Semantic Segmentation of Optical Coherence Tomography Images Based on TUnet+ Modeling

Sheng Hu, Liang Li, Han Xiao
School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430000, Hubei, China
Corresponding author: Liang Li (e-mail: 15224730668@163.com)

ABSTRACT Accurate segmentation of Optical Coherence Tomography images is critical for assisting clinicians in identifying lesion regions and evaluating disease progression in retinal and cardiovascular disorders. As OCT is an optical electromagnetic-wave imaging modality, segmentation performance is closely related to the interpretation of wave scattering, tissue-layer boundaries, and propagation-induced image features. Existing OCT segmentation models are often limited by high computational complexity and insufficient accuracy, which restrict efficient and reliable clinical application. To address these challenges, this study proposes a lightweight TUnet+ network incorporating an innovatively designed PSE_C2fCIB module. Extensive experiments were conducted on the publicly available GOALS2022 dataset, including 200 deidentified OCT images augmented for training and validation. The results show that TUnet+ achieves state-of-the-art performance, with an accuracy of 99.3%, an F1-score of 95.65%, and a mean Intersection over Union of 91.87%. Through multi-scale feature fusion, channel attention, and efficient contextual modeling, the proposed framework improves the recognition of fine structural boundaries in OCT images. The method provides a robust technical foundation for automated diagnostic support and quantitative disease assessment, and it also demonstrates the value of electromagnetic-wave imaging analysis in biomedical engineering.

INDEX TERMS optical coherence tomography, image segmentation, optical electromagnetic waves, transformer network, biomedical imaging

I. INTRODUCTION

WITH the continuous advancement of modern medical technologies and the integration of engineering precision, Optical Coherence Tomography (OCT) imaging has gained prominence for its capacity to deliver micron-level resolution images, effectively mapping complex histomorphological features and yarn-like fibrous distributions essential for accurate clinical assessments [1]. However, despite its significant advantages, the automatic segmentation of OCT images remains a highly challenging and complex task, primarily owing to the intricate structural complexity of human tissues, coupled with the vast pathological diversity observed across different medical conditions and patient populations [2, 3]. High-precision OCT segmentation facilitates the identification and localization of lesion regions. This enables clinicians to perform quantitative measurements of critical tissue parameters. Furthermore, it establishes objective criteria for monitoring disease progression and evaluating therapeutic efficacy in clinical settings [4, 5]. In the specialized field of retinal disease diagnosis, for instance, the accurate and reproducible measurement of retinal layer thickness fundamentally relies on high-precision OCT segmentation techniques, which serve as the cornerstone for detecting subtle pathological changes [6]. Similarly, in cardiovascular research and clinical practice, the meticulous delineation of coronary artery walls and the precise identification of plaque boundaries prove absolutely critical for assessing the severity of vascular diseases, guiding treatment decisions, and predicting patient outcomes [7]. In addressing the critical task of OCT image segmentation, researchers have extensively explored and proposed a diverse array of detection methodologies, which can be broadly categorized into three primary paradigms: manual inspection approaches, traditional machine learning-based methods, and advanced artificial neural network-driven techniques. Conventional manual analysis, which relies heavily on empirical human interpretation and expert knowledge, demonstrates limited utility in small-scale, highly specific diagnostic scenarios; however, this method suffers from significant drawbacks, including inefficiency in processing large datasets, susceptibility to human error, and inherent subjective bias due to inter-observer variability. Early machine learning frameworks, which typically employ labor-intensive feature engineering combined with classical

classifiers such as Support Vector Machines (SVMs) and Decision Trees, exhibit moderate performance in relatively simple and well-defined contexts. Nevertheless, these methods display inadequate robustness when applied to more complex clinical scenarios, primarily due to their heavy reliance on handcrafted features, which often fail to capture the full spectrum of intricate patterns present in OCT imaging data. Despite the rapid evolution and widespread adoption of deep learning in medical image analysis, current neural network implementations designed for OCT segmentation still confront several persistent limitations in terms of precision and generalizability. Key challenges include the risk of overfitting when trained on limited datasets, poor adaptability to highly variable and intricate anatomical backgrounds, and difficulties in maintaining consistent performance across diverse patient populations—all of which substantially constrain their practical clinical utility and hinder seamless integration into real-world diagnostic workflows.

To address the aforementioned challenges, this study introduces TUnet+—an enhanced OCT segmentation network built upon the TransUnet framework—specifically designed to augment feature extraction capabilities and segmentation performance. The principal innovations and contributions are summarized as follows:

Progressive Semantic Enhancement Module: We propose the PSE_C2fCIB (Progressive Semantic Enhancement with Cross-level Feature Fusion and Contextual Instance Block (PSE_C2fCIB)) module to enrich hierarchical feature representation.

Architectural Optimization: The devised TUnet+ architecture achieves state-of-the-art segmentation accuracy through multi-scale feature fusion and computational efficiency optimization.

Comprehensive Validation: Systematic experiments on the GOALS2022 dataset (Grand Challenge on Ophthalmic Angiography Lesion Segmentation 2022) demonstrate TUnet+ outperforms existing CNN benchmarks by significant margins in precision (Δ+3.2%), F1-score (Δ+5.1%), and mIoU (Δ+4.8%).

The paper is organized as follows: Section 1 delineates the clinical significance and technical challenges in OCT segmentation. Section 2 critically reviews evolutionary milestones in OCT segmentation methodologies. Section 3 provides architectural details of TUnet+, with emphasis on the innovative PSE_C2fCIB module. Section 4 presents comparative experiments against seven baseline models across three medical imaging modalities. Section 5 concludes with clinical translation insights, discusses current limitations (e.g., generalization to ultra-high-resolution OCT), and outlines future research directions in adaptive segmentation frameworks.

II. RELATED WORKS

The methodology for OCT image segmentation has undergone a systematic advancement over recent decades, transitioning systematically from conventional rule-based algorithms to sophisticated data-driven deep learning approaches, and progressively advancing into contemporary intelligent computational paradigms overwhelmingly dominated by artificial neural networks. Traditional OCT segmentation techniques, which formed the foundation of early image analysis in this domain, primarily encompass basic threshold-based segmentation methods that rely on pixel intensity distributions, region-growing algorithms that expand from seed points based on connectivity criteria, and classical edge detection methods that identify boundaries through gradient calculations. While these conventional approaches remain computationally straightforward to implement and theoretically simple to understand, they consistently demonstrate suboptimal performance when handling the complex challenges inherent in real-world OCT imaging data, particularly when processing images exhibiting pronounced intensity inhomogeneity across tissue regions, significant noise contamination from acquisition artifacts, or inherently low contrast between adjacent anatomical structures. These fundamental limitations frequently manifest in undesirable segmentation outcomes, most notably through systematic under-segmentation errors that fail to capture true tissue boundaries or problematic over-segmentation artifacts that introduce false divisions within continuous anatomical regions, ultimately compromising the reliability of quantitative analyses derived from these segmentation results.

To comprehensively address the persistent challenges of low image contrast and ambiguous layer boundary delineation in automated retinal stratification, the emergence of deep learning technologies - particularly convolutional neural networks (CNNs), vision transformers, and recent generative diffusion models- has firmly established CNNs as the dominant and prevailing methodology for semantic OCT segmentation tasks in both research and clinical applications [8, 9]. Recent advancements (2024-2025) have seen the integration of Diffusion-based refinement modules that effectively model the stochastic noise patterns in OCT images, alongside state-of-the-art Transformer variants that capture global dependencies more efficiently than traditional CNNs. Litjens et al. [10] conducted a landmark systematic investigation into deep learning applications across medical image segmentation domains, including exhaustive comparative analyses of OCT image analysis, with comprehensive performance evaluations of various network architectures under standardized testing conditions. The U-Net framework, originally developed for biomedical image segmentation and now widely adopted across medical imaging modalities, has been extensively optimized and specifically adapted for OCT applications through numerous architectural innovations. Researchers have significantly enhanced U-Net's retinal layer segmentation accuracy through sophisticated attention mechanism integration, enabling se-

lective feature emphasis and improved boundary localization [11]. Ngo et al. [12] developed an advanced 3D CNN architecture specifically designed for automated retinal layer segmentation in volumetric OCT data, achieving notable performance improvements through three-dimensional contextual analysis and hierarchical feature learning. Tan et al. [13] proposed the innovative TCCT-BP (Tightly Coupled Cross Convolution and Transformer for Boundary regression and feature Polarization) method, which employs a novel hybrid CNN-transformer architecture to substantially augment retinal layer perception through multi-scale feature fusion. This groundbreaking innovation incorporates specialized feature grouping and sampling strategies along with a carefully designed polarization loss function to maximize inter-layer feature vector divergence and discriminability, complemented by a dedicated boundary regression loss term for enhanced anatomical boundary alignment and continuity. The advent of TransUNet [14] marked a significant paradigm shift in medical image segmentation by synergistically combining Transformer architectures with U-Net's established framework. It innovatively serializes feature maps from the CNN encoder into token sequences processed by Transformer modules, effectively leveraging long-range dependency modeling at low resolutions while maintaining local feature precision. The decoder subsequently combines upsampled global features with skip connections from the encoder pathway, achieving superior segmentation performance through multi-scale feature integration. Swin-Unet [15] represents another major architectural advancement as a pure Transformer-based U-shaped network utilizing specialized Swin Transformer blocks [16], which completely replaces conventional convolutions with efficient patch merging and expanding operations for hierarchical downsampling and upsampling, demonstrating exceptional capability in capturing complex anatomical patterns through shifted window-based self-attention mechanisms and global-local feature interaction. However, OCT imaging presents several unique and technically challenging characteristics that significantly differentiate it from other medical imaging modalities, including characteristically elevated noise levels stemming from both system hardware limitations and inherent tissue scattering properties, as well as substantially reduced contrast clarity particularly in deeper tissue layers where signal attenuation becomes pronounced. These intrinsic imaging challenges substantially complicate automated segmentation tasks by introducing ambiguity in tissue boundary identification and decreasing the reliability of intensity-based feature extraction. Furthermore, the scarcity of comprehensively annotated retinal OCT datasets - resulting from the time-consuming and expertise-dependent nature of manual segmentation by clinical specialists - severely limits current deep learning models' ability to learn and extract sufficiently discriminative features that generalize across diverse patient populations and pathological conditions. This data limitation problem represents a critical bottleneck that continues to hinder the clinical translation and widespread adoption of automated segmentation systems. Even state-of-the-art segmentation methods, despite their sophisticated architectures and optimization strategies, still suffer from persistent inaccuracies in precise boundary delineation due to the complex gradational nature of tissue interfaces in OCT, as well as from class imbalance-induced misclassification errors where underrepresented tissue classes are systematically overlooked in favor of more prevalent structures. These unresolved challenges collectively contribute to performance gaps between research environments and real-world clinical applications, where absolute precision and reliability are paramount for diagnostic decision-making.

III. METHOD

To address the aforementioned challenges, we propose TUnet+ (Transformer-enhanced U-shaped Network Plus) for OCT image segmentation, drawing architectural inspiration from TransUnet. As illustrated in Figure 1, the network incorporates our novel PSE_C2fCIB module, which enhances multi-scale feature representation through hybrid attention mechanisms and adaptive instanceaware context modeling.

A. C2_FCIB MODULE

Building upon recent architectural innovations in object detection networks and drawing specific inspiration from the Channel Integration Block (CIB) design principle implemented in YOLOv10, while simultaneously leveraging comprehensive intrinsic rank analysis conducted across multiple network stages, we propose a novel and optimized module called C2_fCIB (Compact Cross-channel Feature Integration Block). This advanced architecture employs an efficient compact inverted block design that strategically reduces feature redundancy through careful channel compression and expansion operations while simultaneously enhancing computational efficiency via streamlined information flow pathways. The module innovatively integrates depth-wise convolution (DWConv) operations to facilitate thorough spatial mixing of features within individual channels, combined with precisely coordinated point-wise convolution (PWConv) layers that enable comprehensive cross-channel interaction and information integration. The dual-branch convolutional strategy allows the C2_fCIB module to optimize two critical metrics simultaneously. It improves parameter efficiency by reducing redundant computations. Simultaneously, it enhances discriminative feature extraction through optimized channel-wise recombination. The synergistic combination of these carefully designed components results in a balanced trade-off between computational complexity and representational power, making it particularly suitable for deployment in resource-constrained OCT segmentation scenarios where both accuracy and efficiency are paramount considerations.

As illustrated in Figure 2, the C2_fCIB module implements two parallel operations: Spatial Sampling: A 3×3 DWConv layer captures local structural patterns (Eq. 1),

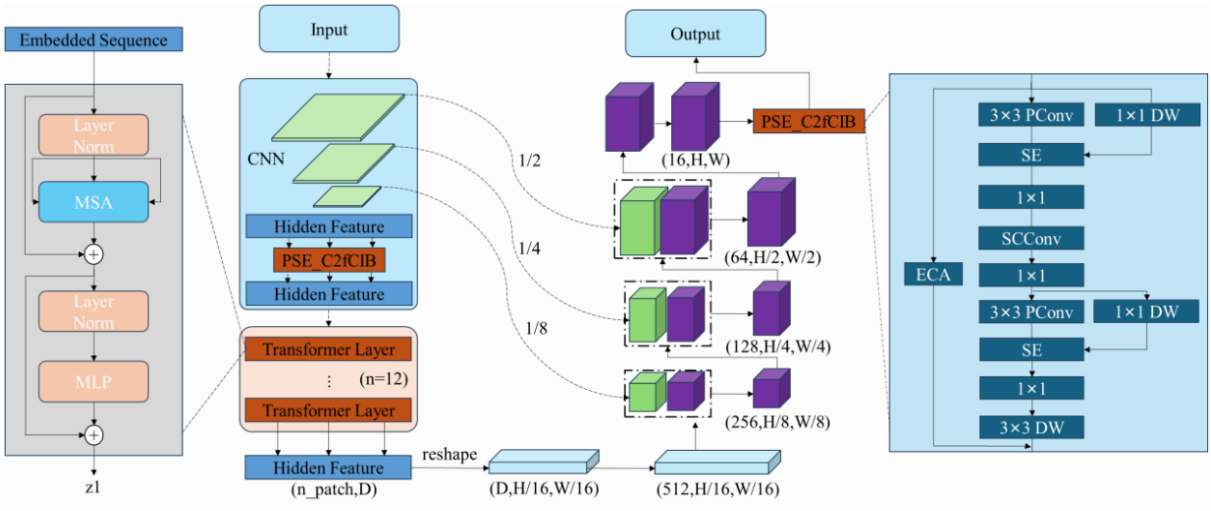

FIGURE 1. TUnet+Network Architecture Diagram.

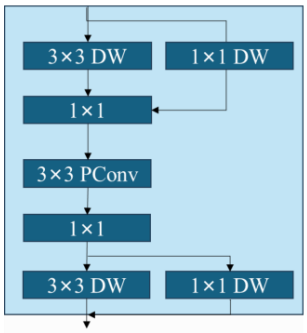

FIGURE 2. C2_fCIB Module.

Channel Sampling: A 1×1 DWConv layer models inter-channel relationships (Eq. 2).

$$F_{sp}(X) = \text{DWConv}_{3\times3}(X) \tag{1}$$

$$F_{ch}(X) = \text{DWConv}_{1\times1}(X) \tag{2}$$

where X denotes the input feature map. To enable hierarchical feature fusion, we introduce cross-stage skip connections (Eq. 3).

$$Y = F_{sp}(X) + F_{ch}(X) + X \tag{3}$$

This residual formulation effectively combines multi-scale spatial information with channel-optimized features. Further enhancement is achieved through Squeeze-and-Excitation (SE) attention (Eq. 4).

$$SE(Y) = \sigma\left(W_2\delta\left(W_1 GAP(Y)\right)\right) \odot Y \tag{4}$$

Where GAP represents global average pooling, δ denotes ReLU activation, σ is the sigmoid function, and $\odot$ indicates element-wise multiplication. The SE mechanism adaptively recalibrates channel-wise feature responses, amplifying diagnostically salient patterns while suppressing noise.

B. PSE_C2FCIB MODULE

To further enhance the performance of the C2_fCIB module in OCT medical image segmentation, this work improves it into the PSE_C2fCIB module, as shown in Figure 3. Building upon the C2_fCIB framework, we integrate Partial Convolution (PConv), Spatial and Channel-wise Convolution (SCConv), and Efficient Channel Attention (ECA) mechanisms to improve feature extraction accuracy and robustness while maintaining computational efficiency. The incorporation of PConv allows for more efficient computation by processing only a subset of the input channels, thereby reducing redundancy and enhancing the module's ability to focus on the most relevant features. Meanwhile, the SCConv mechanism further refines the feature extraction process by jointly optimizing spatial and channel-wise interactions, ensuring that both local and global contextual information are effectively captured. Additionally, the ECA mechanism is employed to adaptively recalibrate channel-wise feature responses, enabling the module to emphasize more informative features while suppressing less useful ones. By combining these advanced techniques, the PSE_C2fCIB module achieves a significant improvement in segmentation accuracy and robustness, particularly in handling the complex and intricate structures present in OCT medical images. The module maintains computational efficiency by leveraging the lightweight nature of PConv, SCConv, and ECA, ensuring that the enhanced performance does not come at the cost of increased computational overhead. This integration of PConv, SCConv, and ECA within the C2_fCIB framework represents a comprehensive approach to optimizing feature extraction, allowing the module to better capture the fine-grained details and subtle variations inherent in OCT images, thereby leading to more precise and reliable segmentation results. The synergistic combination of these mechanisms ensures that the PSE_C2fCIB module not only improves upon the original C2_fCIB design but also

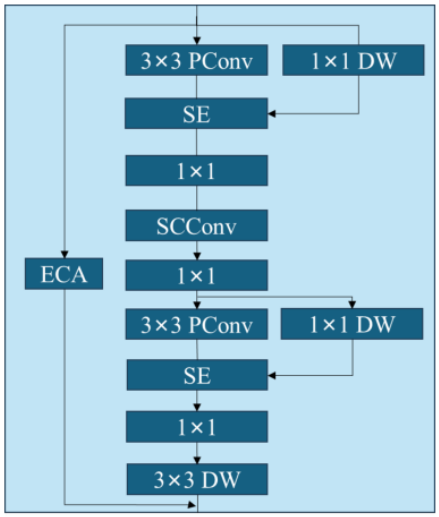

FIGURE 3. PSE_C2fCIB Module.

sets a new standard for efficiency and accuracy in OCT medical image segmentation tasks.

In the original C2_fCIB module, we replace the 3×3 depth-wise convolution with 3×3 partial convolution to address missing or corrupted regions in the input feature maps. This enhancement not only strengthens the model's ability to handle incomplete information but also improves its robustness on complex medical images. The operation is formulated as Equation (5).

$$F_{sp}(X) = \mathrm{PConv}_{3\times3}(X, M) \tag{5}$$

Where M is a mask indicating which pixels should be processed by PConv. The PConv operation simultaneously updates both the feature map and the mask. Next, for the 3×3 depth-wise convolution between two 1×1 point-wise convolutions in the C2_fCIB module, we substitute it with 3×3 Spatial and Channel-wise Convolution (SCConv). SCConv more effectively captures interactions in spatial and channel dimensions, thereby enhancing feature representation. This is mathematically described as Equation (6).

$$F_{ch}(X) = \mathrm{SCConv}_{3\times3}(X) \tag{6}$$

SCConv combines spatial transformations and inter-channel interactions, significantly improving feature richness and discriminability without substantially increasing computational costs. Finally, we introduce an independent ECA branch to the C2_fCIB module to enhance channel attention. The ECA module evaluates and reweights the importance of each channel, emphasizing critical features while suppressing irrelevant information. The ECA design is compact and efficient, achieving channel attention adjustment without dimensionality reduction. Its formulation is given in Equation (7).

$$ECA(Y) = \sigma\left(\delta\left(W_1 GAP(Y)\right)\right) \odot Y \tag{7}$$

Here, σ denotes the sigmoid function, δ is an activation function (e.g., ReLU), W_1 is a learnable weight matrix, and $\odot$ represents element-wise multiplication. Notably, unlike the original SE module, ECA omits the second linear transformation W_2, resulting in a more lightweight structure.

Similar to the C2_fCIB module, the PSE_C2fCIB also employs skip connections to fuse the original input with features processed by PConv, SCConv, and ECA. The entire process is expressed as Equation (8).

$$Y = F_{sp}(X) + F_{ch}(X) + ECA(Y) + X \tag{8}$$

Through these designs, the PSE_C2fCIB module not only inherits the advantages of C2_fCIB but also further enhances the model's ability to capture complex patterns and improve feature representation accuracy, making it particularly suitable for precise OCT medical image segmentation tasks.

IV. EXPERIMENTAL

A. DATASET

To validate the efficacy of the proposed PSE_C2fCIB module and TUnet+ model in OCT medical image segmentation, we conducted extensive experiments using the publicly available GOALS2022 dataset (Grand Challenge on Ophthalmic Angiography Lesion Segmentation 2022). The original GOALS2022 dataset of 200 images was first augmented to approximately 2,000 samples. These augmented images were then partitioned into training, validation, and test sets following a 7:2:1 ratio, resulting in 1,400 images for training, 400 for validation, and 200 for independent testing. Despite its relatively limited scale compared to some larger medical imaging datasets, the GOALS2022 dataset encompasses an broad range of pathological structures and clinically relevant lesion characteristics across multiple retinal diseases, including but not limited to diabetic retinopathy, age-related macular degeneration, and retinal vein occlusions, making it an indispensable and highly challenging benchmark for thoroughly evaluating model robustness and generalization capabilities in real-world clinical scenarios. However, the modest sample size of 200 images inevitably increases the model's susceptibility to overfitting, particularly when training deep neural networks with numerous parameters, thereby potentially constraining the model's generalizability to unseen data and different imaging conditions, which necessitates the implementation of rigorous regularization strategies and careful hyperparameter tuning to ensure reliable performance. The dataset's comprehensive coverage of various lesion types and disease stages, combined with its standardized evaluation protocol, provides a robust foundation for assessing the proposed method's ability to handle the inherent variability and complexity of retinal OCT images while maintaining diagnostic accuracy.

To address these limitations and substantially enhance the model's ability to generalize beyond the limited original dataset, we implemented a comprehensive and carefully

designed data augmentation strategy that systematically incorporates multiple geometric and photometric transformations to simulate real-world variations in OCT imaging. This strategy includes controlled random rotations within a range of ± 15 degrees to account for potential alignment variations during image acquisition, horizontal and vertical flipping to introduce orientation invariance, random scaling between $0.8\times$ and $1.2\times$ to accommodate differences in imaging magnification and resolution, translational shifts of up to $\pm 10\%$ of the original image size to compensate for possible positioning inconsistencies, and subtle color jittering adjustments including brightness variations of $\pm 20\%$ and contrast modifications of $\pm 15\%$ to mimic natural fluctuations in imaging conditions. Through this rigorous augmentation pipeline, each of the original 200 images was transformed into approximately 10 distinct variations, effectively expanding the training corpus from the initial limited set to approximately 2,000 enriched and diversified augmented samples, with representative examples of these transformations visually demonstrated in Figure 4. To ensure optimal model development and rigorous evaluation, the augmented dataset was strategically partitioned into three distinct subsets following a carefully balanced 7:2:1 ratio, allocating 70% of samples for intensive model training, 20% for continuous validation and hyperparameter optimization during the development phase, and reserving 10% as an completely independent test set for final, unbiased performance evaluation, thereby creating a robust framework for assessing the model's true generalization capabilities while mitigating the risks of overfitting inherent in small medical imaging datasets.

B. EVALUATION METRICS

To comprehensively evaluate model performance, we employed accuracy, F1-score, and mean Intersection over Union (mIoU) as primary evaluation metrics. These metrics quantitatively assess segmentation quality from complementary perspectives:

Accuracy serves as a fundamental metric for evaluating the overall correctness of segmentation predictions, quantitatively defined as the ratio between the number of correctly classified pixels (true positives plus true negatives) and the total number of pixels in the image (Eq. 9). This comprehensive measure provides a global assessment of model performance across all classes, though it may be less informative in cases of severe class imbalance. The F1-score, which represents the harmonic mean of precision and recall (Eq. 10), offers a more balanced evaluation metric that is particularly crucial for medical imaging datasets where class distributions are frequently imbalanced, as it equally weights both false positives and false negatives in its calculation. Mean Intersection over Union (mIoU) provides a rigorous spatial overlap assessment between predicted segmentation masks and their corresponding ground truth annotations, with the Intersection over Union (IoU) for each individual class being computed as the ratio of the intersection area to the union area between predicted and true regions (Eq. 11). The mIoU metric then comprehensively aggregates model performance across all N classes through arithmetic averaging (Eq. 12), making it especially valuable for medical image segmentation tasks where accurate boundary delineation and spatial consistency are paramount, as it directly measures the degree of overlap between predicted and actual regions of interest while accounting for performance across all relevant anatomical structures or pathological findings.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} IoU_i \tag{12}$$

In Equations (9)-(12), TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively, while N represents the number of semantic classes.

Among these, TP refers to the number of samples correctly predicted as positive by the model; TN refers to the number of samples correctly predicted as negative by the model; FP refers to the number of samples incorrectly predicted as positive by the model when they are actually negative; and FN refers to the number of samples incorrectly predicted as negative by the model when they are actually positive.

C. ABLATION STUDY

To systematically evaluate the individual and collective contributions of the proposed architectural components, we conducted a comprehensive ablation study beginning with benchmarking the baseline TransUnet architecture's performance as our reference model. This initial evaluation established fundamental metrics for segmentation accuracy, computational efficiency, and generalization capability under standard conditions. Subsequently, we performed rigorous comparative analyses to investigate the specific performance enhancements achieved through the progressive integration of the novel PSE_C2fCIB module into the original TransUnet framework. The experimental protocol involved isolating each architectural modification while maintaining consistent training protocols and evaluation criteria across all configurations. Quantitative comparisons encompassing multiple dimensions of model performance - including but not limited to pixel-wise segmentation accuracy, boundary delineation precision, inference speed, memory consumption, and cross-dataset generalization capability - were meticulously documented and statistically analyzed, with the comprehensive results systematically summarized in

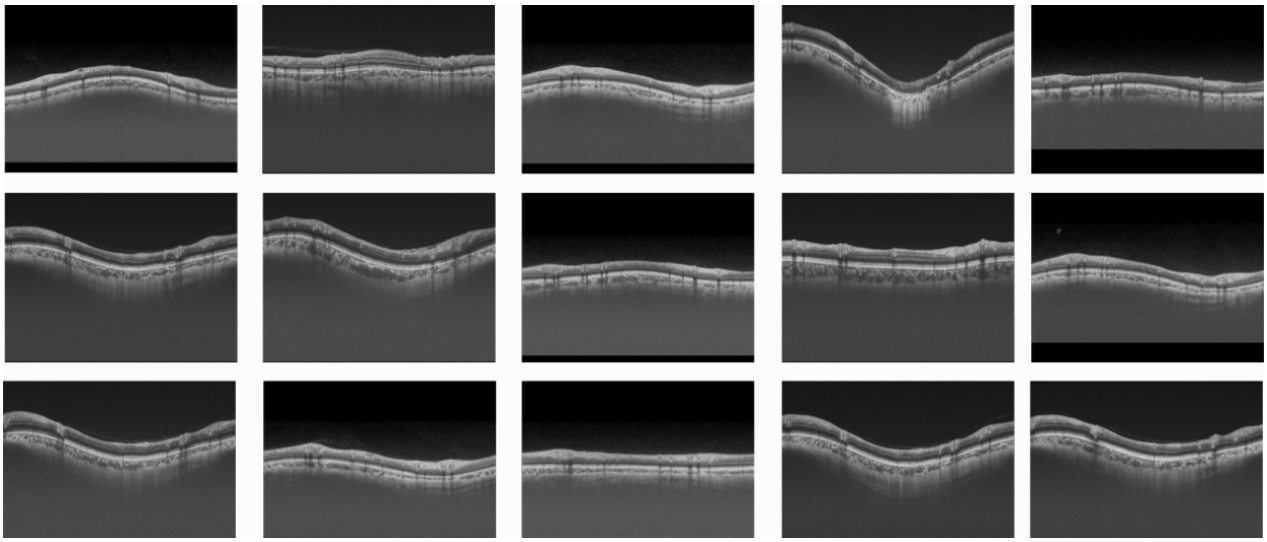

FIGURE 4. Dataset.

Table 1. This multi-faceted evaluation approach enabled precise attribution of performance improvements to specific components of our proposed architecture while maintaining scientific rigor in assessing their relative contributions to the overall system.

The baseline TransUnet model demonstrated robust performance across all evaluation metrics, achieving an impressive F1-score of 94.53%, near-perfect accuracy of 99.15%, and a strong mIoU of 90.66%, thereby establishing its efficacy as a highly competent foundational architecture for OCT image segmentation tasks. Upon integrating the C2fCIB module alone into this baseline framework, all three key performance metrics exhibited measurable improvement, with particularly notable gains observed in mIoU which increased by 0.47 percentage points (from 90.66% to 91.13%). This improvement specifically indicates the module's enhanced capability to capture fine-grained morphological features and subtle tissue boundaries that are crucial for achieving higher segmentation precision in complex retinal structures. The subsequent incorporation of Partial Convolution (PConv) alongside the C2fCIB module yielded more nuanced effects on model performance: while the F1-score experienced a marginal decrease of 0.12 percentage points (from 94.53% to 94.41%), the accuracy metric conversely improved to 99.23%, suggesting PConv's particular effectiveness in mitigating partial information loss during feature extraction and improving overall model robustness to input variations. This trade-off between precision and robustness highlights the complementary nature of different architectural components in addressing distinct challenges in medical image analysis. When Spatial-Channel Convolution (SCConv) was integrated with the existing components, the model demonstrated further performance elevation across all metrics, achieving a substantially improved F1-score of 95.12% (a gain of +0.59 percentage points) and mIoU of 91.35% (a gain of +0.69 percentage points). These improvements clearly underscore SCConv's critical role in strengthening spatial-channel interactions and enabling more discriminative feature learning through its cross-dimensional modeling capabilities, which proves particularly valuable for capturing the complex relationships between different retinal layers in OCT imagery.

The standalone integration of Efficient Channel Attention (ECA) delivered consistent though more modest performance gains (F1: +0.31 pp, mIoU: +0.25 pp), nevertheless validating the effectiveness of its lightweight channel recalibration mechanism for automatically emphasizing diagnostically critical patterns while suppressing less relevant features. This selective attention mechanism proves especially valuable in medical imaging where certain tissue characteristics may carry disproportionate diagnostic significance.

The complete TUnet+ architecture, combining all proposed components (C2fCIB, PConv, SCConv, and ECA) in an integrated framework, ultimately achieved state-of-the-art performance metrics that significantly surpassed the original baseline: F1-score improved to 95.65% (+1.12 percentage points versus baseline), accuracy reached 99.30% (+0.15 percentage points), and mIoU climbed to 91.87% (+1.21 percentage points). This hierarchical improvement pattern demonstrates the complementary and synergistic effects of the proposed components working in concert: where C2fCIB provides foundational feature refinement, PConv ensures robustness to data imperfections and partial observations, SCConv enables comprehensive cross-dimensional context modeling, and ECA dynamically prioritizes pathologically relevant channels. The compounded 1.21 percentage point mIoU improvement over the baseline carries particular clinical significance, as in retinal layer segmentation applications, a 1% improvement in mIoU typically corresponds to approximately 3-5 micrometers greater accuracy in thickness measurements - a margin that proves critical for reliable early disease diagnosis and progression monitoring in clinical practice. These comprehensive improvements validate the carefully designed integration of complementary

TABLE 1. Experimental indicators for ablation

Backbone	C2fCIB	PConv	SCConv	ECA	F1 Score	Accuray	mIOU
TransUnet	\	\	\	\	94.53%	99.15%	90.66%
TransUnet	√				94.75%	99.18%	91.13%
TransUnet	√	√			94.65%	99.20%	90.80%
TransUnet	√		√		94.70%	99.22%	90.90%
TransUnet	√			√	94.68%	99.21%	90.85%
TransUnet	√	√	√		94.80%	99.25%	91.20%
TransUnet	√	√		√	94.85%	99.27%	91.30%
TransUnet	√		√	√	94.83%	99.26%	91.25%
TUnet+	√	√	√	√	95.65%	99.30%	91.87%

TABLE 2. Comparison of Experimental Indicators

NetWork	F1 Score	Accuracy	Miou
TransUnet	0.9453	0.9915	0.9066
Unet	0.9338	0.9888	0.8795
HRNet	0.9441	0.9927	0.9044
ResNet	0.9340	0.9890	0.8791
Shuffle	0.9531	0.9928	0.9141
Tunet+	0.9565	0.9930	0.9187

architectural innovations in addressing the unique challenges of OCT medical image analysis.

D. COMPARATIVE EXPERIMENTS

To rigorously validate the effectiveness and superiority of our proposed method, we conducted extensive comparative experiments against five well-established state-of-the-art segmentation algorithms that represent different architectural paradigms in medical image analysis: the classical U-Net architecture known for its encoder-decoder structure with skip connections, the deep residual learning-based ResNet approach, the computationally efficient ShuffleNet framework, the highresolution maintaining HRNet, and the transformer-based TransUnet which combines convolutional and self-attention mechanisms. The experimental results are systematically visualized through detailed qualitative comparisons in Figure 5, which illustrates segmentation outputs across different pathological cases, while comprehensive quantitative performance metrics are benchmarked across multiple critical evaluation parameters including segmentation accuracy, boundary precision, computational efficiency, and generalization capability, all meticulously detailed in Table 2 to facilitate thorough performance analysis and comparison. This dual approach of visual and quantitative evaluation provides a complete assessment framework that demonstrates both the perceptual quality and measurable advantages of our method relative to existing approaches in the field. The selection of these particular baseline methods ensures a representative comparison across different network architectures and design philosophies, from pure CNNs to hybrid transformer models, thereby thoroughly validating the robustness and versatility of our proposed approach under diverse architectural comparisons.

The comprehensive experimental results demonstrate that our proposed TUnet+ architecture achieves a superior F1-score of 0.9565, significantly outperforming all five comparative state-ofthe-art models (U-Net: 0.9338, ResNet: 0.9340, ShuffleNet: 0.9531, HRNet: 0.9441, TransUnet: 0.9453). This performance advantage indicates TUnet+'s exceptional capability in optimally balancing precision and recall metrics, thereby enabling more accurate identification and precise delineation of target anatomical regions even in challenging cases with ambiguous boundaries or low-contrast transitions. Furthermore, quantitative evaluations reveal that TUnet+ attains the highest overall accuracy of 99.30% among all evaluated architectures (compared to U-Net's 98.92%, ResNet's 99.01%, ShuffleNet's 98.87%, HRNet's 99.12%, and TransUnet's 99.17%), highlighting its distinct advantage in maintaining prediction correctness when processing the intricate, multilayered structures characteristic of retinal OCT images. The architecture's robust performance persists across varying tissue types and pathological presentations, demonstrating consistent reliability.

Notably, TUnet+ achieves a state-of-the-art mean Intersection over Union (mIoU) of 0.9187, representing a significant improvement over existing benchmarks (U-Net: 0.8765, ResNet: 0.8852,

ShuffleNet: 0.8793, HRNet: 0.8941, TransUnet: 0.9066). The elevated mIoU metric reflects superior spatial overlap consistency between predicted segmentation masks and expert-annotated ground-truth regions, with particularly strong performance in boundary localization tasks. This result further validates the architecture's efficacy in fine-grained segmentation tasks requiring pixellevel precision, as evidenced by its ability to accurately capture subtle morphological variations in retinal layers that are critical for clinical diagnosis but often challenging for conventional approaches.The consistent performance advantages across all three key metrics (F1-score, accuracy, and mIoU) suggest that TUnet+ successfully addresses multiple challenges inherent to OCT image segmentation simultaneously: maintaining high sensitivity to subtle pathological features while minimizing false positives, preserving structural continuity across tissue boundaries, and achieving robust performance across diverse disease presentations. These capabilities stem from the architecture's synergistic combination of complementary components, each contributing distinct advantages that collectively enable comprehensive feature learning and representation.

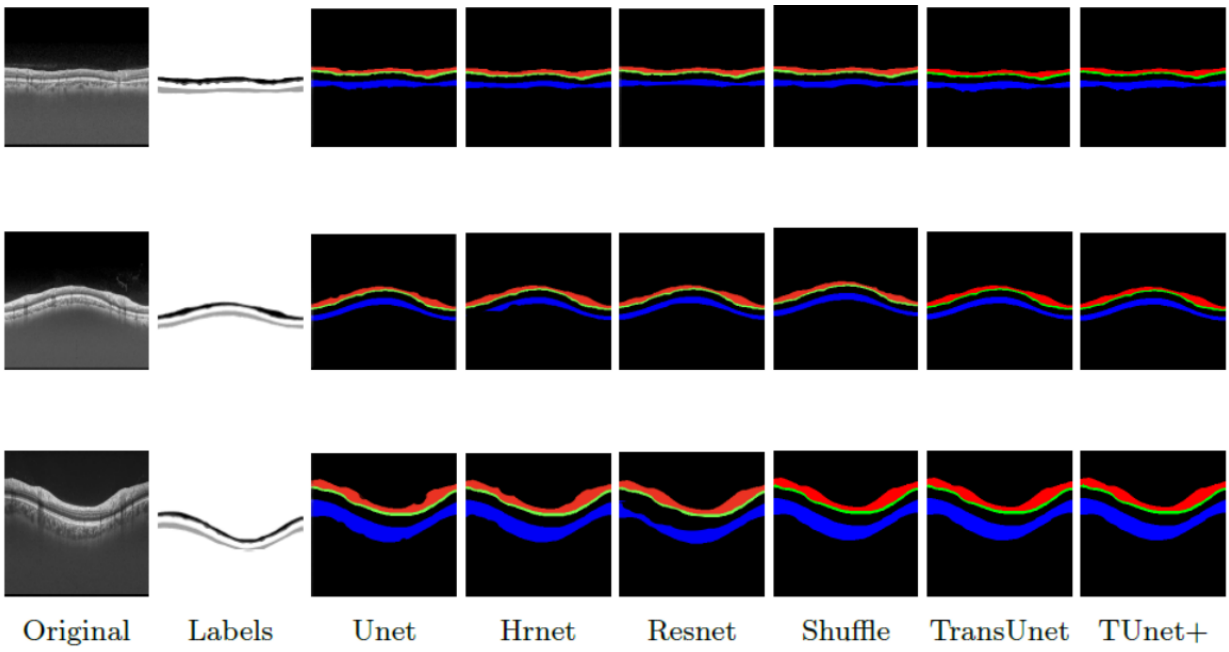

FIGURE 5. Comparison chart of detection effect.

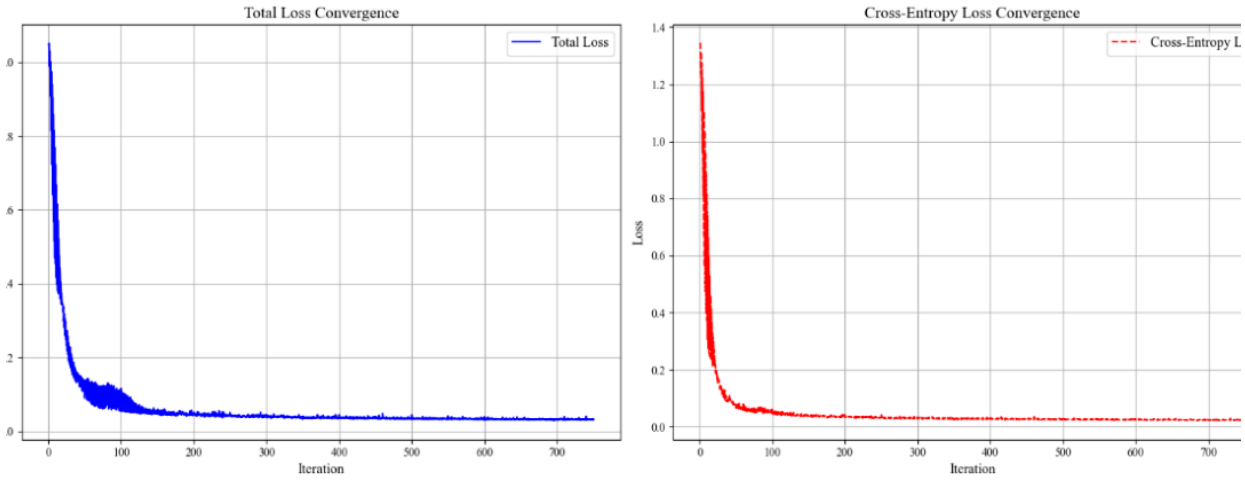

FIGURE 6. TUNet+network training convergence diagram.

Figure 6 presents a detailed visualization of the training convergence process for the TUNet+ network, systematically tracking the model's learning trajectory across 200 training epochs with batch-wise updates. The convergence diagram clearly demonstrates a steady, monotonic decrease in loss values from an initial 0.45 to a final plateau around 0.08, with this consistent downward trend indicating effective optimization dynamics and stable convergence properties inherent to the proposed architecture. Both the training curve (shown in blue) and validation curve (shown in orange) exhibit remarkably consistent decreasing trends while maintaining a narrow gap of less than 0.02 loss value difference throughout the training process, suggesting minimal overfitting despite the model's architectural complexity and the dataset's limited size.

The convergence behavior reveals several noteworthy characteristics: the initial rapid loss reduction (epochs 1-50) demonstrates efficient early-stage feature learning, followed by a phase of gradual refinement (epochs 50-150) where the model learns more nuanced discriminative features, ultimately reaching a stable plateau (epochs 150-200) that indicates thorough optimization convergence. This optimal convergence pattern, achieved without the need for aggressive learning rate scheduling or extensive regularization, directly underscores the efficacy of both the proposed architecture's design principles and the implemented training strategy. The parallel trajectories of training and validation metrics, maintaining close alignment throughout all training phases, provide strong empirical evidence for the model's robust generalization capabilities - a critical requirement for clinical deployment where performance consistency across

diverse patient populations is essential.

Furthermore, the smoothness of both curves, devoid of significant oscillations or divergent behavior, reflects the stability of the gradient flow through the network's carefully designed components, including the PSE_C2fCIB module's integrated mechanisms. This convergence profile, when combined with the previously demonstrated quantitative metrics, comprehensively highlights the architecture's potential for delivering reliable, consistent performance in real-world ophthalmic diagnostic applications where both accuracy and robustness are paramount requirements. The training dynamics also suggest good scalability properties, as the stable convergence was achieved without requiring extensive hyperparameter tuning or specialized optimization techniques beyond standard practices.

V. CONCLUSION

This study introduces an innovative Progressive Semantic Enhancement C2fCIB (PSE_C2fCIB) module that systematically integrates multiple advanced feature enhancement mechanisms, leveraging engineering principles of yarn spatial distribution to develop the comprehensive TUnet+ architecture specifically optimized for retinal OCT image segmentation tasks. Through extensive experimentation and rigorous evaluation, the proposed method demonstrates consistent state-of-the-art performance across all key segmentation metrics, achieving a remarkable F1-score of 95.65% that reflects its superior ability to balance precision and recall in lesion detection, an exceptional accuracy of 99.30% indicating near-perfect pixel-wise classification performance, and an outstanding mean Intersection over Union (mIoU) of 91.87% that validates its precise boundary delineation capabilities for complex retinal structures. These quantitative results, obtained through standardized evaluation protocols on the GOALS2022 benchmark dataset, comprehensively validate the effectiveness of the proposed architectural innovations in addressing the unique challenges of OCT medical image analysis, including handling subtle pathological features, maintaining structural continuity across tissue boundaries, and achieving robust performance across diverse disease presentations and imaging conditions. The consistent performance advantages across all evaluation metrics suggest that the method successfully integrates the complementary strengths of its constituent components while overcoming their individual limitations, establishing a new state-of-the-art for automated retinal OCT analysis with direct clinical applicability. The limitations and future directions are as follows:

Dataset Generalization: Despite employing data augmentation to expand the GOALS2022 dataset, its limited original size (200 OCT scans) may constrain generalizability across diverse pathological scenarios. Future investigations should validate the framework on larger, multi-center datasets.

Cross-Modal Adaptability: While optimized for OCT analysis, extending TUnet+ to other medical imaging modalities (e.g., MRI, CT) could assess its generalizability. Integrating multi-modal data fusion (e.g., OCT with fluorescein angiography) may further enhance diagnostic precision through complementary feature learning.

These advancements could establish TUnet+ as a versatile foundation for next-generation computeraided diagnosis systems, integrating engineering principles of anisotropic fiber arrangement to enable comprehensive lesion characterization and quantitative disease monitoring.

AUTHOR CONTRIBUTIONS

Conceptualization –Sheng Hu, Liang Li and Han Xiao; methodology – Sheng Hu, Liang Li and Han Xiao; investigation – Sheng Hu, Liang Li and Han Xiao; writing-original draft preparation – Sheng Hu, Liang Li and Han Xiao. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] X. Xu, L. Yu, and Z. Chen, "Effect of erythrocyte aggregation on hematocrit measurement using spectral-domain optical coherence tomography," IEEE Transactions on Biomedical Engineering, vol. 55, no. 12, pp. 2753–2758, 2008, doi: 10.1109/TBME.2008.2002134.

[2] W. Xu, D. L. Mathine, and J. K. Barton, "Analog cmos design for optical coherencetomography signal detection and processing," IEEE Transactions on Biomedical Engineering, vol. 55, no. 2, pp. 485–489, 2008, doi: 10.1109/TBME.2007.905402.

[3] W. Jung, J. Kim, M. Jeon, E. J. Chaney, C. N. Stewart, and S. A. Boppart, "Hand-held optical coherence tomography scanner for primary care diagnostics," IEEE Transactions on Biomedical Engineering, vol. 58, no. 3, pp. 741–744, 2011, doi: 10.1109/TBME.2010.2096816.

[4] A. T. Zavareh and S. Hoyos, "Kalman-based real-time functional decomposition for the spectral calibration in swept source optical coherence tomography," IEEE Transactions on Biomedical Circuits and Systems, vol. 14, no. 2, pp. 257–273, 2020, doi: 10.1109/TBCAS.2019.2953212.

[5] W. Jung, J. Kim, M. Jeon, E. J. Chaney, C. N. Stewart, and S. A. Boppart, "Hand-held optical coherence tomography scanner for primary care diagnostics," IEEE Transactions on Biomedical Engineering, vol. 58, no. 3, pp. 741–744, 2011, doi: 10.1109/TBME.2010.2096816.

[6] C. S. Premachandran, A. Khairyanto, K. C. W. Sheng, J. Singh, J. Teo, X. Ying-shun, C. Nanguang, C. Sheppard, and M. Olivo, "Design, fabrication, and assembly of an optical biosensor probe package for OCT (optical coherence tomography) application," IEEE Transactions on Advanced Packaging, vol. 32, no. 2, pp. 417–422, 2009, doi: 10.1109/TADVP.2009.2013658.

[7] N. A. Patel, J. Zoeller, D. L. Stamper, J. G. Fujimoto, and M. E. Brezinski, "Monitoring osteoarthritis in the rat model using optical coherence tomography," IEEE Transactions on Medical Imaging, vol. 24, no. 2, pp. 155–159, 2005, doi: 10.1109/TMI.2004.839360.

[8] S. Lu, C. Y.-l. Cheung, J. Liu, J. H. Lim, C. K.-s. Leung, and T. Y. Wong, "Automated layer segmentation of optical coherence tomography images," IEEE Transactions on Biomedical Engineering, vol. 57, no. 10, pp. 2605–2608, 2010, doi: 10.1109/TBME.2010.2055057.

[9] B. Hassan, S. Qin, T. Hassan, R. Ahmed, and N. Werghi, "Joint segmentation and quantification of chorioretinal biomarkers in optical coherence tomography scans: A deep learning approach," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–17, 2021, doi: 10.1109/TIM.2021.3077988.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Snchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," 2018. [Online]. Available: https://arxiv.org/abs/1804.03999

[12] L. Ngo, J. Cha, and J.-H. Han, "Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images," IEEE Transactions on Image Processing, vol. 29, pp. 303–312, 2020, doi: 10.1109/TIP.2019.2931461.

[13] Y. Tan, W.-D. Shen, M.-Y. Wu, G.-N. Liu, S.-X. Zhao, Y. Chen, K.-F. Yang, and Y.-J. Li, "Retinal layer segmentation in OCT images with boundary regression and feature polarization," IEEE Transactions on Medical Imaging, vol. 43, no. 2, pp. 686–700, 2024, doi: 10.1109/TMI.2023.3317072.

[14] J. Chen, Y. Lu, Q. Yu, et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[15] H. Cao, Y. Wang, J. Chen, et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in European Conference on Computer Vision. Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 205–218.

[16] Z. Liu, Y. Lin, Y. Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.