Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 33(9): 5496–5517.
DOI: 10.3934/era.2025246
Received: 13 May 2025
Revised: 04 July 2025
Accepted: 10 July 2025
Published: 15 September 2025

Research article

OD/OC-semantic FPN: an enhanced optic cup and disc segmentation model in color fundus images using improved MaxViT and semantic feature pyramid network

Xuan Liu[1,†], Qian Ma[2,†], Jiajia Wang[1], Xiaohu Liu[1], Qiuyang Zhang[3], Jin Yao[3,*], Biao Yan[4,*] and Zhenhua Wang[1,*]

[1] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
[2] General Hospital of Ningxia Medical University, Ningxia 750001, China
[3] Department of Ophthalmology and Optometry, The Affiliated Eye Hospital, Nanjing Medical University, Nanjing 210029, China
[4] Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China

* **Correspondence:** Email: jinyao@126.com, yanbiao@sjtu.edu.cn, zh-wang@shou.edu.cn; Tel: +8613052312346; Fax: 02161900627.

† These authors have contributed equally to this work.

Abstract: Glaucoma, a leading cause of irreversible blindness, requires early detection to prevent progressive vision loss. Color fundus photography is a non-invasive and widely accessible modality for glaucoma screening; however, traditional manual interpretation is limited by subjectivity, time inefficiency, and inter-observer variability. This study proposes an optic disc (OD)/optic cup (OC)semantic feature pyramid network, a joint OD and OC segmentation model for glaucoma screening. The model extends the Semantic FPN architecture through three key enhancements: (1) a MaxViT backbone that incorporates multi-axis attention to reinforce local-global feature interaction and preserve boundary information during downsampling; (2) inception depthwise convolution modules embedded within MBConv blocks, which enables multi-scale convolution to expand receptive fields without compromising fine-grained details; (3) an optimized semantic FPN structure to improve the cross-scale feature alignment and multi-scale fusion. The proposed OD/OC-Semantic FPN was evaluated on five publicly available fundus image datasets (Drishti-GS, ORIGA, RIM-ONE DL, RIM-ONE-R3, and REFUGE), and its performance was compared against several state-of-the-art segmentation models (U-Net, DeepLabV3+, PSPNet, APCNet, semantic FPN-PoolFormer, and attention U-Net). The results show that the OD/OC-semantic FPN surpasses existing models across several metrics: dice coefficient, Mean Intersection over Union (mIoU), mean pixal accuracy (MPA), and classification accuracy, thus demonstrating superior structural precision for fundus analysis. Collectively, these results indicate that the OD/OC-Semantic FPN is a robust and generalizable tool for intelligent early detection of glaucoma.

Keywords: glaucoma; color fundus image; deep learning; transformer; segmentation model

1. Introduction

Glaucoma is a chronic and progressive eye disease characterized by irreversible damage to the optic nerve, which, if left undiagnosed or poorly managed, can lead to permanent vision loss and significantly impair the patients' quality of life [1]. Globally, glaucoma ranks among the leading causes of irreversible blindness. Numerous clinical studies have demonstrated that early detection, an accurate diagnosis, and a timely intervention are essential to halt disease progression. Accordingly, the development of efficient and accurate automated diagnostic tools for large-scale screening carries substantial clinical and public health value.

Color fundus photography is a non-contact, cost-effective, and widely accessible retinal imaging technique, extensively used in community-based screenings and primary eye care settings. By delineating the optic disc (OD) and optic cup (OC) from fundus images and calculating the cup-to-disc ratio (CDR), clinicians can effectively assess the risk of glaucoma [2]. As illustrated in Figure 1, glaucomatous fundus images typically exhibit an enlarged OC and elevated CDR compared to healthy eyes.

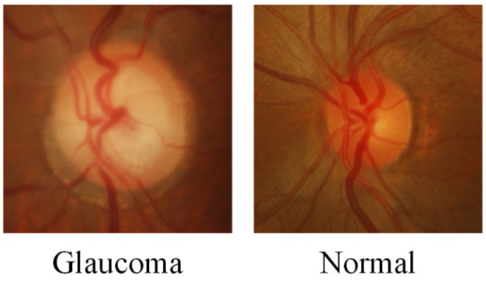

Figure 1. Comparison between glaucomatous and normal fundus images.

In clinical practice, the boundaries of the OD and the OC are typically manually delineated by ophthalmologists, followed by a visual estimation of the cup-to-disc ratio (CDR). However, traditional manual segmentation methods exhibit several limitations, including high subjectivity, low efficiency, and pronounced sensitivity to image quality. Due to their reliance on expert judgment, manual approaches are susceptible to inter-observer variability, which leads to inconsistent and less reproducible results. Moreover, the process of individually annotating each image is time-consuming and labor-intensive, which poses substantial barriers to large-scale screening initiatives. These challenges are further exacerbated when analyzing images with low contrast, uneven illumination, or interference from pathological features, where manual methods often struggle to accurately delineate the blurred boundaries of the OD and the OC [3]. Therefore, the development of computer vision–based algorithms for automated OD and OC segmentation is essential. Such technologies can enhance the consistency and efficiency of fundus image interpretation and enable scalable, automated glaucoma screening in community and primary care settings. Ultimately, these tools can optimize healthcare resource allocation and facilitate early detection and timely interventions.

2. Related work

In recent years, deep learning has revolutionized the field of retinal image analyses, thereby offering robust solutions for tasks such as OD and OC segmentation, glaucoma classification, and their integration. Existing studies can be broadly categorized into three types based on the task objectives.

2.1. Segmentation-only approaches

Segmentation approaches focus on delineating the OD and OC boundaries. For example, Haider et al. [4] proposed a dual sub-network architecture that incorporated depthwise separable convolutions and residual connections to enhance the segmentation performance. Mahrooqi et al. [5] introduced GARDNet, which leverages multi-view fusion from original, cropped, and polar-transformed inputs. Bian et al. [6] embedded anatomical priors and attention mechanisms into a generative adversarial network (GAN)-based cascaded network to improve the small-structure segmentation. Shyamalee et al. [7] further embedded attention gates into a ResNet50-based U-Net to suppress background noise and enhance the target region sensitivity. In response to data scarcity, Bengani et al. [8] adopted transfer learning and semi-supervised learning to improve the generalizability. Jiang et al. [9] utilized dilated convolutions within a region proposal framework to refine segmentation, while other studies employed multi-scale feature pyramids [10], attention modules [11], or Transformer components fused with U-Net [12]. Chen et al. [13] designed a parallel structure that combined boundary-aware and adversarial learning branches to enhance edge localization.

2.2. Classification-only approaches

Classification approaches aim to directly predict the glaucoma status based on global image features. Shyamalee et al. [14] conducted comparative studies of convolutional neural network (CNN) architectures (inception-V3, VGG19, ResNet50), and showed that contrast enhancement, global average pooling, and Dropout significantly improved the generalization. Wassel et al. [15] evaluated multiple vision transformer (ViT) variants, including Swin Transformer and CrossViT, and achieved high area under the curve (AUCs) using aggregated public datasets. Singh et al. [16] applied transfer learning with Inception-V3, performed extensive data preprocessing, and achieved high sensitivity even on limited datasets.

2.3. Hybrid segmentation and classification approaches

Hybrid approaches that combine segmentation and classification have attracted increasing attention. Nawaz et al. [17] employed EfficientDet-D0 and EfficientNet-B0 for OD/OC localization and glaucoma classification. Shyamalee et al. [18] enhanced U-Net by replacing the encoder with pre-trained CNNs and adding attention gates, thus enabling both segmentation and classification. Hervella et al. [19] developed a shared encoder-decoder framework with a lightweight classification head based on 1 × 1 convolutions and global average pooling. Wang et al. [20] used a ResNet101-UperNet architecture for optic disc segmentation and integrated a ViT backbone with masked autoencoder pretraining for classification tasks.

Despite the rapid advancement of deep learning techniques, several key challenges in OD and OC segmentation remain unresolved. The presence of blurred boundaries, overlapping anatomical structures, and significant inter-subject morphological variability continue to hinder the segmentation accuracy and robustness [3]. While existing deep learning approaches, primarily based on CNNs and ViTs have demonstrated promising results, their full potential is often constrained by inherent architectural limitations and data scarcity, particularly in real-world clinical applications. These challenges highlight the need for hybrid architectures capable of combining the fine-grained feature extraction capabilities of CNNs with the global context modeling strength of Transformers, thereby enabling more robust and scalable solutions.

Inspired by [21], this study proposes a hybrid framework that integrates the complementary advantages of CNNs and ViTs for joint OD and OC segmentation. Specifically, a semantic feature pyramid network (semantic FPN) architecture is adopted, thereby incorporating an enhanced multi-axis vision Transformer (MaxViT) as the backbone. This design enables the model to capture both localized structural details and long-range contextual dependencies, thereby improving there boundary delineation and structural integrity. Furthermore, the semantic pyramid facilitates effective multi-scale feature fusion, thus allowing the model to better adapt to morphological variations and cope with challenging imaging conditions such as low contrast or lesion interference. By bridging the gap between local sensitivity and global perception, the proposed model aims to deliver an accurate, robust, and generalizable segmentation performance, particularly in clinically relevant scenarios that involve limited or imbalanced datasets.

3. OD/OC joint segmentation model

Figure 2 presents the architecture of the OD/OC-semantic FPN, a joint segmentation model for the OD and OC based on a Semantic FPN. The model is comprised of three primary components: (1) backbone network: MaxViT [22] is employed as the feature extractor to capture multi-scale representations from fundus images;(2) module enhancement: the MBConv modules within MaxViT are upgraded using inception depthwise convolution [23], which introduces larger convolution kernels to broaden the receptive field and improve feature discrimination; and (3) feature integration: the enhanced MaxViT generates hierarchical feature maps that are passed into the semantic FPN [24], which enables the fusion of global semantics and local details. High-level features are upsampled via convolution and bilinear interpolation to spatially align with low-level features, followed by channel-wise fusion to produce the final OD/OC segmentation output.

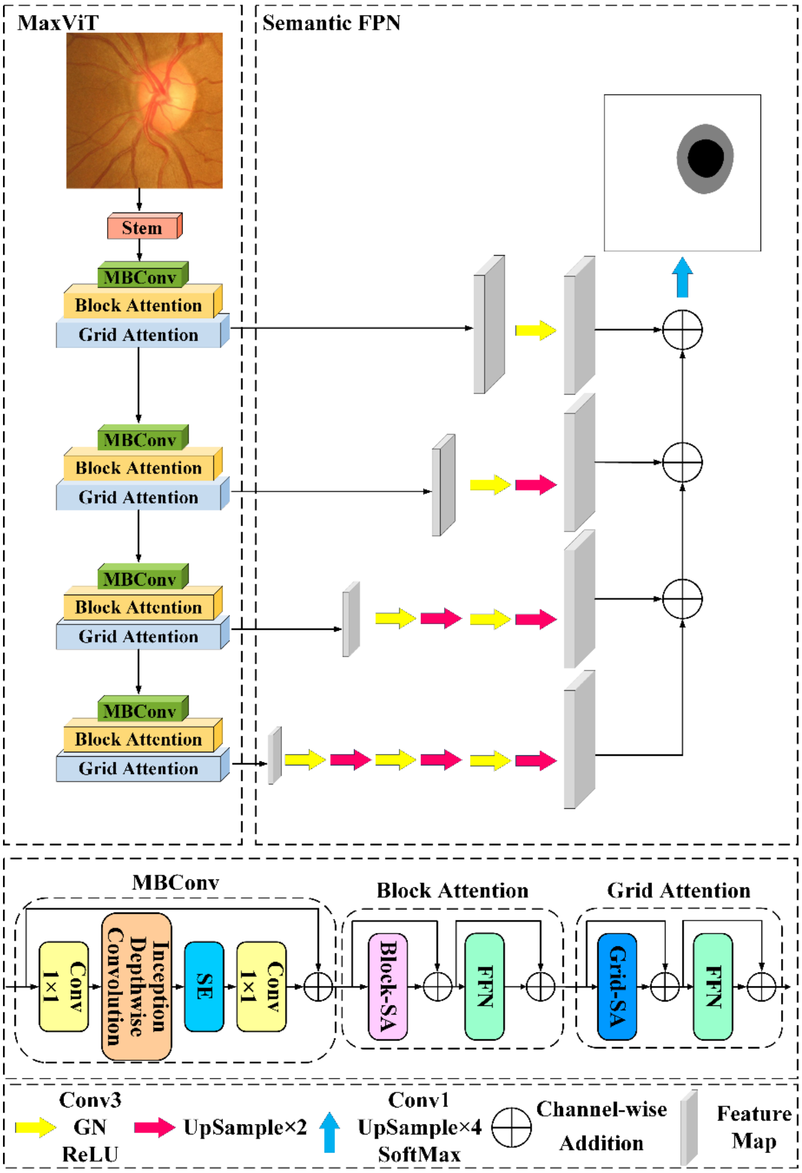

Figure 2. Architecture of the OD/OC-Semantic FPN.

3.1. Backbone network: MaxViT

The MaxViT architecture begins with a Stem module composed of two sequential convolutional layers, which extract low-level visual features while performing an initial spatial down sampling. The core of the network consists of four successive stages, each containing multiple MaxViT Blocks that combine local and global attention mechanisms to effectively model multi-scale contextual information.

As the network deepens, the spatial resolution is halved while the number of channels doubles at each stage, thus increasing both the semantic richness and model capacity. Figure 3 shows the structure of a MaxViT Block, which is comprised of two main components: the Mobile Inverted Bottleneck Convolution (MBConv) and the Multi-Axis Attention module.

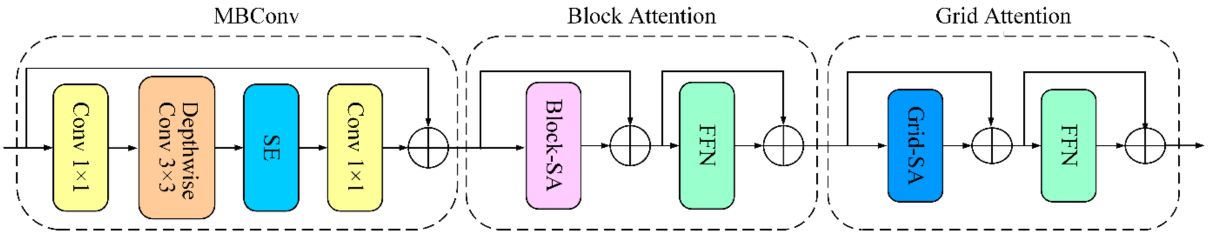

Figure 3. Structure of a MaxViT block.

The MBConv module, especially with depth wise convolution (DWConv), significantly enhances the feature extraction efficiency. DWConv independently operates on the individual input channels, reducing the parameters and computational cost while preserving the spatial detail, which is crucial for to capture fine contours in complex retinal imagery.

The multi-axis attention module integrates block attention and grid attention, thus forming a hybrid attention mechanism that balances fine-grained detail capture with global contextual understanding.

As illustrated in Figure 4, block attention partitions the input feature map into non-overlapping local windows and applies self-attention within each window. This method efficiently captures subtle variations, such as edge irregularities or small lesions at OD/OC boundaries, while maintaining computational efficiency.

Given an input feature map $X \in R^{H\times W\times C}$, block attention reshapes it into a tensor of shape $X \in R^{(\frac{H}{p}\times\frac{W}{p},p\times p,C)}$, thus partitioning it into windows of size $p \times p$. Then, self-attention is independently applied within each window to enable fine-grained local information exchange.

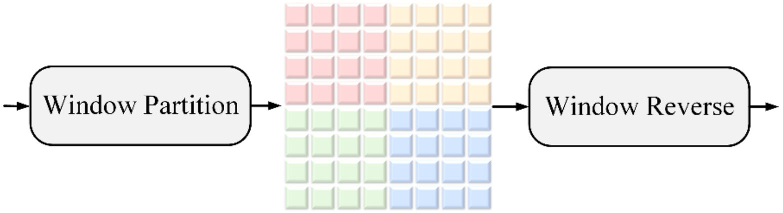

Figure 4. block attention mechanism.

Figure 5 depicts grid attention, which breaks the limitations of local windows by using sparse sampling to reorganize the spatial features into regularly spaced grids. This enhances the model's ability to capture long-range dependencies and global context, even under a high-resolution input.

The input tensor $X \in R^{H\times W\times C}$ is reconstructed into $X \in R^{(G\times\frac{H}{G},G\times\frac{W}{G},C)}$ using a grid size of G×G, thus enabling adaptive window sizes of $\frac{H}{G}\times\frac{W}{G}$ for self-attention.

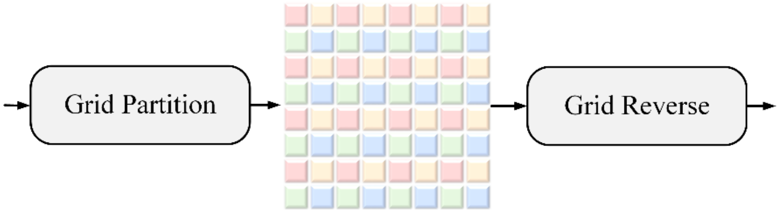

Figure 5. grid attention mechanism.

Together, block and grid attention offer a complementary framework: the former excels in local feature refinement, while the latter boosts global context integration.

Compared to traditional convolutional networks such as ResNet, which primarily rely on local operations and residual connections to mitigate gradient issues, MaxViT introduces a hybrid architecture that combines the efficiency of CNNs in local feature extraction with the global modeling power of Transformers. Each MaxViT block integrates MBConv and Multi-Axis Attention, which enables the model to robustly extract hierarchical spatial and semantic information from complex, low-quality fundus images.

3.2. MBConv enhancement via inception depthwise convolution

The accurate segmentation of OD and OC structures requires the model to effectively capture contextual information across large spatial regions. Conventional methods often adopt large rectangular convolutional kernels to expand the receptive field, but such approaches are computationally intensive and may fail to preserve fine edge details.

In the MaxViT architecture, the MBConv block serves as a fundamental building unit, thereby combining depthwise (DW) and pointwise (PW) convolutions for efficient feature extraction. However, the standard DW convolution in MBConv, which utilizes square kernels, faces limitations when processing OD/OC structures with irregular geometries in fundus images. Specifically, it struggles to capture their characteristic radial distributions. To address this limitation, an enhanced MBConv is designed by introducing an Inception-style DW convolution. This improvement is motivated by the anatomical property that the OD and OC exhibit anisotropic dimensions, on which the vertical and horizontal diameters are often unequal. By incorporating elongated strip-shaped kernels into the DW convolution, the enhanced MBConv block can more accurately model these elongated spatial features. Moreover, increasing the kernel dimensions effectively enlarges the receptive field, which allows the model to capture global contextual information while preserving essential boundary details.

As shown in Figure 6, the enhanced MBConv block applies inception DW convolution to process feature maps in parallel with kernels of different sizes and shapes. Specifically, 1/8 of the input channels are processed with a 3 × 3 DW convolution, another 1/8 with a 1 × 11 vertical strip DW convolution, and a final 1/8 with an 11 × 1 horizontal strip DW convolution. The remaining 5/8 of the channels undergo a standard 3 × 3 convolution. Compared with the original MBConv design in MaxViT, which uses a single-scale square DW convolution, our enhanced block improves the spatial sensitivity and significantly extends the receptive field, all while maintaining a low computational cost and better preserving the fine edge details of the OD and OC structures.

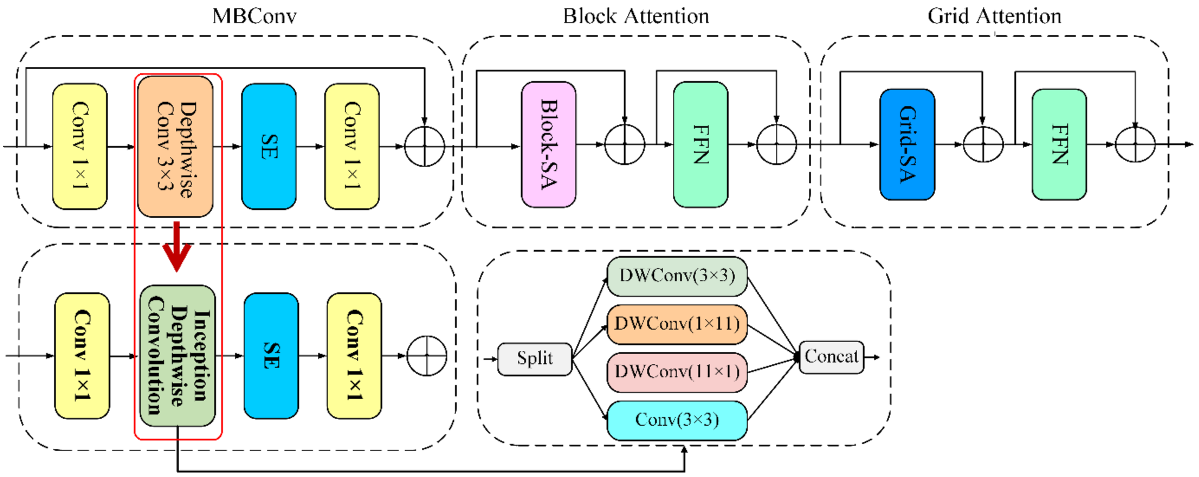

Figure 6. Enhanced MBConv structure with inception DW convolution.

Formally, the input feature map is first divided into four groups along the channel dimension:

$$X_{hw}, X_{w}, X_{h}, X_{id} = X_{:,g}, X_{:g:2g}, X_{:2g:3g}, X_{:3g:}, \tag{1}$$

where g is the number of channels per convolution branch, and C is the total number of input channels. By setting a proportionr_g, the number of channels in each branch can be determined as $g = r_g \ C$. In this case, $r_g = 0.125$. Once the groups are divided, they are processed by their corresponding convolution branches in parallel:

$$X'_{hw} = DWConv_{3\times3}^{g\rightarrow g}g(X_{hw}), \tag{2}$$

$$X'_{w} = DWConv_{1\times11}^{g\rightarrow g}g(X_{w}), \tag{3}$$

$$X'_{h} = DWConv_{11\times1}^{g\rightarrow g}g(X_{h}), \tag{4}$$

$$X'_{id} = Conv_{3\times3}(X_{id}), \tag{5}$$

Finally, the outputs of all branches are concatenated along the channel dimension to form the fused feature representation:

$$X' = Concat(X'_{hw}, X'_{w}, X'_{h}, X'_{id}), \tag{6}$$

3.3. OD/OC segmentation using semantic FPN

As illustrated in Figure 7, the four stages of MaxViT generate multi-scale feature maps denoted as F1, F2, F3, and F4, thus capturing hierarchical representations from coarse contours to fine edges. Among them, F1 retains the highest spatial resolution and fine structural details, while F4 encodes richer semantic information at a lower resolution. Then, these feature maps are fed into the semantic FPN, which employs a hierarchical feature fusion strategy to enable effective multi-scale interactions. For each stage, the features are first aligned in the channel dimension using a 3 × 3 convolution followed by group normalization, and further processed through ReLU activation for nonlinear mapping. To ensure resolution consistency across stages, a bilinear interpolation is applied (upscaling by a factor of 2), which allows higher-level feature maps to match the spatial resolution of the preceding stage before aggregation. During the key fusion stage, channel-wise addition is utilized to combine features from different levels. This design ensures that the rich semantic context encoded by deeper layers is complemented by the fine-grained spatial details preserved in shallower layers. Finally, the fused features undergo 4× upsampling and a 1 × 1 convolution to compress the channel dimension and produce the final segmentation map. Throughout the entire process, a progressive upsampling strategy helps minimize information loss, while group normalization contributes to maintaining the feature stability and model robustness. This top-down fusion pathway effectively integrates global contextual semantics with local geometric structures, thereby enhancing the segmentation precision and consistency in retinal fundus images.

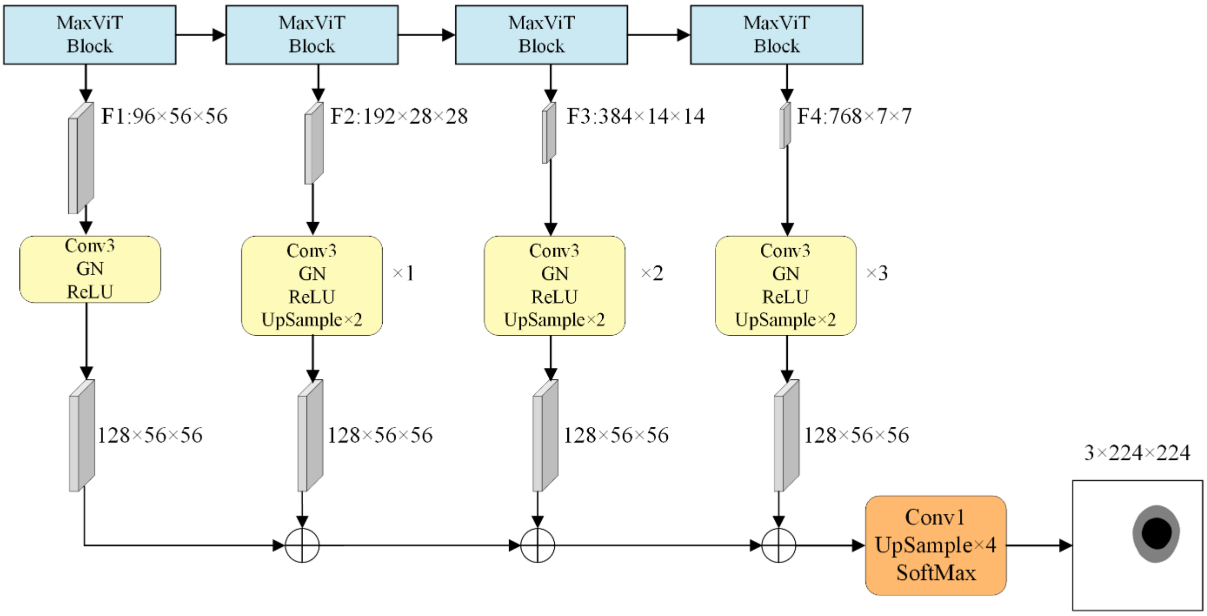

Figure 7. Semantic FPN structure.

4. Experiment

4.1. Dataset description and preprocessing

In this study, five publicly available retinal fundus image datasets were used: DRISHTI-GS, ORIGA, RIM-ONE DL, RIM-ONE-R3, and REFUGE. These datasets were selected to ensure diversity in the image quality, glaucoma prevalence, and annotation protocols, thus providing a comprehensive benchmark for both OD/OC segmentation and glaucoma detection tasks. All datasets include expert-labeled OD and OC boundaries, and some additionally provide glaucoma diagnosis labels, making them suitable for multi-task evaluation. A detailed summary is provided below.

DRISHTI-GS [25] contains 101 retinal fundus images, including 70 glaucoma cases and 31 non-glaucoma cases, and is divided into 50 training and 51 testing images. OD and OC annotations were provided by four ophthalmologists with different levels of clinical experience. The final ground truth was obtained by averaging the boundary annotations from all experts. ORIGA [26] includes 650 images, equally divided into 325 images for training and 325 images for testing, and is comprised of 168 glaucoma and 482 non-glaucoma cases. Elliptical OD/OC boundaries were manually annotated by ophthalmologists at the Singapore Eye Research Institute. The final segmentation masks were generated by fusing multiple expert annotations into soft boundary maps. RIM-ONE DL [27] consists of 485 images (172 glaucomatous and 313 non-glaucomatous), with 311 images used for training and 174 images used for testing. Annotations were provided by five ophthalmologists, and the final reference masks were obtained by averaging their manually annotated boundaries. RIM-ONE-R3 [28] contains 159 images, including 74 glaucoma and 85 non-glaucoma cases, split into 111 training, 24 validation, and 24 testing images. Manual OD/OC segmentations were provided by two ophthalmologists, and the final annotation was computed by averaging their results. REFUGE [29] consists of 1,200 images, including 120 glaucoma and 1080 non-glaucoma cases. The dataset is evenly divided into 400 training, 400 validation, and 400 testing images. Annotations were independently performed by seven glaucoma-specialized ophthalmologists, and the final ground truth was produced through majority voting.

In summary, the five datasets collectively provide 2595 retinal fundus images, comprised of 604 glaucoma and 1991 non-glaucoma cases. The annotations were all performed by qualified ophthalmologists, with final labels derived via averaging or consensus-based strategies, thus ensuring high-quality and reliable ground truth for segmentation and classification tasks.

To reduce variability caused by differences in illumination and contrast across datasets, all images were first normalized. In addition, contrast limited adaptive histogram equalization (CLAHE) was applied to enhance the local contrast while suppressing noise amplification.

4.2. Experimental environment

All experiments were carried out on a Windows 10 operating system using PyCharm as the development environment. The deep learning framework employed was PyTorch, version 1.10.2, configured to leverage CUDA 12.0 for GPU acceleration. The hardware setup included an 11th-generation Intel Core i7-11700 CPU with 16 GB of system memory, paired with an NVIDIA GeForce RTX 3060 GPU featuring 12 GB of dedicated VRAM. This configuration ensured efficient parallel processing capabilities for computationally intensive deep learning tasks.

4.3. Evaluation metrics

To comprehensively evaluate the segmentation performance of the model, multiple metrics were employed: (1) dice coefficient: measures the overlap between the predicted and ground truth regions, where a higher value indicates greater overlap and better segmentation performance; (2) mean intersection over union (MIoU): assesses the overall similarity between the predicted and ground truth labels across all three classes (OD, OC, and background), where higher MIoU values indicate better segmentation performance; (3) mean pixel accuracy (MPA): calculates the average pixel accuracy for background, OC, and OD, thus reflecting the model's precision at the pixel level; (4) accuracy: represents the proportion of correctly classified pixels in the test dataset, thus reflecting the overall segmentation accuracy of the model; and (5) frames per second (FPS): measures the number of images the model can process per second, thus indicating its real-time processing capability.

$$Dice = \frac{2 \cdot p_{ii}}{\left(\sum_{j=0}^{k-1} p_{ij} + \sum_{j=0}^{k-1} p_{ji} + p_{ii}\right)} \tag{7}$$

$$MIoU = \frac{1}{k}\sum_{i=0}^{k-1}\left(\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}+\sum_{j=0}^{k} p_{ji}-p_{ii}}\right) \tag{8}$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{n} \tag{9}$$

$$Accuracy = \frac{1}{n}\sum_{i=0}^{k} p_{ii} \tag{10}$$

$$FPS = \frac{1}{t} \tag{11}$$

Here, p_{ii}denotes the number of pixels correctly classified as OC/OD regions, p_{ij} refers to the number of pixels incorrectly classified as background, and p_{ji} indicates the number of background pixels misclassified as OC/OD regions. k represents the different annotated categories, where $k = 0$ denotes the background class, $k = 1$ denotes the OC, and $k = 2$ denotes the OD. n is the total number of pixels. t represents the average time required to process a single image.

4.4. Comparative analysis of experimental results

To rigorously evaluate the effectiveness of the proposed model, ablation and comparative experiments were systematically designed. All models were trained and tested under consistent experimental settings, including identical data splits, preprocessing pipelines, training epochs, optimization strategies, and loss functions, to ensure fairness and reproducibility. Specifically, all input images were resized to 224 × 224 pixels. U-Net and Semantic FPN were trained using the Adam optimizer with an initial learning rate of 0.0001. DeepLabV3+ was optimized using stochastic gradient descent (SGD) with a learning rate of 0.007, while APCNet and PSPNet were trained using SGD with a learning rate of 0.01. A batch size of 4 and 100 training epochs were uniformly applied across all models to ensure sufficient convergence.

4.4.1. Ablation study

To validate the effectiveness of the proposed improvements, three experimental models were designed and compared: (1) a baseline model using the traditional ResNet50 backbone; (2) an upgraded model that replaces ResNet50 with MaxViT; and (3) the proposed OD/OC-Semantic FPN model that incorporates both the MaxViT backbone and the enhanced MBConv module. All models were evaluated on five publicly available datasets: DRISHTI-GS, ORIGA, RIM-ONE DL, RIM-ONE-R3, and REFUGE. Detailed quantitative results are summarized in Tables 1–5, and demonstrate the progressive performance gains from each architectural enhancement.

Table 1. Ablation study on the Drishti-GS dataset.

MaxViT	Improved MaxViT	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
		87.70±1.17	80.31±0.82	81.38±0.83	90.51±0.35	95.24±0.25	3.40±0.18
√		90.10±0.28	82.48±0.42	83.52±0.35	92.31±0.22	95.86±0.09	2.87±0.17
	√	93.32±0.05	88.09±0.53	88.18±0.04	93.69±0.18	97.15±0.02	2.60±0.30

On the Drishti-GS dataset, using MaxViT as the backbone network yields significant improvements across all metrics compared to the original ResNet50. When employing MaxViT alone, OC Dice and OD Dice increase from 87.70% and 80.31% to 90.10% and 82.48%, respectively. Additionally, the mIoU, mPA, and overall accuracy improve by 2.14%, 1.80%, and 0.62%, respectively, thus demonstrating MaxViT's enhanced capability in feature extraction and segmentation performance. After further enhancements to MaxViT, the model performance is greatly boosted, with OC Dice and OD Dice reaching 93.32% and 88.09%, respectively, and the mIoU and accuracy rising to 88.18% and 97.15%, respectively, thus confirming the effectiveness of the improved strategy. Although the inference speed decreases slightly, with FPS dropping from 3.40 to 2.60, the segmentation gains justify this trade-off.

Table 2. Ablation study on the ORIGA dataset.

MaxViT	Improved MaxViT	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
		89.77±0.25	88.18±0.64	85.75±0.19	92.75±0.05	95.83±0.21	7.79±0.59
√		91.56±0.21	89.96±0.18	87.95±0.16	94.11±0.13	96.62±0.01	6.05±0.43
	√	93.23±0.05	92.00±0.05	90.17±0.02	94.69±0.07	97.33±0.01	5.86±0.14

On the ORIGA dataset, MaxViT alone improves OC Dice and OD Dice from 89.77% and 88.18% (ResNet50 baseline) to 91.56% and 89.96%, respectively. Concurrently, the mIoU, mPA, and accuracy increase by 2.20%, 1.36%, and 0.79%, respectively, thus illustrating MaxViT's robust feature extraction capabilities across different datasets. Following further improvements, the model achieves OC Dice and OD Dice of 93.23% and 92.00%, respectively with the mIoU and accuracy increasing to 90.17% and 97.33%, respectively. Despite a decrease in the inference speed (FPS from 7.79 to 5.86), the performance improvements, especially in OD Dice (3.82%) and accuracy (1.50%).

Table 3. Ablation study on the RIM-ONE DL dataset.

MaxViT	Improved MaxViT	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
		77.70±0.28	82.48±0.69	73.48±0.82	86.02±0.67	88.08±0.19	12.28±0.27
√		84.02±0.70	88.32±0.46	81.36±0.42	91.37±0.11	92.61±0.09	9.41±0.23
	√	86.12±0.04	90.11±0.03	83.99±0.03	92.09±0.01	93.98±0.04	9.09±0.44

On the RIM-ONE DL dataset, the MaxViT backbone alone provides significant enhancements over the ResNet50 baseline, increasing OC Dice from 77.70% to 84.02% and OD Dice from 82.48% to 88.32%. At the same time, the mIoU, mPA, and accuracy improve by 7.88%, 5.35%, and 4.53%, respectively, thus demonstrating MaxViT's powerful feature representation in handling diverse fundus images. A further optimization of MaxViT leads to OC Dice and OD Dice of 86.12% and 90.11%, respectively, with the mIoU and accuracy rising to 83.99% and 93.98%. Notably, the OD Dice surpasses 90%. While the inference speed decreases from 12.28 FPS to 9.09 FPS, the significant performance gains-particularly an 8.42% relative improvement in OC Dice-outweigh this reduction.

Table 4. Ablation study on the RIM-ONE-R3 dataset.

MaxViT	Improved MaxViT	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
		74.82±1.90	81.34±0.59	74.57±1.15	85.37±1.62	93.61±0.27	9.60±0.03
√		83.31±0.46	89.08±1.17	83.00±1.04	89.65±0.57	96.28±0.48	7.76±0.04
	√	85.50±0.55	89.93±0.86	84.69±0.87	90.45±1.02	96.64±0.30	7.32±0.18

On the RIM-ONE R3 dataset, replacing the ResNet50 backbone with MaxViT results in substantial improvements across all metrics. OC Dice increases from 74.82% to 83.31%, which is an 8.49 percentage point gain, while OD Dice rises from 81.34% to 89.08%. the mIoU and accuracy also improve by 8.43% and 2.67%, respectively, thus highlighting MaxViT's superior feature representation for fundus image segmentation. After further refinement, the model achieves OC Dice of 85.50%, OD Dice near 90% at 89.93%, and an accuracy of 96.64%. All metrics reach new highs, with the relative improvement in OC Dice reaching an impressive 14.27%, thus demonstrating the efficacy of the proposed improvements.

Table 5. Ablation study on the REFUGE dataset.

MaxViT	Improved MaxViT	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
		90.18±0.03	90.31±0.23	87.09±0.15	93.47±0.11	96.37±0.11	22.21±0.28
√		89.57±0.11	91.28±0.02	87.55±0.07	92.73±0.25	96.86±0.01	18.45±0.33
	√	92.22±0.01	92.83±0.01	90.03±0.01	94.41±0.01	97.41±0.01	16.99±0.02

On the REFUGE dataset, ablation studies confirm the model's strong performance on large-scale data. When using MaxViT alone, OC Dice slightly decreases from 90.18% to 89.57%, while OD Dice improves from 90.31% to 91.28%. The accuracy and mIoU increase by 0.49% and 0.46%, respectively, thus suggesting that MaxViT enhances the structural recognition while maintaining the overall performance. With further improvements, OC Dice and OD Dice reach 92.22% and 92.83%, respectively, the mIoU surpasses 90% at 90.03%, and accuracy rises to 97.41%. Although the inference speed drops from 22.21 FPS to 16.99 FPS, the segmentation performance remains excellent.

These experimental results clearly demonstrate the effectiveness of the MaxViT backbone: the model achieved notable improvements across all evaluation metrics compared to the original semantic FPN, thus confirming the superior capability of the MaxViT backbone in extracting structural details. The multi-axis attention mechanism in MaxViT enhances the model's ability to capture fine-grained features in fundus images, thereby improving the segmentation accuracy. Moreover, these experimental results clearly demonstrate the significant advantage of the improved MBConv module: OD/OC-semantic FPN achieved the best segmentation performance across all five fundus image datasets, especially in key metrics such as Dice scores, MIoU, and the overall classification accuracy. This strongly validates the effectiveness of the improved MBConv module. The incorporation of Inception DW Convolution successfully integrates multi-scale receptive fields, thereby enhancing the model's precision in capturing OC and OD boundary features.

Although the introduction of multi-scale convolutions results in a slight drop in FPS, the segmentation accuracy is typically more important than the processing speed in medical image analyses. For fundus image segmentation, precise delineation of the OC and OD is crucial for disease diagnosis. The excellent performance of the OD/OC-semantic FPN demonstrates its great potential in fundus image analyses. High-precision segmentation of the OC and OD can provide critical auxiliary information for the early diagnosis of ocular diseases such as glaucoma, thus aiding clinicians in formulating more effective treatment strategies.

4.4.2. Comparison study

In order to comprehensively evaluate the advanced features and practicality of the OD/OC-Semantic FPN, a systematic and in-depth comparative analysis was conducted with current mainstream image segmentation models. The selected comparison models include the following representative medical image segmentation models: U-Net [30]: a classic U-shaped CNN architecture, widely used in medical image segmentation due to its symmetric encoder-decoder design and skip connection mechanism; DeepLabV3+ [31]: incorporates dilated convolutions and the atrous spatial pyramid pooling (ASPP) module, thus providing strong multi-scale contextual modeling capabilities; PSPNet [32]: aggregates information from different receptive fields through the pyramid pooling module, thus enhancing the model’s ability to capture both global and local features; APCNet [33]: combines attention mechanisms with contextual information guidance to effectively enhance feature representation and improve the segmentation accuracy; Semantic FPN based on PoolFormer [34]: utilizes PoolFormer’s minimalist feature extraction combined with the FPN structure to achieve efficient semantic information modeling; and Attention U-Net [35]: integrates attention mechanisms into the U-Net architecture to improve the segmentation accuracy of the OC and OD by enhancing feature representation.

These models were all evaluated on the five representative public fundus image datasets (DRISHTI-GS, ORIGA, RIM-ONE DL, RIM-ONE-R3, and REFUGE) to ensure the comprehensiveness and objectivity of the experimental results. The quantitative performance comparisons of each model in the OC and OD segmentation tasks are shown in Tables 6–10, and the visual results of the segmentation are presented in Figures 8–12.

Table 6. Comparison of experiments on the Drishti-GS dataset.

Models	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
U-Net	84.34±0.14	77.86±0.21	78.00±0.07	87.08±0.15	94.36±0.01	3.90±0.03
DeepLabV3+	85.41±0.37	79.71±0.83	79.44±0.46	88.83±0.36	94.62±0.06	3.92±0.03
PSPNet	87.20±0.54	77.93±1.13	79.62±0.23	88.47±0.21	94.55±0.07	3.85±0.07
APCNet	89.62±0.23	79.81±0.62	81.43±0.13	89.46±0.15	95.07±0.03	3.11±0.01
Semantic FPN	87.68±0.01	80.26±0.02	80.73±0.02	89.40±0.02	94.97±0.01	3.71±0.01
Attention U-Net	89.00±0.06	80.19±0.20	81.35±0.09	90.69±0.10	95.02±0.06	4.10±0.06
OD/OC-Semantic FPN	93.32±0.05	88.09±0.53	88.18±0.04	93.69±0.18	97.15±0.02	2.60±0.30

Table 7. Comparison of experiments on the ORIGA dataset.

Models	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
U-Net	87.60±0.04	85.92±0.07	83.31±0.05	91.49±0.01	95.28±0.01	8.63±0.14
DeepLabV3+	91.16±0.06	88.44±0.05	86.62±0.07	93.58±0.05	96.01±0.01	9.30±0.01
PSPNet	90.36±0.03	88.04±0.17	85.94±0.07	92.52±0.04	95.85±0.03	8.44±0.04
APCNet	90.79±0.13	88.01±0.31	86.03±0.26	93.01±0.05	95.82±0.04	6.22±0.44
Semantic FPN	87.93±0.01	84.13±0.01	82.19±0.01	91.53±0.01	94.47±0.01	7.90±0.20
Attention U-Net	91.56±0.16	88.64±0.20	86.93±0.21	93.91±0.05	96.09±0.07	9.28±0.01
OD/OC-Semantic FPN	93.23±0.05	92.00±0.05	90.17±0.02	94.69±0.07	97.33±0.01	5.86±0.14

Table 8. Comparison of experiments on the RIM-ONE DL dataset.

Models	OC Dice(%)	OD Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
U-Net	73.31±0.21	78.06±0.31	69.07±0.46	82.48±0.46	85.72±0.38	13.40±0.04
DeepLabV3+	79.44±0.82	85.87±0.07	77.05±0.23	88.01±0.08	90.51±0.03	14.07±0.14
PSPNet	82.31±0.11	86.07±0.12	78.44±0.29	89.45±0.01	90.76±0.10	13.31±0.13
APCNet	81.58±0.31	86.43±0.31	78.44±0.61	89.06±0.05	91.12±0.25	10.06±2.61
Semantic FPN	74.21±2.26	76.14±0.03	66.81±0.08	83.50±0.23	83.19±0.11	11.83±0.31
Attention U-Net	77.31±0.46	84.64±0.03	75.54±0.05	85.89±0.17	90.06±0.15	14.30±0.02
OD/OC-Semantic FPN	86.12±0.04	90.11±0.03	83.99±0.03	92.09±0.01	93.98±0.04	9.09±0.44

Table 9. Comparison of experiments on the RIM-ONE-R3 dataset

Models	Cup Dice(%)	Disc Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
U-Net	60.21±2.70	77.38±0.47	66.78±1.21	76.29±1.30	91.86±0.28	10.54±0.17
DeepLabV3+	73.02±0.75	81.38±0.35	73.75±0.45	85.65±0.28	93.34±0.06	11.37±0.04
PSPNet	76.50±1.21	82.35±0.57	75.84±0.80	84.95±0.53	94.02±0.16	10.75±0.07
APCNet	74.26±0.07	81.97±0.00	74.72±0.04	85.01±0.60	93.85±0.08	8.55±0.06
Semantic FPN	75.62±0.16	82.64±0.16	75.62±0.04	84.68±0.03	94.05±0.06	9.54±0.04
Attention U-Net	70.84±0.05	77.93±0.31	70.96±0.17	83.41±0.48	92.19±0.09	10.32±0.05
OD/OC-Semantic FPN	85.50±0.55	89.93±0.86	84.69±0.87	90.45±1.02	96.64±0.30	7.32±0.18

Table 10. Comparison of experiments on the REFUGE dataset.

Models	Cup Dice(%)	Disc Dice(%)	MIoU(%)	MPA(%)	Accuracy(%)	FPS(img/s)
U-Net	88.77±0.94	89.53±0.47	85.93±0.82	91.11±1.23	96.26±0.20	27.63±0.24
DeepLabV3+	90.90±0.07	91.83±0.01	88.60±0.05	93.99±0.02	97.03±0.01	30.61±0.05
PSPNet	90.03±0.10	90.69±0.04	87.35±0.02	92.84±0.11	96.63±0.02	28.29±0.24
APCNet	89.85±0.12	90.54±0.10	87.12±0.01	92.58±0.16	96.54±0.02	23.94±0.15
Semantic FPN	90.33±0.00	91.19±0.01	87.88±0.01	93.05±0.01	96.83±0.01	22.14±0.12
Attention U-Net	89.46±0.18	90.25±0.16	86.76±0.18	93.08±0.09	96.48±0.04	26.67±0.10
OD/OC-Semantic FPN	92.22±0.01	92.83±0.01	90.03±0.01	94.41±0.01	97.41±0.01	16.99±0.02

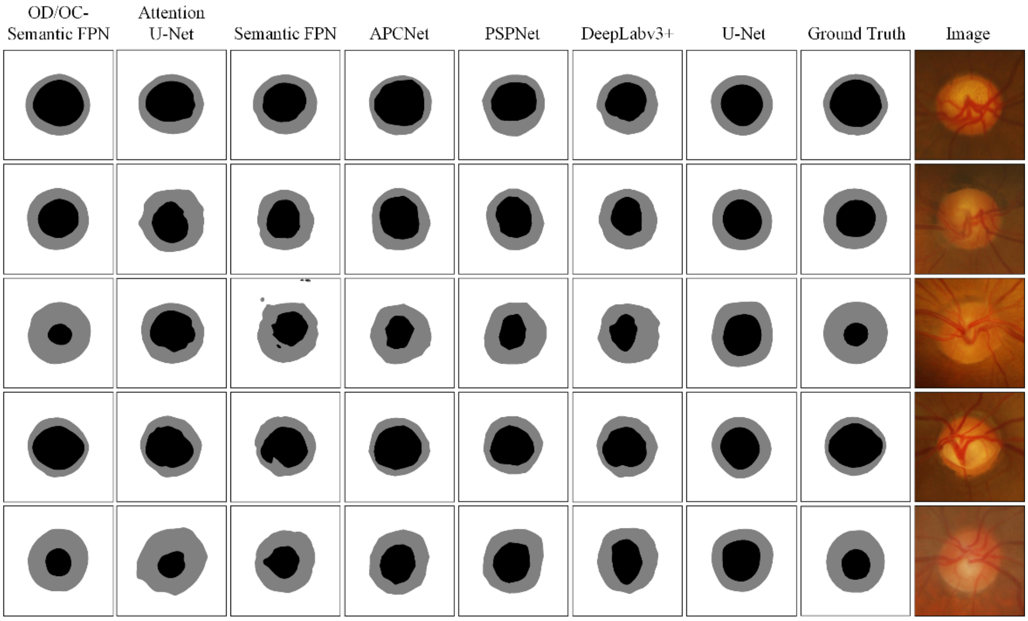

Figure 8. Segmentation results of different models on the Drishti-GS dataset.

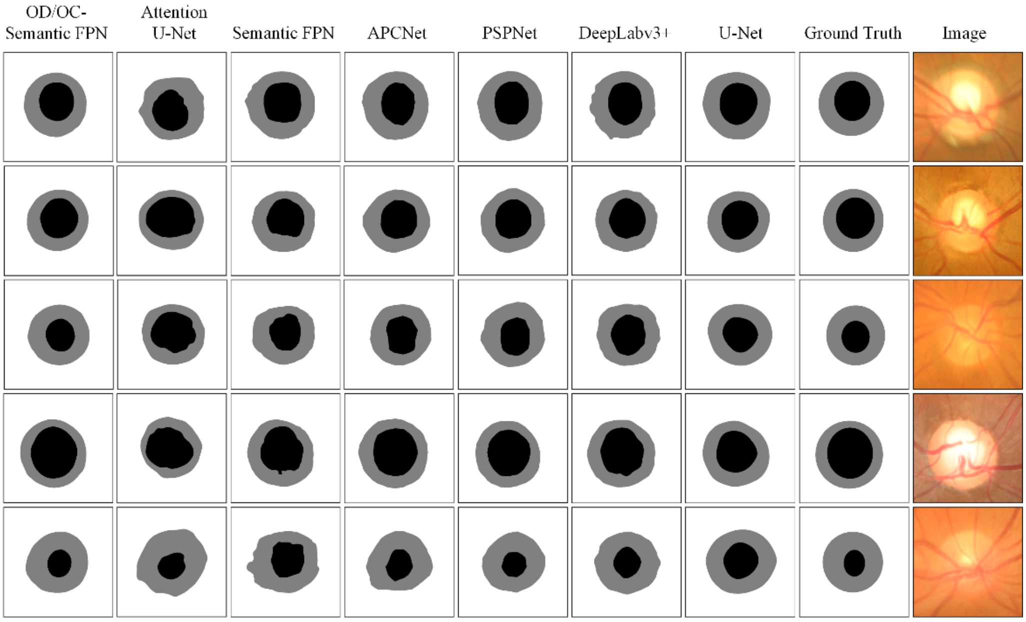

Figure 9. Segmentation results of different models on the ORIGA dataset.

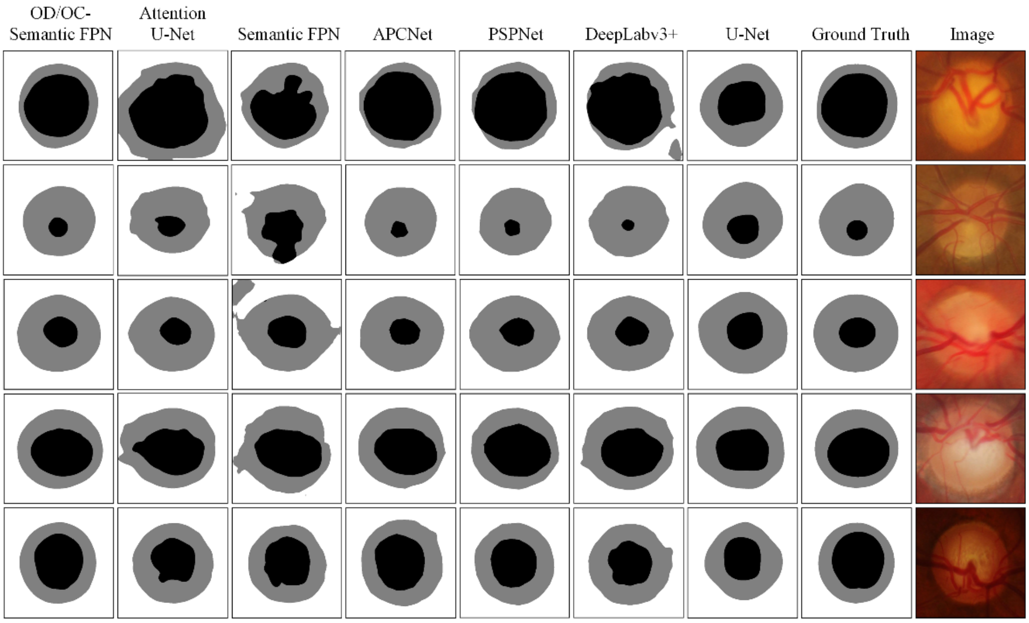

Figure 10. Segmentation results of different models on the RIM-ONE DL dataset.

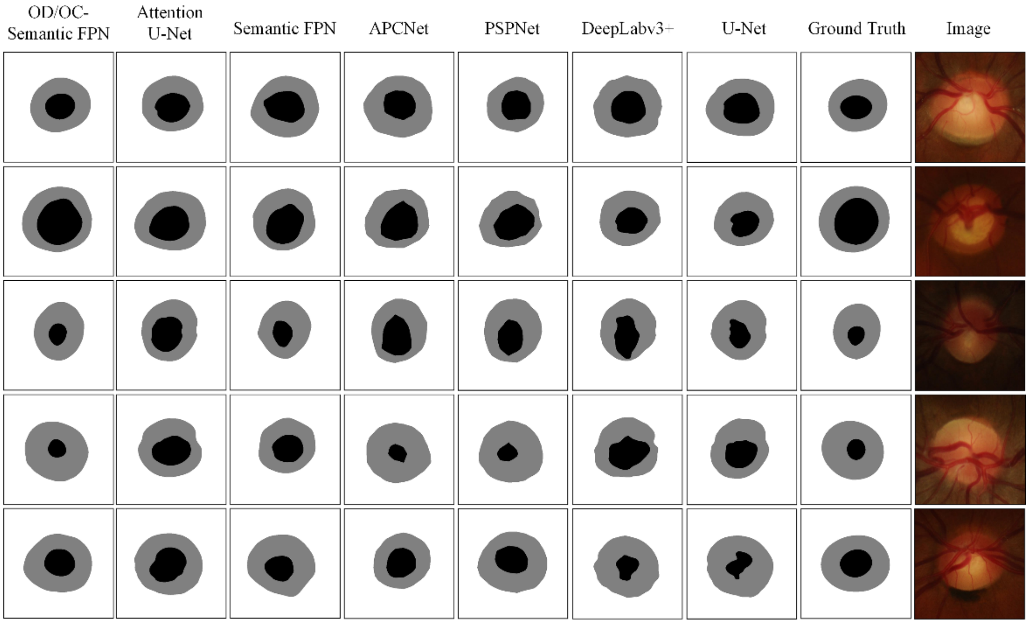

Figure 11. Segmentation results of different models on the RIM-ONE-R3 dataset.

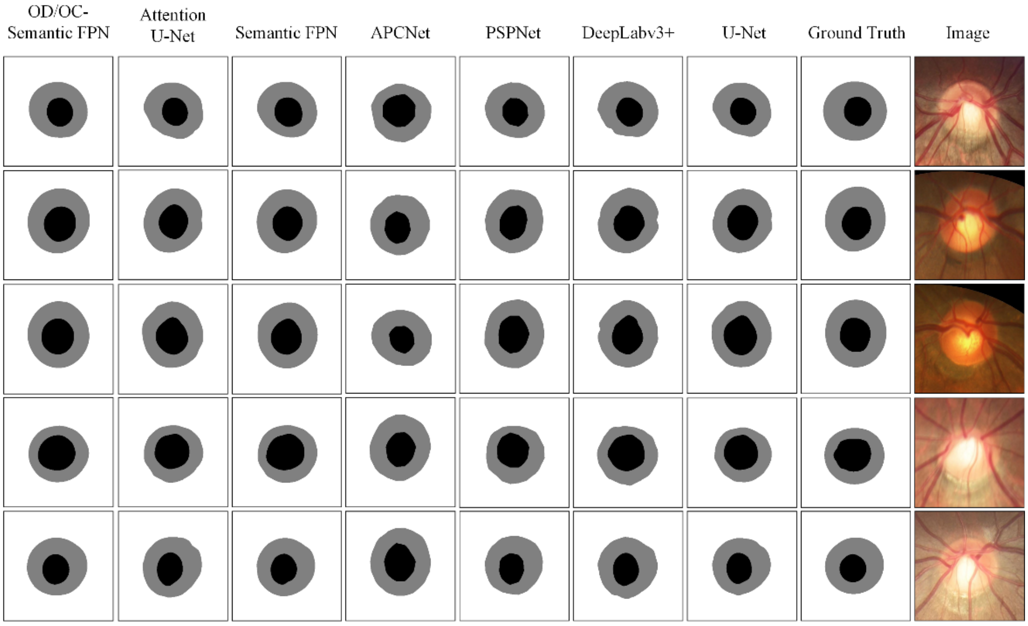

Figure 12. Segmentation results of different models on the REFUGE dataset.

The experimental results demonstrate that the OD/OC-Semantic FPN consistently achieves a state-of-the-art segmentation performance across all five benchmark datasets, significantly surpassing mainstream models such as U-Net, DeepLabV3+, PSPNet, APCNet, the PoolFormer-based Semantic FPN, and Attention U-Net. Across key evaluation metrics, including OC Dice, OD Dice, MIoU, MPA, and overall accuracy, the OD/OC-Semantic FPN exhibits superior performance, thus highlighting its strengths in fine structure recognition and precise boundary localization.

Although the proposed model shows a slight decrease in FPS compared to some lightweight architectures, its substantial gains in segmentation accuracy and structural completeness make it especially well-suited for clinical decision-support scenarios that demand high precision. In medical image analyses, particularly for tasks such as glaucoma screening and optic nerve assessment, the segmentation accuracy is typically prioritized over real-time performance, which makes this trade-off both acceptable and justifiable.

Furthermore, as shown in Figures 8–12, the segmentation maps generated by the OD/OC-Semantic FPN exhibit a high degree of consistency with the ground truth annotations. The model effectively delineates the boundaries of the OC and OD with smooth and continuous contours, thus markedly reducing both over-segmentation and under-segmentation artifacts. These visual results further validate the model’s robustness in maintaining the segmentation stability and structural integrity, even under complex imaging conditions, thus highlighting its strong potential for clinical application.

5. Discussion and conclusions

In this study, we proposed a joint segmentation model for OD and OC in fundus images to improve the accuracy and robustness for glaucoma screening. Building upon the Semantic FPN architecture, the model incorporates three key innovations: (1) MaxViT backbone integration, which employs a multi-axis attention mechanism to capture both local and global contextual information, thus preserving critical edge details and enhancing feature representation; (2) MBConv module enhancement, which employs Inception-style DW convolutions with parallel kernels of varying sizes to expand receptive field, thereby improving multi-scale feature extraction and boundary delineation, particularly in challenging imaging conditions; and (3) multi-scale feature fusion via semantic FPN's bottom-up and top-down pathways, thus enabling precise spatial alignment and strengthening boundary continuity and segmentation robustness. Extensive experiments on five public fundus image datasets demonstrated that these architectural enhancements yielded accurate and consistent segmentation results. Although the enhanced model incurs higher computational costs and reduced inference speed, it achieves state-of-the-art performance with superior generalizability across diverse datasets.

Despite these promising results, several critical challenges must be addressed to facilitate real-world clinical translation. A future study will prioritize the following directions:

(1) Enhancing real-time efficiency and deployment

Despite the great accuracy and generalization, the inference speed is limited by MaxViT and MBConv complexities. Future work will focus on architectural simplification, pruning, quantization, and hardware-aware optimization to achieve millisecond-level latency. Inspired by platforms such as GlaucoCare [35], efforts will aim to develop lightweight, end-to-end solutions compatible with electronic health records that are deployable on portable or edge devices in resource-constrained environments.

(2) Improving generalization through multimodal learning and few-shot adaptation

To enhance the model's generalization across diverse devices, populations, and pathologies, future work will explore multimodal fusion, few-shot learning with cycle-consistency to address annotation scarcity, task-aware meta-learning for rapid domain adaptation, cross-domain transferable few-shot models inspired by infrastructure diagnostics, anatomy-guided geometric regularization embedding OD/OC priors for physiological consistency, GAN-based data augmentation for generating diverse synthetic samples, and model distillation and compression to develop lightweight variants suitable for scalable edge deployment [36–38].

(3) Enhancing Clinical Interpretability and Fostering User Trust

Although the model demonstrates a high segmentation accuracy, its deep learning-based decision-making often appears opaque to clinicians, thus limiting a widespread clinical adoption. Future work is expected to incorporate visual explanation tools, such as attention heatmaps and Grad-CAM, to highlight key image regions and provide intuitive visual rationale for the model's outputs. Additionally, transparent and clinically meaningful reasoning pathways should be established by linking image features—such as lesion characteristics and structural boundaries—to segmentation results and their diagnostic significance in ophthalmology. By aligning the model's behavior with clinical reasoning, these advancements are anticipated to enhance user trust and confidence.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was generously supported by the grants from the National Natural Science Foundation of JiangSu (Grant No BK20221186 to J.Y.). We would like to thank the anonymous reviewers for their valuable suggestions.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. Shinde, Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms, *Intell. Med.*, **5** (2021), 100038. https://doi.org/10.1016/j.ibmed.2021.100038
2. L. Zhang, C. P. Lim, Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models, *Appl. Soft Comput.*, **92** (2020), 106328. https://doi.org/10.1016/j.asoc.2020.106328
3. D. Meedeniya, T. Shyamalee, G. Lim, P. Yogarajah, Glaucoma identification with retinal fundus images using deep learning: systematic review, *Inf. Med. Unlocked*, **56** (2025), 101644. https://doi.org/10.1016/j.imu.2025.101644
4. A. Haider, M. Arsalan, M. B. Lee, M. Owais, T. Mahmood, H. Sultan, et al., Artificial intelligence-based computer-aided diagnosis of glaucoma using retinal fundus images, *Exp. Syst. Appl.*, **207** (2022), 117968. https://doi.org/10.1016/j.eswa.2022.117968
5. A. Al-Mahrooqi, D. Medvedev, R. Muhtaseb, M. Yaqub, GARDNet: Robust multi-view network for glaucoma classification in color fundus images, in *Ophthalmic Medical Image Analysis: 9th International Workshop*, Springer, **13576** (2022), 152–161. https://doi.org/10.1007/978-3-031-16525-2_16
6. X. Bian, X. Luo, C. Wang, W. Liu, X. Lin, Optic disc and optic cup segmentation based on anatomy guided cascade network, *Comput. Methods Programs Biomed.*, **197** (2020), 105717. https://doi.org/10.1016/j.cmpb.2020.105717
7. T. Shyamalee, D. Meedeniya, Attention U-net for glaucoma identification using fundus image segmentation, in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, (2022), 6–10. https://doi.org/10.1109/DASA54658.2022.9765303
8. S. Bengani, J. A. A. Jothi, S. Vadivel, Automatic segmentation of optic disc in retinal fundus images using semi-supervised deep learning, *Multimed. Tools Appl.*, **80** (2021), 3443–3468. https://doi.org/10.1007/s11042-020-09778-6
9. Y. Jiang, L. Duan, J. Cheng, Z. Gu, H. Xia, H. Fu, et al., JointRCNN: A region-based convolutional neural network for optic disc and cup segmentation, *IEEE Trans. Biomed. Eng.*, **67** (2019), 335–343. https://doi.org/10.1109/TBME.2019.2913211
10. X. Yuan, L. Zhou, S. Yu, M. Li, X. Wang, X. Zheng, A multi-scale convolutional neural network with context for joint segmentation of optic disc and cup, *Artif. Intell. Med.*, **113** (2021), 102035. https://doi.org/10.1016/j.artmed.2021.102035
11. R. Bhattacharya, R. Hussain, A. Chatterjee, D. Paul, S. Chatterjee, D. Dey, PY-Net: Rethinking segmentation frameworks with dense pyramidal operations for optic disc and cup segmentation from retinal fundus images, *Biomed. Signal Process. Control*, **85** (2023), 104895. https://doi.org/10.1016/j.bspc.2023.104895
12. Y. Yi, Y. Jiang, B. Zhou, N. Zhang, J. Dai, X. Huang, et al., C2FTFNet: Coarse-to-fine transformer network for joint optic disc and cup segmentation, *Comput. Biol. Med.*, **164** (2023), 107215. https://doi.org/10.1016/j.compbiomed.2023.107215

13. Y. Chen, Z. Liu, Y. Meng, J. Li, Lightweight optic disc and optic cup segmentation based on MobileNetv3 convolutional neural network, *Biomimetics*, **9** (2024), 637. https://doi.org/10.3390/biomimetics9100637
14. T. Shyamalee, D. Meedeniya, CNN based fundus images classification for glaucoma identification, in *2022 2nd International Conference on Advanced Research in Computing (ICARC)*, IEEE, (2022), 200–205. https://doi.org/10.1109/ICARC54489.2022.9754171
15. M. Wassel, A. M. Hamdi, N. Adly, M. Torki, Vision transformers based classification for glaucomatous eye condition, in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, (2022), 5082–5088. https://doi.org/10.1109/ICPR56361.2022.9956086
16. A. Singh, S. Sengupta, V. Lakshminarayanan, Glaucoma diagnosis using transfer learning methods, in *Applications of Machine Learning, SPIE*, (2019), 27. https://doi.org/10.1117/12.2529429
17. M. Nawaz, T. Nazir, A. Javed, U. Tariq, H. S. Yong, M. A. Khan, et al., An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization, *Sensors*, **22** (2022), 434. https://doi.org/10.3390/s22020434
18. T. Shyamalee, D. Meedeniya, Glaucoma detection with retinal fundus images using segmentation and classification, *Mach. Intell. Res.*, **19** (2022), 563–580. https://doi.org/10.1007/s11633-022-1354-z
19. Á. S. Hervella, J. Rouco, J. Novo, M. Ortega, End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images, *Appl. Soft Comput.*, **116** (2022), 108347. https://doi.org/10.1016/j.asoc.2021.108347
20. H. Wang, H. Sun, Y. Fang, S. Li, M. Feng, R. Wang, A workflow for computer-aided diagnosis of glaucoma, in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, IEEE, (2022), 1–4. https://doi.org/10.1109/ISBIC56247.2022.9854585
21. Y. Xu, C. Zhang, H. Li, Transformer-based large vision model for universal structural damage segmentation, *Autom. Constr.*, **176** (2025), 106256. https://doi.org/10.1016/j.autcon.2025.106256
22. Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, et al., Maxvit: Multi-axis vision transformer, in *Computer vision–ECCV 2022: 17th European Conference*, Springer, **13684** (2022), 459–479. https://doi.org/10.1007/978-3-031-20053-3_27
23. W. Yu, P. Zhou, S. Yan, X. Wang, InceptionNeXt: When inception meets ConvNeXt, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2024), 5672–5683. https://doi.org/10.1109/cvpr52733.2024.00542
24. A. Kirillov, R. Girshick, K. He, P. Dollar, Panoptic feature pyramid networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2019), 6392–6401. https://doi.org/10.1109/CVPR.2019.00656
25. J. Sivaswamy, S. R. Krishnadas, G. D. Joshi, M. Jain, A. U. S. Tabish, Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation, in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE, (2014), 53–56. https://doi.org/10.1109/ISBI.2014.6867807
26. Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, et al., Origa-light: An online retinal fundus image database for glaucoma analysis and research, in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, (2010), 3065–3068. https://doi.org/10.1109/IEMBS.2010.5626137

27. F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayón, R. Arnay, D. Angel-Pereira, Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning, *Image Anal. Stereol.*, **39** (2020), 161–167. https://doi.org/10.5566/ias.2346
28. F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, RIM-ONE: An open retinal image database for optic nerve evaluation, in *2011 24th International Symposium on Computer-based Medical Systems (CBMS)*, IEEE, (2011), 1–6. https://doi.org/10.1109/CBMS.2011.5999143
29. J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, et al., REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, *Med. Image Anal.*, **59** (2020), 101570. https://doi.org/10.1016/j.media.2019.101570
30. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Lecture Notes in Computer Science*, Springer, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
31. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Computer Vision—ECCV 2018*, Springer, **11211** (2018), 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
32. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2017), 6230–6239. https://doi.org/10.1109/CVPR.2017.660
33. J. He, Z. Deng, L. Zhou, Y. Wang, Y. Qiao, Adaptive pyramid context network for semantic segmentation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2019), 7511–7520. https://doi.org/10.1109/CVPR.2019.00770
34. W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, et al., MetaFormer is actually what you need for vision, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2022), 10809–10819. https://doi.org/10.1109/cvpr52688.2022.01055
35. T. Shyamalee, D. Meedeniya, G. Lim, M. Karunarathne, Automated tool support for glaucoma identification with explainability using fundus images, *IEEE Access*, **12** (2024), 17290–17307. https://doi.org/10.1109/ACCESS.2024.3359698
36. Y. Fan, H. Li, Y. Bao, Y. Xu, Cycle-consistency-constrained few-shot learning framework for universal multi-type structural damage segmentation, *Struct. Health Monit.*, **2024** (2024), 14759217241293467. https://doi.org/10.1177/14759217241293467
37. Y. Xu, Y. Fan, Y. Bao, H. Li, Task-aware meta-learning paradigm for universal structural damage segmentation using limited images, *Eng. Struct.*, **284** (2023), 115917. https://doi.org/10.1016/j.engstruct.2023.115917
38. Y. Xu, Y. Fan, Y. Bao, H. Li, Few-shot learning for structural health diagnosis of civil infrastructure. *Adv. Eng. Inf.*, **62** (2024), 102650. https://doi.org/10.1016/j.aei.2024.102650