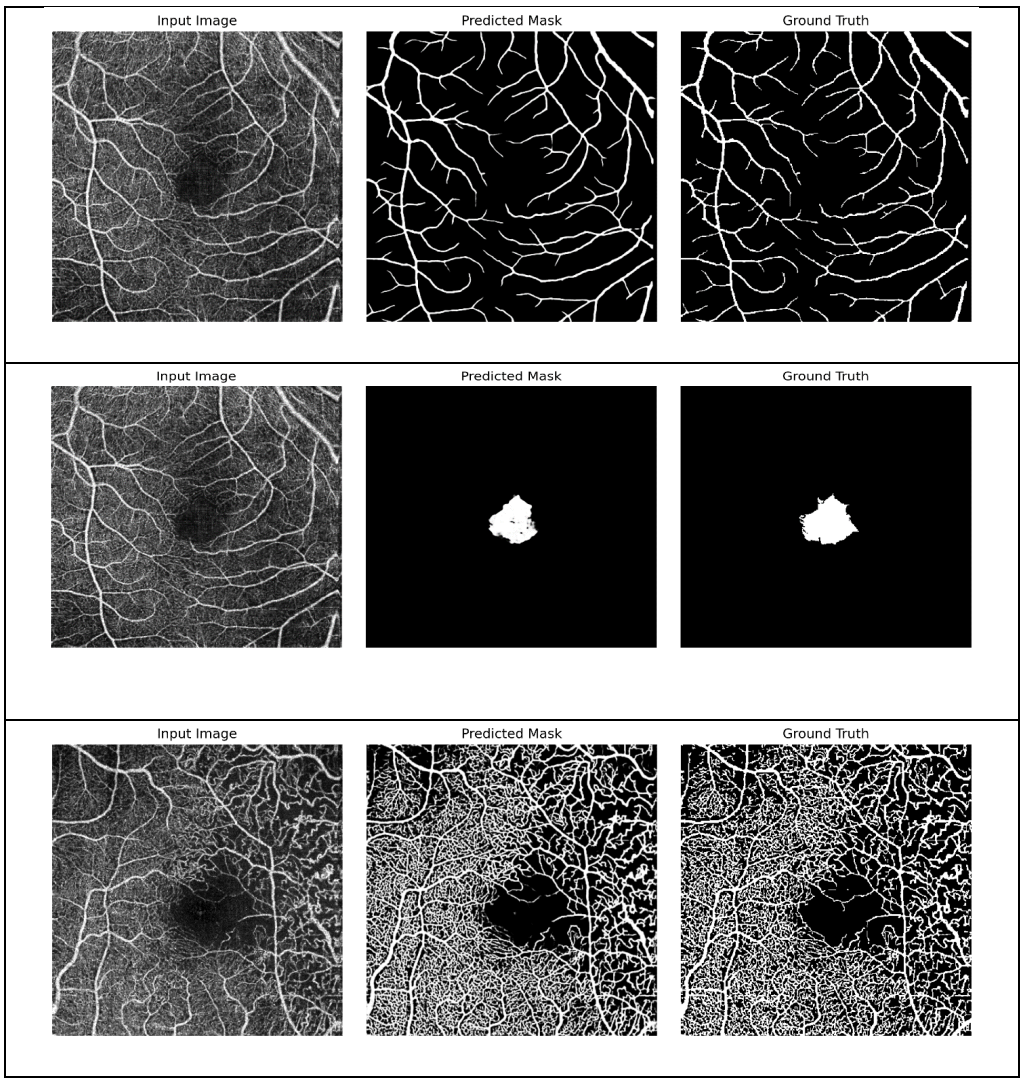

Fig. 11 *Qualitative segmentation performance of the CGOctaNet model*

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 387-402

The qualitative analysis reported the promising performance of the CGOctaNet approach for different classes. Further, we present a quantitative performance analysis in which the parameters above measure the proposed model's performance, and we compare it with existing methods. Tables 2 and 3 below present a comparative study of retinal vessel segmentation.

Table 2 *Comparison of segmentation performance on OCTA 3M, OCTA 6M, and ROSE datasets*

Method	Dice	IoU	Sen	Pr	Spe	F1-score
OCTA 3M						
Unet	79.36	65.86	79.81	78.92	97.63	79.36
R2UNet	66.32	46.31	69.33	64.10	92.30	66.32
UNet++	82.83	70.86	82.31	83.36	98.17	82.83
SwinUNet	73.66	58.40	72.01	75.39	97.39	73.66
H2Former	80.13	66.90	82.32	NA	NA	80.13
MISSFormer	80.63	67.62	80.72	NA	NA	80.63
UMamba	75.76	61.07	77.56	NA	NA	75.76
VM-UNet	71.98	56.34	69.67	NA	NA	71.98
AC-Mamba	80.44	67.36	79.13	NA	NA	80.44
H-vmunet	67.15	50.66	69.10	NA	NA	67.15
OCTAMamba [18]	84.50	73.23	84.00	NA	NA	84.50
SAM OCTA [16]	91.99	0.852	NA	NA	NA	91.99
SAM OCTA-2	92.07	0.842	NA	NA	NA	92.07
Proposed	93.50	0.8670	85.50	99.10	99.25	93.50
OCTA 6M						
Method	Dice	IoU	Sen	Pr	Spe	F1-score
Unet	77.32	63.11	78.88	75.82	97.20	77.32
R2UNet	51.78	35.16	44.82	58.70	89.40	51.78
UNet++	79.60	66.21	80.06	79.15	97.66	79.60
SwinUNet	72.39	56.84	72.10	72.68	96.99	72.39
H2Former	74.19	59.04	78.07	NA	NA	74.19
MISSFormer	78.03	64.07	78.97	NA	NA	78.03
UMamba	70.24	54.22	73.41	NA	NA	70.24
VM-UNet	71.55	55.81	70.54	NA	NA	71.55
AC-Mamba	78.42	64.61	77.19	NA	NA	78.42
H-vmunet	64.18	47.36	66.61	NA	NA	64.18
OCTAMamba [18]	82.31	70.03	82.75	NA	NA	82.31
SAM OCTA [16]	88.69	0.79	NA	NA	NA	88.69
SAM OCTA-2	89.23	0.804	NA	NA	NA	89.23
Proposed	90.25	84.50	88.50	92.07	99.15	90.25
ROSE						
Method	Dice	IoU	Sen	Pr	Spe	F1-score
Unet	83.82	72.25	84.76	82.90	98.06	83.82
R2UNet	78.27	64.42	79.21	77.50	95.80	78.27
UNet++	88.88	80.14	87.57	90.23	98.95	88.88
SwinUNet	79.19	65.68	76.86	81.67	98.08	79.19
H2Former	83.74	72.13	83.37	NA	NA	83.74
MISSFormer	84.27	72.93	85.34	NA	NA	84.27
UMamba	77.46	63.33	80.43	NA	NA	77.46
VM-UNet	81.10	68.33	81.64	NA	NA	81.10
AC-Mamba	88.85	80.10	87.57	NA	NA	88.85
H-vmunet	70.33	54.38	71.10	NA	NA	70.33
OCTAMamba [18]	90.04	82.03	88.86	NA	NA	90.04
SAM OCTA [16]	NA	NA	NA	NA	NA	NA
SAM OCTA-2	NA	NA	NA	NA	NA	NA
Proposed	89.50	0.851	89.50	89.50	98.83	89.50

According to this experiment, the proposed CGOctaNet technique consistently improved performance across all datasets. In particular, it achieved the highest Dice values of 93.50%, 90.25 and 89.50% on the OCTA_3M, OCTA_6M, and ROSE datasets, respectively. The corresponding leading IoU scores were 86.70%, 84.50%, and 85.10%, and the Sensitivity was high at 89.50% on the ROSE dataset, indicating that the network can map vessel

boundaries. The CGOctaNet model provided performance gains compared to other current competing techniques, including SAM-OCTA, SAM-OCTA2, transformation-based architectures (SwinUNet and MISSFormer), and Other Techniques. Furthermore, for OCTA 3M, the proposed model has reported average performance of 99.10%, 99.25%, and 93.50% for precision, specificity, and F1-score, respectively. For OCTA 6M, the precision, specificity, and F1-score are reported as 92.07%, 99.15%, and 90.25%, respectively. Similarly, for the ROSE dataset, these performance metrics report 89.50%, 98.83%, and 89.50%, respectively. These outcomes demonstrate the robust performance of the proposed model when compared with the standard segmentation methods.

Table 3 *Comparison of SAM and proposed method on OCTA 500 (3M) dataset*

Class	OCTA 500 (3M)			
	Dice SAM	Dice Proposed	Jaccard SAM	Jaccard Proposed
RV	0.9199	0.9250	0.8520	0.8915
FAZ	0.9838	0.9880	0.9692	0.9750
Capillary	0.8785	0.8815	0.7837	0.8015
Artery	0.8747	0.9015	0.7785	0.8680
Vein	0.8817	0.9125	0.7897	0.8860

To validate the performance of the CGOctaNet approach, we conducted a detailed class-wise comparative analysis of the OCTA-500 (3M) and OCTA-500 (6M) datasets. The results show that the proposed model consistently outperformed the SAM baseline across all classes. According to this experiment, the proposed approach achieved better performance in artery and vein segmentation, with Dice scores increasing from 0.8747 to 0.9015 and from 0.8817 to 0.9125, respectively, and corresponding Jaccard scores improving from 0.7785 to 0.8680 and from 0.7897 to 0.8860, as shown in Table 4. Similarly, the FAZ region segmentation also reported the consistent improvement, achieving a Dice of 0.9880 and an IoU of 0.9750, indicating near-perfect segmentation.

Table 4 *Comparison of Dice and Jaccard scores between SAM and the proposed method on OCTA-500 (6M)*

Class	OCTA 500 (6M)			
	Dice SAM	Dice Proposed	Jaccard SAM	Jaccard Proposed
RV	0.8869	0.9015	0.7975	0.8215
FAZ	0.9073	0.9250	0.8473	0.8650
Capillary	0.8379	0.8560	0.7213	0.8515
Artery	0.8602	0.9115	0.7572	0.8510
Vein	0.8526	0.9230	0.7474	0.8215

Similar trends were observed on the OCTA-500 (6M) dataset, where the proposed CGOctaNet model reported improved performance for all segmentation classes. For instance, the Dice score for artery segmentation increased from 0.8602 (SAM) to 0.9115, while the vein segmentation improved from 0.8526 to 0.9230, ensuring the model's robustness in handling deeper retinal scans. The most significant IoU improvement was observed for capillaries, rising from 0.7213 to 0.8515, highlighting the proposed model's superior ability to segment fine-grained microvascular networks. The FAZ Dice improved from 0.9073 to 0.9250, further emphasizing consistent enhancement across both datasets.

Relative to other available OCTA segmentation models, CGOctaNet is more robust to image noise, illumination changes, and motion artefacts, which are typical in retinal imaging. The Transformer bottleneck is being integrated to achieve global spatial coherence by generating long-range dependencies between vessel structures, thereby reducing fragmentation of microvascular networks that often occurs in traditional CNN-based architectures. Also, a hybrid Dice-Cross Entropy loss is used to mitigate class imbalance and to create smoother vessel boundaries and better define the foveal avascular zone (FAZ).

Compared with the state-of-the-art models (SAM-OCTA) [16] and (OCTAMamba) [18], CGOctaNet showed better stability and accuracy, especially on capillary fine-caliber segmentation and FAZ. CGOctaNet has similar or better accuracy than SAM-based constructions, which are heavily dependent on pretraining and highly computationally intensive. On the same note, the proposed model also demonstrated higher segmentation continuity and lower topological inconsistency than HAIC-Net [20] and LA-Net [21], which both use hybrid CNN-Transformer frameworks to model local vessel geometry and global contextual information.

Generally, CGOctaNet offers the best balance between computational efficiency and segmentation performance. Its multi-scale feature extraction and attention-based reconstruction can provide consistent results across a variety of imaging conditions and datasets and are highly generalizable. The model is sensitive to extreme illumination artefacts, but it requires a moderate amount of graphics card memory (Transformer layers), and its accuracy and flexibility are more significant. The existing 2D model can also be strengthened in future research by the extension to 3D volumetric OCTA segmentation. Overall, CGOctaNet creates a state-of-the-art solution that

replaces retinal imaging with the use of convolutional encoding and Transformer-based global reasoning, providing a reliable, scalable, and efficient way to detect early neurodegenerative disease using retinal imaging.

5. Conclusion and Future Work

This article introduced a new deep learning method for vessel segmentation in Optical Coherence Tomography Angiography (OCTA) images. The proposed architecture has a U-Net-like encoder-decoder backbone and a bottleneck module based on a Transformer to effectively model and capture the global contextual relations among spatial features. Moreover, the decoder provides an attention module called Convolutional Block Attention Modules (CBAM), which refines the feature and adaptively focuses the structure of interest in upsampling, i.e., vessel structure. The efficiency of the CGOctaNet mode was also evaluated on the OCTA-500 dataset, where the CGOctaNet model outperformed existing approaches, achieving high Dice coefficients and IoUs across all vascular classes, including arteries, veins, and capillaries. Future research can aim to optimize computational efficiency through lightweight Transformer variants, extend the architecture to 3D OCTA segmentation for volumetric vascular analysis, and conduct multi-center clinical validation to establish diagnostic reliability. The incorporation of explainable AI techniques can further enhance interpretability and facilitate clinical adoption. In summary, CGOctaNet offers a scientifically justified, computationally efficient, and generalizable method for OCTA vessel segmentation, presenting a promising foundation for early detection of neurodegenerative disorders and broader applications in ophthalmic image analysis.

Acknowledgement

The authors would like to thank NMAM Institute of Technology (NMAMIT), NITTE, Karkala, and Malnad College of Engineering, Hassan, India, for their support and encouragement in carrying out this research work.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

Conceptualization, methodology, software implementation, validation, formal analysis, investigation, resource management, data curation, original draft preparation, review and editing, and visualization were performed by the 1st, 4th, and 5th authors. Supervision and project administration were carried out by the 2nd and 3rd authors. The 2nd author made a significant contribution in designing the study framework and analyzing the outcomes.

References

[1] Ibrahim, Y., Xie, J., Macerollo, A., Sardone, R., Shen, Y., Romano, V., & Zheng, Y. (2023). A systematic review on retinal biomarkers to diagnose dementia from OCT/OCTA images. *Journal of Alzheimer's Disease Reports*, *7*(1), 1201-1235.

[2] World Health Organization. (2022). *A blueprint for dementia research*. World Health Organization.

[3] NHS-UK, About dementia - dementia guide, NHS, https://www.nhs.uk/conditions/dementia/about/?tabname= about-dementia

[4] NIA-scientists, How biomarkers help diagnose dementia, National Institute on Aging, https://www.nia.nih.gov/health/how-biomarkers-helpdiagnose-dementia

[5] NHS-UK, MRI scan, https://www.nhs.uk/conditions/mriscan/

[6] Snyder, P. J., Alber, J., Alt, C., Bain, L. J., Bouma, B. E., Bouwman, F. H., ... & Snyder, H. M. (2021). Retinal imaging in Alzheimer's and neurodegenerative diseases. *Alzheimer's & Dementia*, *17*(1), 103-111.

[7] Ngolab, J., Honma, P., & Rissman, R. A. (2019). Reflections on the utility of the retina as a biomarker for Alzheimer's disease: a literature review. *Neurology and therapy*, *8*(Suppl 2), 57-72.

[8] García-Bermúdez, M. Y., Vohra, R., Freude, K., Wijngaarden, P. V., Martin, K., Thomsen, M. S., ... & Kolko, M. (2023). Potential retinal biomarkers in Alzheimer's disease. *International Journal of Molecular Sciences*, *24*(21), 15834.Vij, R., & Arora, S. (2022).

[9] Vij, R., & Arora, S. (2022). A systematic survey of advances in retinal imaging modalities for Alzheimer's disease diagnosis. *Metabolic Brain Disease*, *37*(7), 2213-2243.

[10] Meiburger, K. M., Salvi, M., Rotunno, G., Drexler, W., & Liu, M. (2021). Automatic segmentation and classification methods using optical coherence tomography angiography (OCTA): A review and handbook. *Applied Sciences*, *11*(20), 9734.

[11] Chen, K., Gao, G., Yang, X., Wang, W., & Na, J. (2025). Denoising, segmentation and volumetric rendering of optical coherence tomography angiography (OCTA) image using deep learning techniques: a review. *arXiv preprint arXiv:2502.14935*.

[12] Totolici, G., Miron, M., & Culea-Florescu, A. L. (2024). Automatic Segmentation and Statistical Analysis of the Foveal Avascular Zone. *Technologies*, *12*(12), 235.

[13] Liu, X., Zhu, H., Zhang, H., & Xia, S. (2024). The framework of quantifying biomarkers of OCT and OCTA images in retinal diseases. *Sensors*, *24*(16), 5227.

[14] Akbar, S. (2025). A Hybrid Multi-Level Segmentation-Based Ensemble Classification Model for Multi-Class Diabetic Retinopathy Detection. In *Computational Techniques for Biological Sequence Analysis* (pp. 112-131). CRC Press.

[15] Udayaraju, P., Jeyanthi, P., & Sekhar, B. V. D. S. (2025). Hierarchical convolution neural network models for classifying the segmented OCT and OCTA images using U-Net model. *Multimedia Tools and Applications*, *84*(19), 20311-20337.

[16] Wang, C., Chen, X., Ning, H., & Li, S. (2024, April). Sam-octa: A fine-tuning strategy for applying foundation model octa image segmentation tasks. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1771-1775). IEEE.

[17] Wang, C., Chen, X., Ning, H., & Li, S. (2024, April). Sam-octa: A fine-tuning strategy for applying foundation model octa image segmentation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1771-1775). IEEE.

[18] Zou, S., Zhang, Z., & Gao, G. (2025, April). Octamamba: A state-space model approach for precision octa vasculature segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[19] Wittmann, B., Glandorf, L., Paetzold, J. C., Amiranashvili, T., Wälchli, T., Razansky, D., & Menze, B. (2024, October). Simulation-based segmentation of blood vessels in cerebral 3D OCTA images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 645-655). Cham: Springer Nature Switzerland.

[20] Shen, H., Tang, Z., Li, Y., Duan, X., & Chen, Z. (2024). HAIC-NET: Semi-supervised OCTA vessel segmentation with self-supervised pretext task and dual consistency training. *Pattern Recognition*, *151*, 110429.

[21] Yang, C., Li, B., Xiao, Q., Bai, Y., Li, Y., Li, Z., ... & Li, H. (2024). LA-Net: layer attention network for 3D-to-2D retinal vessel segmentation in OCTA images. *Physics in Medicine & Biology*, *69*(4), 045019.

[22] Jiang, H., & Jiang, Y. (2024, May). Octa retinal image segmentation method based on improved SegNet. In *2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 362-367). IEEE.

[23] Liu, Y., Shen, J., Yang, L., Bian, G., & Yu, H. (2023). ResDO-UNet: A deep residual network for accurate retinal vessel segmentation from fundus images. *Biomedical Signal Processing and Control*, *79*, 104087.