EdgeFormerNet++: A Hybrid Attention-Guided Transformer Network for Accurate Retinal Layer Segmentation in OCT Images

Anju Thomas, Farhin Janath S. J., Vijayalakshmi P., Nisan Pranavah Raja, P. Palanisamy, and Varun P. Gopi, *Senior Member, IEEE*

***Abstract*—Accurate retinal layer segmentation in optical coherence tomography (OCT) is essential for the early detection and monitoring of retinal and neurodegenerative diseases. However, this task remains challenging due to the presence of thin, low-contrast structures and closely spaced layers in multi-class settings. In this work, we propose EdgeFormerNet++, a hybrid framework that integrates Res2Net-based multi-scale encoding with a lightweight Transformer bottleneck to capture both local and global contextual features. To further enhance feature representation, dual attention mechanisms (CBAM and scSE) are employed for effective spatial and channel-wise recalibration, along with an edge attention module for improved boundary refinement. The proposed model is evaluated on two publicly available datasets, NR206 and MGU, covering both macular and peripapillary regions. Experimental results demonstrate that EdgeFormerNet++ achieves competitive performance, with Dice scores of 91.35% and 81.21% on NR206 and MGU, respectively, along with consistent IoU and pixel accuracy across datasets. Both qualitative and quantitative analyses indicate improved boundary delineation and robustness, particularly in challenging regions with high anatomical variability. These findings highlight the effectiveness of the proposed framework as a reliable and scalable solution for automated OCT segmentation and computer-assisted diagnosis.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10551

***Index Terms*—Deep Learning, Edge Attention, Hybrid CNN-Transformer Architecture, Multi-class Semantic Segmentation, Optical Coherence Tomography (OCT), Res2Net, Retinal Layer Segmentation.**

I. Introduction

OPTICAL COHERENCE TOMOGRAPHY (OCT) is a widely used, non-invasive imaging modality that provides depth-resolved visualization of ocular microstructures. Using near-infrared low-coherence interferometry, OCT acquires one-dimensional reflectivity profiles (A-scans), which are laterally assembled into cross-sectional B-scans; stacking B-scans yields volumetric retinal representations with micron-scale detail [1]. Although originally developed for ophthalmology, OCT has expanded to multiple clinical domains, including dermatology, oncology, gynecology, and intraoperative imaging [2]. Furthermore, as the retina is an extension of the central nervous system, OCT is increasingly used to study neurological and psychiatric disorders [3]. Fig. 1 illustrates the major retinal layers considered in this work. Quantitative alterations in retinal layer thickness and morphology can serve as early biomarkers for retinal and systemic diseases, including Diabetic Retinopathy (DR), Multiple Sclerosis (MS), Alzheimer's Disease (AD), Wolfram Syndrome (WS), and Parkinson's Disease (PD) [4]. Accurate layer segmentation enables reliable thickness measurement and comparison against normative references, supporting early detection and longitudinal monitoring. Manual delineation by experts is time-consuming and subject to inter-observer variability, which motivates the development of automated approaches. While classical methods often lacked robustness (e.g., low Dice/IoU and inconsistent boundaries), deep learning-based models have substantially improved pixel-wise accuracy and agreement with clinical annotations, enabling near-human-level performance in retinal layer segmentation.

Despite recent progress, automated retinal layer segmentation remains a challenging task, particularly in multi-class scenarios that require the simultaneous delineation of multiple retinal layers (typically 9–10 depending on dataset annotations). Unlike binary segmentation, layer-wise segmentation demands pixel-accurate separation of tightly packed, low-contrast, and morphologically variable structures with subtle transitions and overlapping intensity profiles, which are further exacerbated by pathology and poor scan quality. Numerous deep learning models have been explored, including U-Net [6], ResU-Net [7], Attention U-Net [8], and more recent transformer-based variants such as Swin-UNet [9]. Although these architectures can capture contextual and long-range dependencies, their performance is strongly influenced by the availability of high-quality expert annotations, which are limited by privacy constraints and labeling costs. To address these issues, we propose a hybrid architecture that couples convolutional feature extraction with Transformer modules and complementary attention mechanisms, enabling robust layer segmentation under limited data and degraded image conditions. We validate the framework using both quantitative

The associate editor coordinating the review of this manuscript and approving it for publication was Anabel Martin (*Corresponding author: Varun P. Gopi*).

A. Thomas, N. P. Raja, Palanisamy P., and Varun P. Gopi are with the Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tamil Nadu, India (e-mails: anju@nitt.edu, 408122004@nitt.edu, palan@nitt.edu, and varun@nitt.edu).

S. J. F. Janath and P. Vijayalakshmi are with the Department of Electronics and Electronics Engineering, Government Engineering College, Barton Hill, Trivandrum, Kerala, India (e-mails: farhin.trv23ec023@gecbh.ac.in, and vijayalakshmi.trv23ec069@gecbh.ac.in).

A. Thomas is also with the Department of Sensor & Biomedical Technology, School of Electronics Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India (e-mail: nju.t@vit.ac.in).

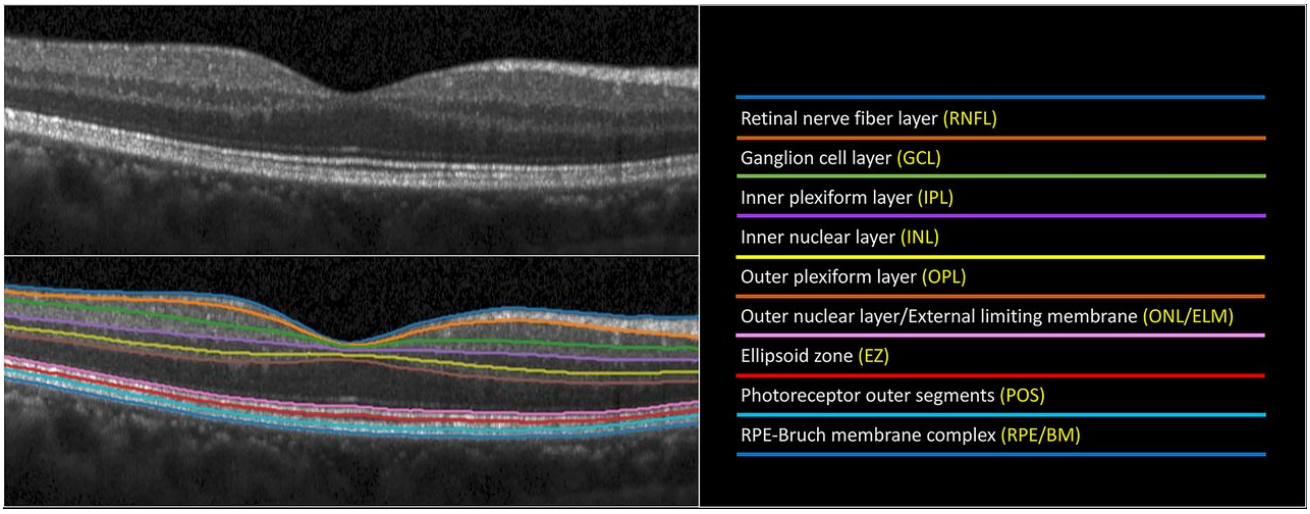


Fig. 1. Layered structure of the retina [5].

metrics and qualitative assessments.

II. Literature Review

OCT enables non-invasive, high-resolution visualization of retinal microstructures; however, manual layer delineation is time-consuming and prone to inter-observer variability. Accordingly, OCT segmentation has evolved from classical image processing to learning-based and transformer-driven methods. Early handcrafted pipelines used thresholding [10], [11], edge detection (e.g., Canny) [12], active contours [13], and graph-based optimization such as Graph Cuts [14]. While computationally tractable, these methods are often sensitive to speckle noise, vessel shadows, and anatomical variability. Traditional machine learning approaches (e.g., SVMs [15] and Random Forests [16]) have improved robustness but rely on feature engineering and provide limited spatial context, degrading performance under class imbalance and cross-device shifts. Deep learning enabled end-to-end segmentation from OCT images, with U-Net [17] and its variants becoming standard. ReLayNet [18] and Attention U-Net [8] improved multi-layer segmentation via boundary-aware designs and attention gates, while Residual Attention U-Net [19] enhanced boundary precision at higher complexity. SD-LayerNet [20] explored semi-supervised learning with anatomical priors for low-label settings. More recently, CNN-Transformer hybrids have been introduced to capture long-range context. TransUNet [21] improves global modeling but increases computation, whereas Swin-UNet [9] and UTNet [22] offer more efficient attention mechanisms. UNetR [23] extends these ideas to 3D volumes using ViT encoders, typically requiring substantial GPU resources. Transformer architectures have also been tailored to the OCT structure. Domain-specific and lightweight models include MGUNet [24], MT-Net [25], LightReSeg [26], and RC-FEDM [27], which emphasize efficiency and boundary regularization. OCT segmentation has moved toward hybrid designs that combine multi-scale feature learning, global attention, and anatomical constraints. However, cross-device generalization, robustness to noise, and accurate delineation of thin or pathological layers remain open challenges, motivating the framework proposed in this work. The main contributions are

- A hybrid encoder-decoder architecture (EdgeFormerNet++) for accurate retinal layer segmentation in OCT images.
- An Edge Attention Module that strengthens boundary delineation using gradient-informed cues.
- Res2Net-based encoders for within-block multi-scale feature extraction and improved structural representation.
- A Transformer-CBAMM fusion bottleneck for global dependency modeling with local attention refinement.
- scSE blocks in the decoder for spatial and channel feature recalibration.

III. Methodology

We propose EdgeFormerNet++, a hybrid encoder-decoder network for retinal layer segmentation in OCT. The Complete data processing pipeline of the proposed model is in Fig. 2. The design focuses on (i) multi-scale feature extraction, (ii) boundary-aware learning, and (iii) spatial-channel recalibration during decoding. An Edge-Aware Attention Module is applied at the input to strengthen boundary cues, especially in regions with weak or ambiguous layer transitions. The encoder uses Res2Net blocks to learn hierarchical multi-scale representations. At the bottleneck, a Transformer-CBAM module combines global dependency modeling with local attention refinement. The decoder integrates scSE blocks to recalibrate spatial and channel responses during upsampling, improving layer reconstruction and boundary precision. Fig. 3 provides an overview of the proposed architecture.

A. Input Preprocessing

As shown in Fig. 4, the input OCT image is first processed by a preprocessing module to enhance structural cues before encoding. An Edge Attention block highlights anatomical boundaries via convolutional filtering and multiplicative enhancement. Its output is concatenated channel-wise with the original image to retain both texture and boundary-aware information. The fused

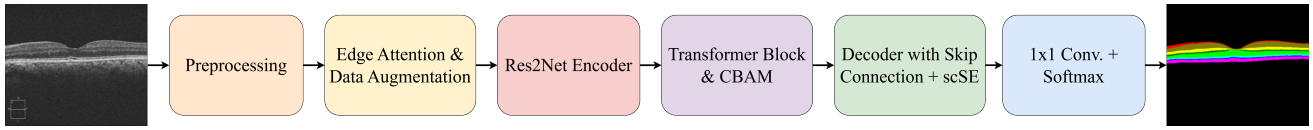


Fig. 2. Complete data processing pipeline of the proposed EdgeFormerNet++.

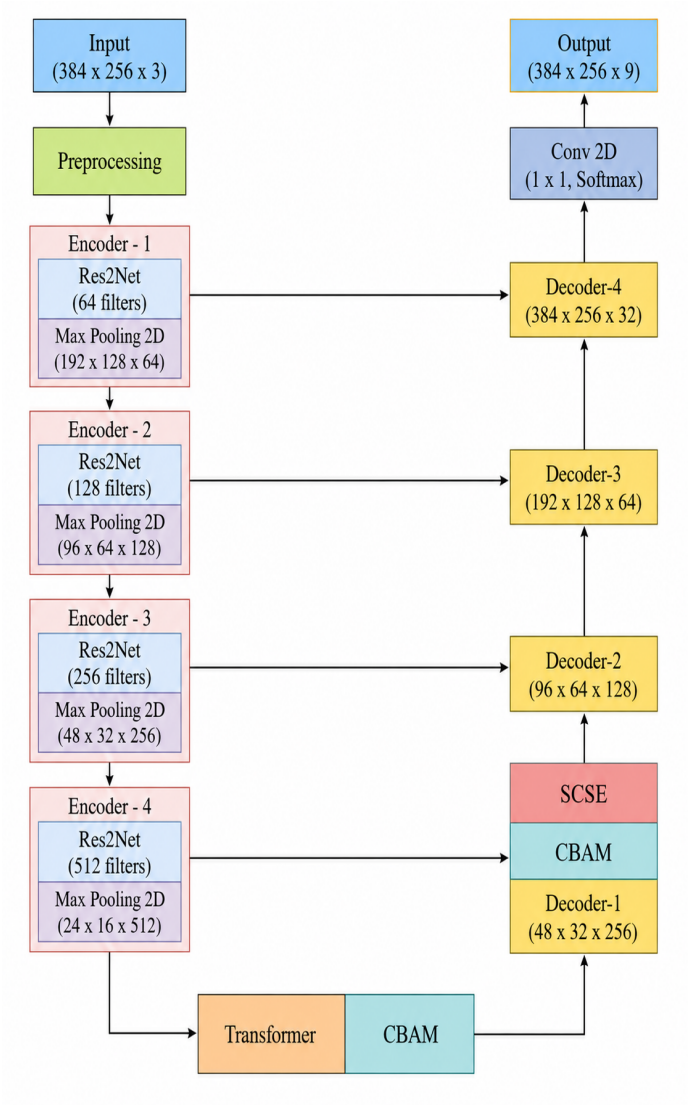


Fig. 3. Overall architecture of the proposed EdgeFormerNet++ for retinal layer segmentation. The framework integrates edge-aware preprocessing, a Res2Net encoder, a Transformer-CBAM bottleneck, and an scSE-enhanced decoder to generate pixel-wise retinal layer segmentation maps.

tensor is then passed through a convolutional stem consisting of a 3×3 convolution, Batch Normalization (BN) [28], and ReLU activation [29], producing a normalized feature representation for subsequent encoder stages.

1) Edge Attention Module: Boundary-Aware Enhancement: To improve sensitivity to retinal layer boundaries, an **Edge Attention Module** is applied in preprocessing (Fig. 5). A Sobel operator is first applied to the input image X to extract gradient responses. The resulting edge map is refined by two successive 3×3 convolutions with ReLU and sigmoid activations:

$$E = \sigma(\mathrm{Conv}_2(\mathrm{ReLU}(\mathrm{Conv}_1(\mathrm{Sobel}(X))))) \,. \tag{1}$$

The refined attention map E is then used to modulate the input via element-wise multiplication:

$$X' = E \odot X. \tag{2}$$

This edge-guided gating emphasizes high-gradient regions and improves the separation of subtle transitions between adjacent retinal layers.

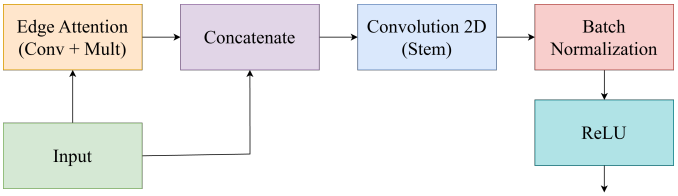


Fig. 4. Preprocessing stage of EdgeFormerNet++, where edge-enhanced features are fused with the input OCT image and refined before encoder processing.

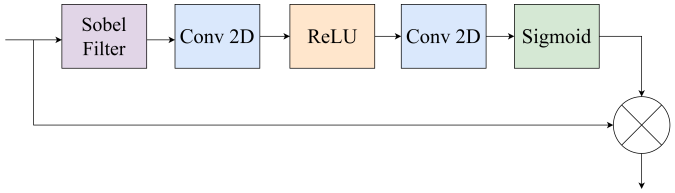


Fig. 5. Block diagram of the Edge Attention Module, which enhances retinal layer boundaries through gradient-guided attention and feature refinement

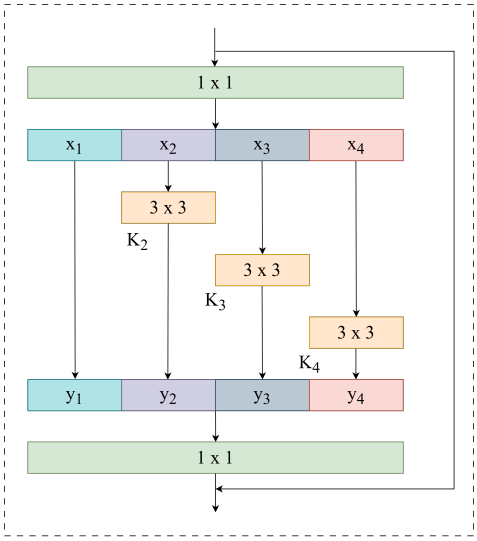


Fig. 6. Res2Net architecture employing four channel splits for hierarchical multi-scale feature learning

B. Res2Net Encoder: Multi-Scale Feature Representation

To capture features at multiple receptive fields, each encoder stage employs a **Res2Net block** [30] (Fig. 6). In contrast to standard residual blocks, Res2Net partitions channels into groups and processes them hierarchically, enabling multi-scale representation within a single residual unit.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, a 1×1 convolution first produces

$$\mathbf{X}' = \mathrm{Conv}_{1\times1}(\mathbf{X}), \quad \mathbf{X}' \in \mathbb{R}^{C'\times H\times W}. \tag{3}$$

The output is split into $s = 4$ groups $\{x_i\}_{i=1}^{4}$, where $x_i \in \mathbb{R}^{\frac{C'}{4}\times H\times W}$:

$$\mathbf{X}' = \{x_1, x_2, x_3, x_4\}. \tag{4}$$

Hierarchical residual-like aggregation is then applied:

$$y_1 = x_1, \tag{5}$$
$$y_2 = \mathrm{Conv}_{3\times3}(x_2 + y_1), \tag{6}$$
$$y_3 = \mathrm{Conv}_{3\times3}(x_3 + y_2), \tag{7}$$
$$y_4 = \mathrm{Conv}_{3\times3}(x_4 + y_3). \tag{8}$$

The group outputs are concatenated and fused using a final 1×1 convolution:

$$Y = \mathrm{Concat}(y_1, y_2, y_3, y_4), \qquad \tilde{Y} = \mathrm{Conv}_{1\times1}(Y). \tag{9}$$

Finally, a residual connection preserves the input signal:

$$\hat{Y} = \tilde{Y} + \mathbf{X}. \tag{10}$$

This design enhances encoder capacity by progressively learning larger receptive fields across groups while maintaining computational efficiency and stable optimization through residual addition.

After each Res2Net block, a 2×2 max pooling operation is applied to progressively downsample the spatial resolution:

$$\mathbf{X}_{\text{pooled}} = \mathrm{MaxPool}_{2\times2}(\hat{Y}) \tag{11}$$

This hierarchical downsampling allows the encoder to capture both fine and coarse retinal structures effectively. The integration of Res2Net blocks facilitates internal multi-scale fusion at each stage, enabling efficient and robust extraction of retinal features.

C. Transformer-CBAM Bottleneck

To connect the encoder and decoder, we use a bottleneck that sequentially applies a **Transformer block** followed by **CBAM**. The Transformer models long-range dependencies in the encoded representation, while CBAM enhances local discrimination through channel and spatial attention. This global-local integration is placed after the deepest encoder output (Fig. 3), which is beneficial for retinal OCT, where both overall anatomy and fine layer cues are critical.

1) Transformer Block for Global Context Modeling: Given the encoder output $\mathbf{F}_e \in \mathbb{R}^{C\times H\times W}$ (here, $\mathbf{F}_e \in \mathbb{R}^{512\times24\times16}$), we flatten it into a sequence of $N = H \cdot W$ tokens, forming $\mathbf{X} \in \mathbb{R}^{N\times d}$. Queries, keys, and values are obtained via linear projections:

$$\mathbf{Q} = \mathbf{X}W^Q, \quad \mathbf{K} = \mathbf{X}W^K, \quad \mathbf{V} = \mathbf{X}W^V. \tag{12}$$

Scaled dot-product attention is computed as

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{13}$$

where d_k is the key dimensionality. Residual connections and normalization stabilize training, followed by a feed-forward network (FFN):

$$\mathbf{Z} = \mathrm{FFN}(\mathrm{LayerNorm}(\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{X})) + \mathbf{X}. \tag{14}$$

The output $\mathbf{Z}$ is reshaped back to $(C \times H \times W)$ for subsequent processing by CBAM (Fig. 7).

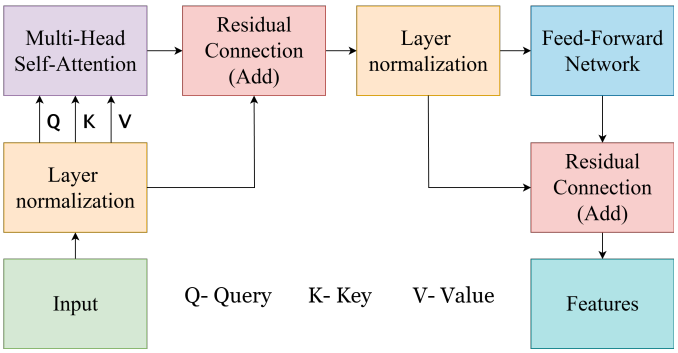


Fig. 7. Transformer bottleneck architecture used to capture long-range contextual dependencies and enhance global feature representation

2) CBAM for Local Feature Emphasis: Although the Transformer captures global context, it does not explicitly emphasize locally salient regions. Therefore, the bottleneck output is refined using CBAM [31], which applies channel attention followed by spatial attention (Fig. 8).

Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the CBAM input. For *channel attention*, global average pooling and max pooling over spatial dimensions produce two descriptors in $\mathbb{R}^{C\times 1\times 1}$, which are passed through a shared MLP and combined:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F}))) . \tag{15}$$

The feature map is reweighted channel-wise:

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}. \tag{16}$$

For *spatial attention*, average and max pooling across channels yield two maps in $\mathbb{R}^{1\times H\times W}$, which are concatenated and convolved with a 7×7 kernel:

$$\mathbf{M}_s(\mathbf{F}') = \sigma\big(f^{7\times 7}([\mathrm{AvgPool}(\mathbf{F}'); \mathrm{MaxPool}(\mathbf{F}')])\big) . \tag{17}$$

The final refined output is

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}'. \tag{18}$$

This sequential attention reweights informative channels and spatial locations, improving focus on diagnostically relevant retinal regions.

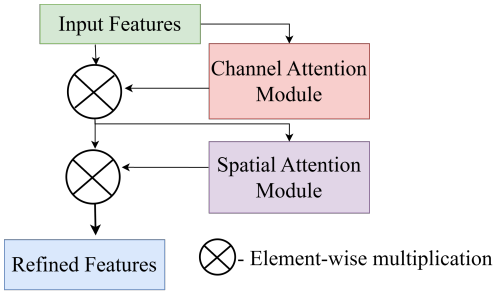


Fig. 8. CBAM architecture for adaptive channel and spatial attention-based feature refinement

3) Combined Effect of Transformer and CBAM: Placing the Transformer and CBAM sequentially at the bottleneck combines global dependency modeling with local feature refinement. The Transformer contextualizes each token using information from the full feature map, while CBAM selectively reweights informative channels and spatial locations. This global-local synergy improves sensitivity to subtle layer variations and weak pathological cues in retinal OCT.

D. Decoder with scSE Attention

The decoder progressively reconstructs high-resolution predictions from the latent representation by upsampling, fusing encoder features via skip connections, and refining the merged responses (Fig. 9). Each stage begins with a transposed convolution:

$$\mathbf{F}_{\mathrm{up}} = \mathrm{ConvTranspose}(\mathbf{F}_{\mathrm{in}}), \tag{19}$$

where $\mathbf{F}_{\mathrm{in}} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_{\mathrm{up}} \in \mathbb{R}^{C'\times 2H\times 2W}$. The upsampled features are concatenated with the corresponding encoder output $\mathbf{F}_{\mathrm{enc}}$:

$$\mathbf{F}_{\mathrm{cat}} = \mathrm{Concat}(\mathbf{F}_{\mathrm{up}}, \mathbf{F}_{\mathrm{enc}}), \tag{20}$$

followed by two 3×3 convolutional layers with batch normalization and ReLU:

$$\begin{aligned} \mathbf{F}_1 &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times 3}(\mathbf{F}_{\mathrm{cat}}))) , \\ \mathbf{F}_2 &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times 3}(\mathbf{F}_1))) . \end{aligned} \tag{21}$$

To further recalibrate decoder features, we apply a spatial and channel squeeze-and-excitation (scSE) block [32] (Fig. 10) after the convolutional refinement at each stage.

a) Channel Squeeze-and-Excitation (cSE): Global average pooling produces a channel descriptor,

$$\mathbf{z}_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}_2(c,i,j), \quad \forall c \in \{1,\ldots,C\}, \tag{22}$$

which is passed through two fully connected layers with ReLU (δ) and sigmoid (σ):

$$\mathbf{s}_c = \sigma(\mathbf{W}_2\,\delta(\mathbf{W}_1\mathbf{z}_c)) , \qquad \mathbf{F}_{\mathrm{cSE}} = \mathbf{s}_c \otimes \mathbf{F}_2. \tag{23}$$

b) Spatial Squeeze-and-Excitation (sSE): A 1×1 convolution followed by a sigmoid yields a spatial attention map

$$\mathbf{M}_s = \sigma(\mathrm{Conv}_{1\times 1}(\mathbf{F}_2)) \in \mathbb{R}^{1\times H\times W}, \qquad \mathbf{F}_{\mathrm{sSE}} = \mathbf{M}_s \odot \mathbf{F}_2. \tag{24}$$

c) scSE Fusion: The final scSE output is obtained by combining both branches:

$$\mathbf{F}_{\mathrm{scSE}} = \mathbf{F}_{\mathrm{cSE}} + \mathbf{F}_{\mathrm{sSE}}. \tag{25}$$

This dual recalibration emphasizes salient spatial regions and informative channels, improving boundary-sensitive layer reconstruction during decoding.

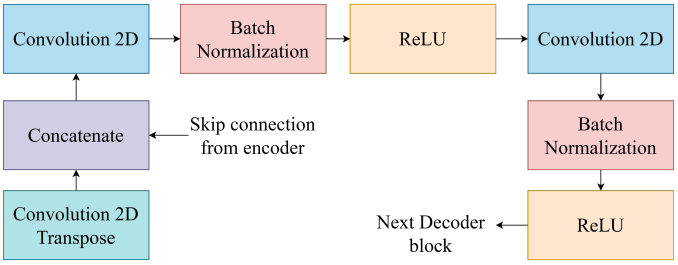


Fig. 9. Decoder block for progressive feature reconstruction using skip connections and convolutional refinement.

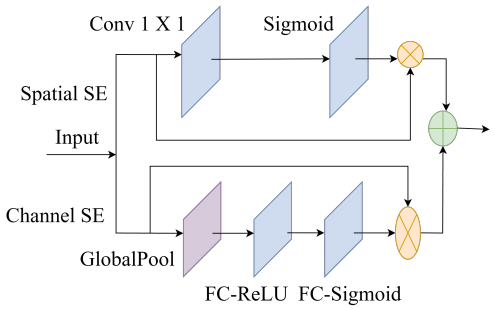


Fig. 10. scSE block for simultaneous spatial and channel-wise feature recalibration during decoding

E. Final Segmentation Layer

The final scSE-refined decoder output $\mathbf{F}_{\text{dec}}$ is projected to class logits using a 1×1 convolution and converted to per-pixel class probabilities with softmax:

$$\mathbf{S} = \text{Softmax}(\text{Conv}_{1\times1}(\mathbf{F}_{\text{dec}}))\,. \tag{26}$$

This yields $\mathbf{S} \in \mathbb{R}^{C\times H\times W}$, where each channel represents a retinal-layer class. The output dimension is adapted based on the number of annotated classes in the dataset (e.g., 9 classes for MGU). Softmax enforces a valid probability distribution at each pixel, enabling interpretable multi-class predictions.

IV. Result

A. Dataset Description

We evaluated the proposed method using two publicly available OCT datasets. NR206 contains 206 high-resolution macular B-scans of structurally normal eyes from OCTID [33], acquired at Sankara Nethralaya Eye Hospital (Chennai, India) using a Cirrus HD-OCT system with a 2 mm scan protocol (840 nm; $\sim$5 μm axial and $\sim$15 μm lateral resolution). Each image is 500×750 and a fovea-centered B-scan, manually selected per volume; all scans were de-identified and used for retinal layer segmentation. The MGU dataset [24] provides peripapillary B-scans centered on the optic nerve head from 87 OCT volumes collected at Shanghai General Hospital; 244 B-scans were used in this study. Images have a resolution of 1024×992 pixels (approximately 20.48×7.94 mm). Expert annotations include nine retinal substructures (RNFL to choroid) and the optic disc boundary, generated in ITK-SNAP and clinically verified. The number of segmentation classes is dataset-dependent; for instance, the MGU dataset contains nine annotated retinal substructures, and the model output is configured accordingly. Table I reports the train/validation/test splits for both datasets.

TABLE I
Dataset Split Details for NR206 and MGU

Dataset	Training Samples	Validation Samples	Test Samples
NR206 [33]	126	40	40
MGU [24]	148	48	48

B. Data Augmentation

To improve robustness and reduce overfitting, the same augmentation pipeline was applied to all training samples from NR206 and MGU, reflecting the variability in clinical OCT acquisition. Spatial augmentations included horizontal/vertical flips as well as random 90° rotations. Elastic deformations and grid distortions were added to simulate anatomical variability and device-induced warping. Intensity augmentations (brightness/contrast changes) and image degradations (blur and noise) were used to emulate differing scan quality. All transformations were implemented with Albumentations [34]. Images were resized to 384×256 and normalized to the $[0, 1]$ range.

C. Evaluation Metrics

Segmentation performance of EdgeFormerNet++ is evaluated using three standard metrics: Dice Similarity Coefficient (DSC) [35], [36], Pixel Accuracy (PA) [37], and Intersection over Union (IoU). Together, they provide complementary views of overlap quality, pixel-wise correctness, and region agreement with expert annotations.

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}. \tag{27}$$

Here, $|X|$ and $|Y|$ denote the number of foreground pixels in the prediction and reference, respectively, and $|X \cap Y|$ is their intersection. DSC ranges from 0 to 1, with higher values indicating better agreement.

$$\text{Pixel Accuracy} = \frac{\text{Correctly Predicted Pixels}}{\text{Total Pixels}}. \tag{28}$$

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|}. \tag{29}$$

For methods where IoU was not explicitly reported, it is estimated from the Dice score using the relation:

$$\text{IoU} = \frac{\text{DSC}}{2 - \text{DSC}} \tag{30}$$

Pixel Accuracy is reported for completeness; however, Dice and IoU are emphasized as primary metrics due to their robustness to class imbalance.

D. Experimental Setup

All experiments were conducted on a high-performance workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), 128 GB system memory, and a multi-core CPU. The proposed EdgeFormerNet++ model was implemented in Python (version 3.x) using the TensorFlow deep learning framework (version 2.x) with the Keras high-level API. GPU acceleration was enabled through CUDA (version 11.2) and cuDNN libraries, ensuring efficient training and inference. Data preprocessing and augmentation were performed using standard scientific computing libraries such as NumPy and PIL, along with the Albumentations library, which provides advanced augmentation strategies, including geometric transformations, elastic deformations, and intensity variations. The model architecture was developed using TensorFlow/Keras layers, integrating convolutional operations, normalization layers, transposed convolutions, and attention mechanisms to effectively capture both local and global contextual information. Training was carried out using a mini-batch optimization strategy with custom-defined loss functions, specifically a hybrid combination of Dice and Tversky losses implemented in TensorFlow. Detailed configurations of the model architecture, training parameters, and optimization settings are summarized in Table II.

TABLE II
MODEL & TRAINING PARAMETERS OF EDGEFORMERNET++

Parameter	Value
Input Image Size	$384 \times 256 \times 3$
Number of Classes	Derived from mask color mapping
Normalization	Pixel values scaled to $[0, 1]$
Batch Size	4
Epochs	400
Optimizer	Adam
Learning Rate	1×10^{-3}
Weight Initialization	Glorot Uniform (seed = 42)
Dice Loss	Included
Tversky Loss	$\alpha = 0.6$, $\beta = 0.4$
Final Loss	$0.5\times$ Dice + $0.5\times$ Tversky
Attention Heads	4
FFN Dimension	128
LR Scheduler	ReduceLROnPlateau
LR Factor	0.5
Patience	30
Minimum LR	1×10^{-6}

E. Qualitative Analysis

Fig. 11 shows a qualitative comparison of input OCT scans, ground truth labels, and the segmentations produced by the proposed network. Each row corresponds to a different test case, highlighting consistent layer delineation across varying anatomies and contrasts. The predictions closely match the annotations, capturing both thick and thin layers with strong visual agreement. Notably, the boundaries of delicate structures, such as the retinal nerve fiber layer and the photoreceptor layer, remain well-defined, indicating reliable detection of layer transitions even in low-contrast regions. This behavior is supported by multi-scale feature extraction in the encoder, long-range dependency modeling at the bottleneck, and attention-based refinement during decoding. Overall, the examples confirm anatomical fidelity and support the quantitative gains.

F. Per-layer Segmentation Performance

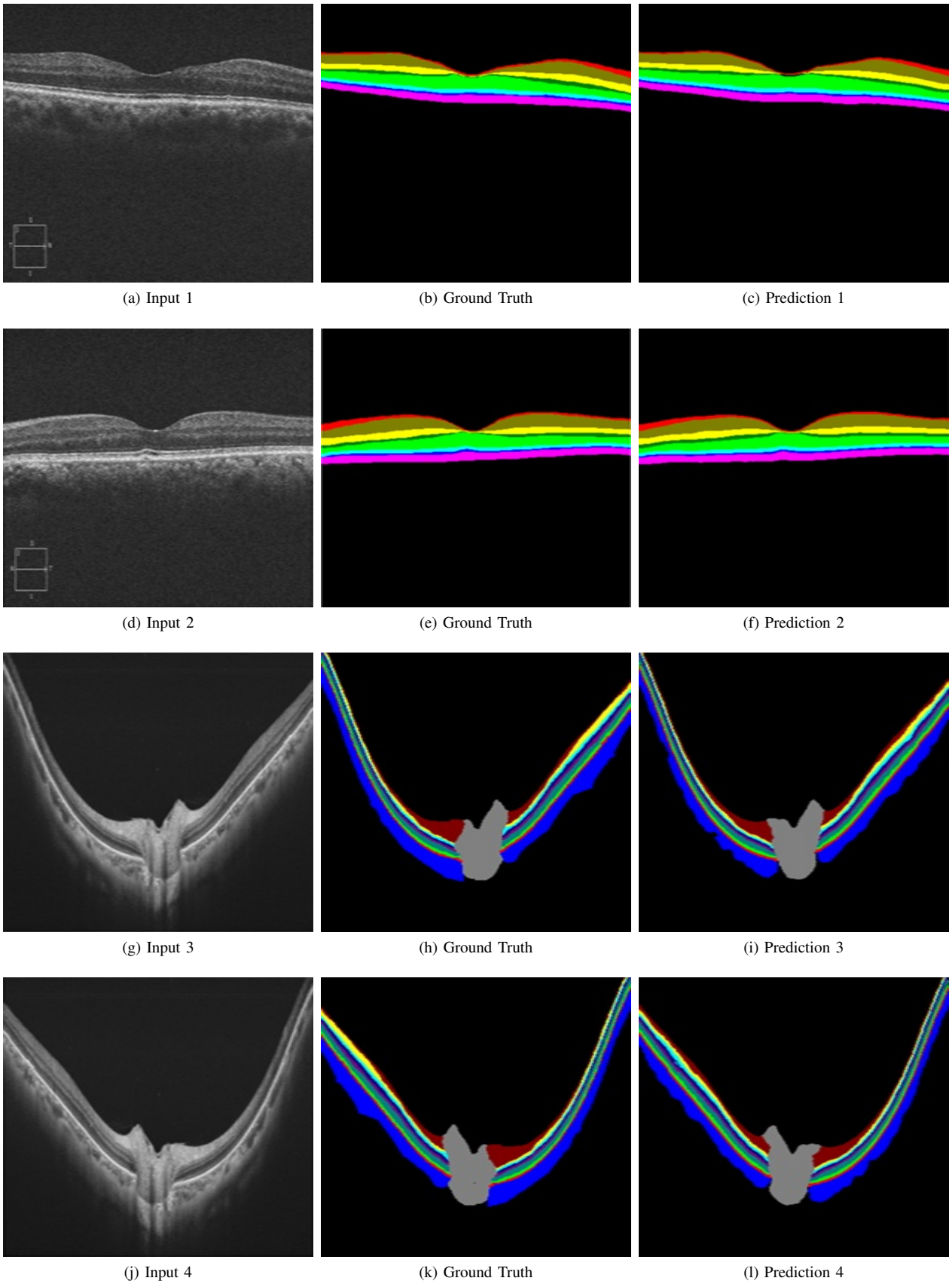
Tables III and IV show the layer-wise results for the proposed segmentation model on the NR206 and MGU datasets. The model on NR206 achieves strong segmentation across most retinal layers. ONL and OS layers achieve the highest performance, with Dice and IoU exceeding 96% and 92%, respectively, indicating more accurate segmentation and more efficient boundary marking than the other layers. Even for the model's segmentation performance on the ELM+IS and GCL+IPL more challenging layers, the results remain high, achieving an average Dice of 91.35% and a mean IoU of 84.47%. On the MGU dataset, the model results show a drop in performance well below average, due to the more complex architecture and high image variation. The results for the deeper layers, ONL and Choroid, continue to show high performance results of above 89%. However, segmentation performance for the thinner, less-defined layers, GCL and IPL, is lower and highly variable. The model on the MGU dataset shows an overall drop in performance, with a Dice of 81.21% and a mean IoU of 69.57%. Results on the MGU dataset and NR206 further show that segmenting fine-grained layers of the retinal architecture on complex datasets remains challenging for the proposed segmentation model.

TABLE III
LAYER-WISE PERFORMANCE ON NR206

Layer	Dice (%)	IoU (%)	PA (%)
NFL	86.35 ± 3.87	76.18 ± 5.83	87.58 ± 5.77
GCL+IPL	82.99 ± 4.15	71.14 ± 6.00	82.64 ± 5.80
INL	95.45 ± 1.38	91.33 ± 2.50	95.70 ± 2.55
OPL	92.45 ± 1.64	86.00 ± 2.82	92.77 ± 3.46
ONL	96.40 ± 0.86	93.06 ± 1.60	96.27 ± 1.65
ELM+IS	89.90 ± 2.33	81.74 ± 3.72	89.56 ± 4.17
OS	96.00 ± 0.87	92.32 ± 1.60	95.74 ± 1.68
RPE	91.24 ± 2.50	83.98 ± 4.16	91.92 ± 3.76
Average	**91.35 ± 1.19**	**84.47 ± 1.89**	**91.52 ± 1.41**

G. Quantitative Analysis

The quantitative evaluation of the proposed method on NR206 and MGU datasets is summarized in Table V. On the NR206 dataset, the segmentation method achieves a Dice score of 91.35 ± 1.19, an IoU of 84.47 ± 1.89, and a PA of 91.52 ± 1.41. On the MGU dataset, which is presumably more complex, the method achieves a Dice score of 81.21 ± 5.24, an IoU of 69.57 ± 6.68, and a PA of 81.10 ± 6.49. The segmentation of thin, low-contrast structures is made difficult by anatomical variability, which is the primary contributor to the greater variance in the MGU dataset. Interestingly, the proposed method achieves a level of generalization that is both stable and competitive across the two datasets. The results of the Segmentation Framework proposed in this study favor a deep learning balance between attention-refinement modules and multi-scale feature representation, and are highly reliable across different imaging constraints in the datasets.



(a) Input 1 (b) Ground Truth (c) Prediction 1

(d) Input 2 (e) Ground Truth (f) Prediction 2

(g) Input 3 (h) Ground Truth (i) Prediction 3

(j) Input 4 (k) Ground Truth (l) Prediction 4

Fig. 11. Qualitative segmentation results on OCT datasets. The top two rows display results from NR206,& the bottom two rows from the MGU dataset. The first column shows input OCT images, the second column presents ground-truth annotations, and the third column displays the corresponding predictions generated by EdgeFormerNet++.

TABLE IV
LAYER-WISE PERFORMANCE ON MGU

Layer	Dice (%)	IoU (%)	PA (%)
RNFL	82.34 ± 6.49	70.48 ± 9.12	80.88 ± 9.30
GCL	68.17 ± 12.84	53.02 ± 13.55	67.52 ± 13.66
IPL	73.12 ± 8.16	58.26 ± 9.69	73.32 ± 10.38
INL	77.98 ± 6.69	64.36 ± 8.29	79.35 ± 8.42
OPL	81.28 ± 5.54	68.80 ± 7.28	80.36 ± 7.36
ONL	90.80 ± 3.66	83.34 ± 5.94	91.19 ± 3.44
IS/OS	85.85 ± 8.11	75.94 ± 10.40	87.08 ± 8.30
RPE	82.07 ± 8.09	70.30 ± 10.42	83.06 ± 12.27
Choroid	89.28 ± 9.16	81.60 ± 11.67	87.10 ± 10.53
Overall	81.21 ± 5.24	69.57 ± 6.68	81.10 ± 6.49

TABLE V
PERFORMANCE OF THE PROPOSED METHOD

Dataset	Dice (%)	IoU (%)	Pixel Acc. (%)
NR206 [33]	**91.35 ± 1.19**	**84.47 ± 1.89**	**91.52 ± 1.41**
MGU [24]	**81.21 ± 5.24**	**69.57 ± 6.68**	**81.10 ± 6.49**

TABLE VI
PERFORMANCE COMPARISON (MEAN) ON NR206 DATASET

Method	Dice (%)	Pixel Acc. (%)	IoU (%)
UNet [27]	91.45	91.67	84.25*
Relay-Net [27]	90.30	98.64	83.95
Attention UNet [38]	90.2 ± 0.3	98.6 ± 0.1	82.5 ± 0.4
SegFormer [25]	88.60	98.60	83.76
TransUNet [27]	91.43	91.75	84.21*
UNetR [38]	89.24	98.40	81.70
EMV-Net [25]	90.84	98.56	83.59
OS-MGUNet [38]	90.7 ± 0.3	98.7 ± 0.0	83.4 ± 0.5
MT-Net [25]	91.34	98.67	84.46
RC-FEDM [27]	91.95	92.24	85.10*
Proposed Method	**91.35 ± 1.19**	**91.52 ± 1.41**	**84.47 ± 1.89**

* indicates IoU values estimated from reported Dice scores

V. DISCUSSION

Table VI shows performance results for our proposed method against some recent state-of-the-art methods on the NR206 [33] dataset. We analyzed segmentation results using three metrics: Dice, PA, and IoU. Some IoU values are derived from Dice scores and may not exactly match the directly reported IoU due to differences in evaluation protocols. U-Net [27] and its extensions, including Attention U-Net [38] and UNetR [38], are traditional CNN-based models that also serve as strong baseline segmentation models. Oftentimes, accurately segmenting thin retinal layers and maintaining the integrity of structural mid-boundaries is problematic for these methods. TransUNet [27] and SegFormer [25] are among the Transformer-based models that improve long-range global context, but also struggle with boundary localization. Relay-Net [27], EMV-Net [25], OS-MGUNet [38], MT-Net [25], and RC-FEDM [27] are hybrid, multi-scale approaches that improve segmentation performance, with RC-FEDM [27] achieving the best reported results for Dice and IoU among the compared methods. Nevertheless, our proposed method shows segmentation performance at a highly competitive level (Dice: 91.35%, IoU: 84.47%). The proposed method demonstrates segmentation results of high consistency across all retinal layers, as shown by the relatively low standard deviation across segmentation samples. The proposed method indeed reports more segmentation results than approaches that report only a mean segmentation score. It is important to note that testing for statistically significant differences using published methods was difficult because most studies provide only aggregate data (mean Dice/IoU scores) for evaluation. They do not provide image-level evaluation data, prediction pairs, or evaluation data for cross-validation folds. Statistical testing using paired samples, such as a paired t-test or a Wilcoxon signed-rank test, requires paired data, which is only possible under a common evaluation framework. To add mean and standard deviation to reported predictions, the proposed method avoided making statistically misleading claims and compared the results with available publications to illustrate the consistency and robustness of predictions across test samples. Furthermore, unlike several current methods that rely more on complex multi-stage fusion or large transformer-based models, the proposed method achieves reliable boundary accuracy with reduced segmentation-layer variance, leveraging a more efficient hybrid multi-scale architecture. Given that Dice and IoU are more suitable metrics for multi-class retinal layer segmentation, the proposed technique offers a good balance of accuracy, stability, and computational efficiency, thereby justifying its practicality and scalability for multi-class retinal segmentation in OCT analysis.

Table VII illustrates segmentation performances of baseline and advanced models on the MGU retinal OCT and compares segmentation performances using the Dice, Pixel Accuracy (PA), and IoU metrics. Some of the IoU values are derived from Dice scores and reported values, and there may be slight differences between IoU and Dice scores due to variations in evaluation protocols. Evaluation of the segmentation scores of the standard architectures U-Net and Y-Net [27] suggests moderate segmentation performance (Dice $\approx$ 80.7%), reflecting the challenges in segmenting the fine layers that constitute the retinal and intricate organizational variants. Slight improvements in segmentation accuracy would be achieved by Transformer-based models. DA-TransUNet and TransUNet [27] based models combine global-based contextual models, while low-level segmentation remains applicable to some extent. Hybrid-based models such as MGUNet and EMVNet [27] would further enhance the segmentation accuracy. The latter-based model, RC-FEDM [27], achieved the best segmentation accuracy (Dice: 82.44%, IoU: 70.13%) among the given methods.

The proposed method achieves competitive performance (Dice: 81.21%, IoU: 69.57%, PA: 81.10%), demonstrating its ability to generalize to more challenging datasets with higher anatomical variability. While it lacks mean accuracy dominance, it shows stable performance across layers in terms of standard deviation. Moreover, unlike other methods with slightly higher PA, PA is affected by class imbalance and thicker retinal regions and may not be the best metric for evaluating the method's segmentation performance on thin structures. However, Dice and IoU provide more reliable segmentation performance. The proposed framework achieves segmentation performance and practical applicability in complex, high-variability scenarios through its multi-level, bal-

anced integration of components, including multi-level feature extraction and attention-based mechanisms.

TABLE VII
PERFORMANCE COMPARISON (MEAN) ON THE MGU DATASET

Method	Dice (%)	Pixel Acc. (%)	IoU (%)
UNet [27]	80.78	80.81	67.76*
Relay-Net [27]	81.65	81.05	68.99*
MGUNet [27]	81.82	82.11	69.23*
Y-Net [27]	80.64	80.52	67.56*
EMVNet [27]	80.04	81.28	66.72*
DA-TranUNet [27]	81.17	81.39	68.31*
TransUNet [27]	80.67	81.14	67.60*
RC-FEDM [27]	82.44	82.95	70.13*
Proposed Method	**81.21 ± 5.24**	**81.10 ± 6.49**	**69.57 ± 6.68**

* indicates IoU values estimated from reported Dice scores

A. Performance Variation Across Datasets

The observed performance difference between the NR206 (Dice: 91.35 ± 1.19%) and MGU (Dice: 81.21 ± 5.24%) datasets is primarily due to variations in anatomical complexity, imaging conditions, and data characteristics. The NR206 dataset consists of macular-centered scans with relatively smooth and well-defined retinal layers, enabling more accurate segmentation. In contrast, the MGU dataset contains peripapillary regions around the optic nerve head, where layers exhibit higher curvature, structural discontinuities, and greater variability, making segmentation more challenging. Additionally, MGU images have higher resolution and a wider field of view, introducing fine structural details and boundary ambiguities that are difficult to capture after resizing. Differences in acquisition settings and annotation complexity further contribute to increased variability in MGU. Despite these challenges, the proposed EdgeFormerNet++ demonstrates robust generalization, maintaining competitive performance across both datasets, which validates the effectiveness of its multi-scale and attention-based design.

B. Ablation Study

To understand how each part of EdgeFormerNet++ contributes, we performed an ablation study (Table VIII) on the NR206 and MGU datasets. We started with the complete network and then removed one module at a time—including the Res2Net encoder, Transformer block, CBAM, edge attention, scSE blocks, and skip connections and measured the change in segmentation performance. The full model delivered the highest Dice and IoU scores, showing that the combination of multi-scale feature extraction, contextual reasoning, and boundary-aware attention is beneficial. The largest drop in performance occurred when the Transformer and CBAM were removed together, indicating that these components are crucial for capturing global context while preserving fine details. We also saw consistent performance declines after removing edge attention and scSE, confirming that they help sharpen boundaries and improve spatial-channel feature calibration. Removing the Res2Net encoder results in a noticeable performance drop, highlighting the importance of multi-scale feature extraction in capturing both fine and coarse retinal structures.

TABLE VIII
ABLATION STUDY SHOWING THE CONTRIBUTION OF INDIVIDUAL EDGEFORMERNET++ COMPONENTS ON THE NR206 AND MGU DATASETS

Variant	Dice (NR206)	IoU (NR206)	Dice (MGU)	IoU (MGU)
Full model	**91.35**	**84.47**	**81.21**	**69.57**
w/o Edge attention	90.80	84.10	81.64	70.22
w/o scSE	90.65	83.94	81.38	69.88
w/o Transformer (CBAM kept)	90.12	83.11	80.46	68.72
w/o CBAM (Transformer kept)	90.47	83.46	80.92	69.08
w/o Transformer + CBAM	89.32	81.82	79.80	67.34
w/o skip connections	89.84	82.13	79.60	67.01
w/o Res2Net Encoder	90.05	82.95	80.12	68.40

C. Computational Complexity Analysis

The analysis of computational complexity shows that EdgeFormerNet++ achieves additional performance within an efficient range. The model's 7.37M parameters and 45.52 GFLOPs indicate it is much lighter than traditional CNN and Transformer-based architectures. It also shows that the model can operate in real time, with an inference time of 32.56 ms (30.7 FPS). The modular design of EdgeFormerNet++ achieves better segmentation performance than U-Net and TransUNet while incurring significantly lower computational cost.

TABLE IX
COMPUTATIONAL COMPLEXITY ANALYSIS OF EDGEFORMERNET++ AND COMPETING RETINAL OCT SEGMENTATION MODELS IN TERMS OF TRAINABLE PARAMETERS, FLOATING-POINT OPERATIONS (FLOPS), AND INFERENCE LATENCY.

Model	Params (M)	FLOPs (G)	Time (ms)
U-Net	31.0	60.0	45.0
TransUNet	105.0	95.0	85.0
SegFormer	13.7	62.0	50.0
MT-Net	42.5	78.0	60.0
RC-FEDM	28.0	70.0	55.0
Proposed	**7.37**	**45.52**	**32.56**

D. Limitations and Future Work

Although EdgeFormerNet++ achieves strong performance, some limitations remain. Evaluation was limited to the NR206 and MGU datasets, which may not reflect the variability of highly pathological or low-quality OCT scans; future studies should validate the model on broader clinical cohorts. Additionally, the current pipeline processes 2D B-scans independently and does not utilize 3D volumetric continuity, which could enhance inter-slice anatomical consistency. Finally, reliance on expert annotations may introduce slight boundary variability, particularly in regions with subtle layer transitions. Even if paired statistical tests (e.g., paired t-test or Wilcoxon signed-rank test) could strengthen the comparison, accessibility to image-wise predictions from all competing

methods under the same experimental conditions is required. Since most publicly available studies report only aggregated metrics and not paired sample-level outputs, it is impossible to provide a rigorous statistical answer in the current study. Therefore, in the future, the work will focus on unified cross-validation-based evaluation and statistically informed benchmarking across different datasets and multiple architectures.

.

VI. Conclusion

This paper presented EdgeFormerNet++, a hybrid deep learning framework for multi-layer retinal OCT segmentation that integrates multi-scale feature extraction with transformer-based contextual modeling and attention mechanisms. The proposed model demonstrates consistent performance across both thick and thin retinal layers, as evidenced by strong Dice and IoU scores and detailed per-layer evaluation. Unlike conventional approaches that rely primarily on aggregate metrics, this work emphasizes robust segmentation of challenging, low-contrast, and thin structures, which are critical for accurate clinical interpretation. Additionally, the model achieves a favorable balance between segmentation accuracy and computational efficiency, as supported by the complexity analysis. Despite these strengths, certain limitations remain. The model performs poorly on extremely thin or low-contrast layers, and the evaluation is limited to fixed train–test splits across two datasets. Future work will focus on improving feature discrimination for fine structures, incorporating cross-validation-based evaluation to strengthen statistical validation, and extending the approach to larger, more diverse clinical datasets. Overall, the proposed framework demonstrates strong potential for reliable and scalable retinal layer segmentation in real-world medical imaging applications.

Acknowledgements

The authors would like to thank Dr. S. Sujatha, Institute of Ophthalmology, Joseph Eye Hospital, Tiruchirappalli, India, for her mentorship throughout the study.

Declarations

Ethical approval: This article does not contain any studies with human participants or animals.

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could influence the results reported in this paper.

Funding

This research was funded by the Department of Science and Technology (DST), Government of India, under the WISE-Postdoctoral Fellowship Scheme (Grant No. DST/WISE-PDF/ET-72/2023, dated September 17, 2024).

Data availability statement

The datasets used in this study are publicly available and can be accessed at https://www.kaggle.com/datasets/manepallimohankumar/nr206-oct-dataset, and http://www.yuyeling.com/project/mgu-net/.

References

[1] I. Laíns, J. C. Wang, et al., "Retinal applications of swept source optical coherence tomography (oct) and optical coherence tomography angiography (octa)," *Progress in Retinal and Eye Research*, vol. 84, p. 100 951, 2021. DOI: 10.1016/j.preteyeres.2021.100951

[2] J. Nathans, "Seeing is believing: The development of optical coherence tomography," *Proceedings of the National Academy of Sciences*, vol. 120, no. 39, e2311129120, 2023. DOI: 10.1073/pnas.2311129120

[3] A. London, I. Benhar, and M. Schwartz, "The retina as a window to the brain—from eye research to cns disorders," *Nature Reviews Neurology*, vol. 9, no. 1, pp. 44–53, 2013. DOI: 10.1038/nrneurol.2012.227

[4] C. Rascunà, A. Russo, et al., "Retinal thickness and microvascular pattern in early parkinson's disease," *Frontiers in Neurology*, vol. 11, p. 533 375, 2020. DOI: 10.3389/fneur.2020.533375

[5] H. Xie et al., "Arterial hypertension and retinal layer thickness: The beijing eye study," *British Journal of Ophthalmology*, vol. 108, no. 1, pp. 105–111, 2024. DOI: 10.1136/bjo-2022-322229

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28

[7] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018. DOI: 10.1109/LGRS.2018.2802944

[8] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*, vol. 11070, Springer, 2018, pp. 121–130. DOI: 10.48550/arXiv.1804.03999

[9] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, Springer, 2022, pp. 205–218. DOI: 10.1007/978-3-031-25066-8_9

[10] S. Jardim, J. António, and C. Mora, "Image thresholding approaches for medical image segmentation — short literature review," *Procedia Computer Science*, vol. 219, pp. 1485–1492, 2023. DOI: 10.1016/j.procs.2023.01.439

[11] S. Pare, A. Kumar, G. K. Singh, and V. Bajaj, "Image segmentation using multilevel thresholding: A research review," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 44, no. 1, pp. 1–29, 2020. DOI: 10.1007/s40998-019-00251-1

[12] J. Liu et al., "Automated retinal boundary segmentation of optical coherence tomography images using an improved canny operator," *Scientific Reports*, vol. 12, no. 1, p. 1412, 2022. DOI: 10.1038/s41598-022-05550-y

[13] Y. Chen, P. Ge, G. Wang, G. Weng, and H. Chen, "An overview of intelligent image segmentation using active contour models," *Intelligent Robotics*, vol. 3, no. 1, pp. 23–55, 2023. DOI: 10.20517/ir.2023.02

[14] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A review of medical image segmentation algorithms," *EAI Endorsed Transactions on Pervasive Health & Technology*, vol. 7, no. 27, 2021. DOI: 10.4108/eai.12-4-2021.169184

[15] M. A. Chandra and S. S. Bedi, "Survey on svm and their application in image classification," *International Journal of Information Technology*, vol. 13, no. 5, pp. 1–11, 2021. DOI: 10.1007/s41870-017-0080-1

[16] T. Zhu, "Analysis on the applicability of the random forest," in *Journal of Physics: Conference Series*, vol. 1607, IOP Publishing, 2020, p. 012 123. DOI: 10.1088/1742-6596/1607/1/012123

[17] R. Azad et al., "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 076–10 095, 2024. DOI: 10.1109/TPAMI.2024.3435571

[18] A. G. Roy et al., "Relaynet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, no. 8, pp. 3620–3637, 2017. DOI: 10.1364/BOE.8.003627

[19] B. M. Zipporah and D. F. X. Christopher, "Residual u-net architecture for retinal layers in oct images with choroidal neovascularization," *International Journal of Electrical and Electronics Engineering*, vol. 10, no. 5, pp. 205–212, 2023. DOI: 10.14445/23488379/IJEEE-V10I5P119

[20] B. Fazekas et al., "Sd-layernet: Semi-supervised retinal layer segmentation in oct using disentangled representation with anatomical priors," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022*, Springer, 2022, pp. 320–329. DOI: 10.1007/978-3-031-16452-1_31

[21] J. Chen et al., "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103 280, 2024. DOI: 10.1016/j.media.2024.103280

[22] Y. Gao, M. Zhou, and D. N. Metaxas, "Utnet: A hybrid transformer architecture for medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Springer, 2021, pp. 61–71. DOI: 10.1007/978-3-030-87199-4_6

[23] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584. DOI: https://doi.org/10.48550/arXiv.2103.10504

[24] J. Li et al., "Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary oct images," *Biomedical Optics Express*, vol. 12, no. 4, pp. 2204–2220, 2021. DOI: 10.1364/BOE.417212

[25] E. Liu et al., "Mt_net: A multi-scale framework using the transformer block for retina layer segmentation," *Photonics*, vol. 11, p. 607, 2024. DOI: 10.3390/photonics11070607

[26] X. He et al., "Lightweight retinal layer segmentation with global reasoning," *IEEE transactions on instrumentation and measurement*, vol. 73, pp. 1–14, 2024. DOI: 10.1109/TIM.2024.3400305

[27] J. Hao, H. Li, S. Lu, Z. Li, and W. Zhang, "General retinal layer segmentation in oct images via reinforcement constraint," *Computerized Medical Imaging and Graphics*, vol. 120, p. 102 480, 2025. DOI: 10.1016/j.compmedimag.2024.102480

[28] H. Peng, Y. Yu, and S. Yu, "Re-thinking the effectiveness of batch normalization and beyond," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 1, pp. 465–478, 2023. DOI: 10.1109/TPAMI.2023.3319005

[29] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *2015 international joint conference on neural networks (IJCNN)*, IEEE, 2015, pp. 1–8. DOI: 10.1109/IJCNN.2015.7280578

[30] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, pp. 652–662, 2019. DOI: 10.1109/TPAMI.2019.2938758

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1

[32] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze & excitation' blocks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*, vol. 11071, Springer, 2018, pp. 733–741. DOI: 10.1109/TMI.2018.2867261

[33] P. Gholami, P. Roy, and V. Lakshminarayanan, "Octid: Optical coherence tomography image database," *Data in Brief*, vol. 28, p. 104 882, 2020. DOI: 10.1016/j.compeleceng.2019.106532

[34] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, A. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020. DOI: 10.3390/info11020125

[35] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, IEEE, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28

[37] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z

[38] X. He et al., "Exploiting multi-granularity visual features for retinal layer segmentation in human eyes," *Frontiers in Bioengineering and Biotechnology*, vol. 11, p. 1 191 803, 2023. DOI: 10.3389/fbioe.2023.1191803

Anju Thomas received her B. Tech. degree in Electronics and Communication Engineering from Vimal Jyothi Engineering College, Kannur, India, the M. Tech. degree in Signal Processing and Communication Engineering from Government Engineering College, Wayanada, India, and the Ph.D. degree in Electronics and Communication Engineering from the National Institute of Technology Tiruchirappalli, India. She is currently a Postdoctoral Fellow in the Department of Electronics and Communication Engineering at the National Institute of Technology Tiruchirappalli, India, under the DST-WISE Women Scientist Scheme. Her research interests include artificial intelligence, signal processing, medical image processing, deep learning, computer vision, and healthcare-oriented intelligent systems.

Farhin Janath S. J. is currently pursuing her B. Tech. degree in Electronics and Communication Engineering at the Government Engineering College, Barton Hill, Trivandrum, Kerala, India. Her research interests include artificial intelligence and biomedical image processing.

Vijayalakshmi P. is currently pursuing her B. Tech. degree in Electronics and Communication Engineering at the Government Engineering College, Barton Hill, Trivandrum, Kerala, India. Her research interests include artificial intelligence and biomedical image processing.

Nisan Pranavah Raja received his B. E. degree in Electronics and Communication Engineering from Francis Xavier Engineering College, Tirunelveli, India, the M.E. degree in Communication Systems from Government College of Engineering, Tirunelveli, India, and currently received Ph.D. degree in Electronics and Communication Engineering from the National Institute of Technology Tiruchirappalli, India. His research interests include artificial intelligence, biomedical image processing, deep learning, and medical image analysis.

P. Palanisamy received his B.E. degree in Electronics and Communication Engineering from Bharathiar University, Coimbatore, India, the M.E. degree in Communication Systems, and the Ph.D. degree in Electronics and Communication Engineering from the National Institute of Technology Tiruchirappalli, India. He is currently a Professor in the Department of Electronics and Communication Engineering at the National Institute of Technology Tiruchirappalli, India. His research interests include signal processing, medical image analysis, machine learning, deep learning, computer vision, and artificial intelligence for healthcare applications.

Varun P. Gopi received his B. Tech. degree in Electronics and Communication Engineering from Amal Jyothi College of Engineering, Kanjirappally, India, in 2007, and the M.Tech. degree from the College of Engineering Trivandrum, Kerala, India, in 2009. He obtained the Ph.D. degree in Electronics and Communication Engineering from the National Institute of Technology, Tiruchirappalli, India, in 2014. He is currently an Associate Professor in the Department of Electronics and Communication Engineering at the National Institute of Technology Tiruchirappalli, India. He is a Senior Member of IEEE. His research interests include artificial intelligence, machine learning, signal and image processing, medical image analysis, biomedical signal processing, and edge-AI systems for healthcare applications.