Electronic Research Archive

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

ERA, 33(4): 2216–2245.
DOI: 10.3934/era.2025097
Received: 20 November 2024
Revised: 10 March 2025
Accepted: 02 April 2025
Published: 16 April 2025

Research article

DST-Net: Dual self-integrated transformer network for semi-supervised segmentation of optic disc and optic cup in fundus image

Yanxia Sun[1], **Tianze Xu**[2], **Jing Wang**[1] **and Jinke Wang**[1,2,*]

[1] Weihai Research Institute, Harbin University of Science and Technology, Weihai 264300, China
[2] School of Automation, Harbin University of Science and Technology, Harbin 150080, China

* **Correspondence:** Email: jkwang@hitwh.edu.cn.

Abstract: Accurate and efficient optic disc and cup segmentation from fundus images is significant for glaucoma screening. However, current neural network-based optic disc (OD) and optic cup (OC) segmentation tend to prioritize the image's local edge features, thus limiting their capacity to model long-term relationships, with errors in delineating the boundaries. To address this issue, we proposed a semi-supervised dual self-integrated transformer network (DST-Net) for joint segmentation of the OD and OC. First, we introduce a dual-view co-training mechanism to construct the encoder and decoder of the self-integrated network from the mutually enhanced feature learning modules of Vision Transformer (ViT) and convolutional neural networks (CNN), which are co-trained with dual views to learn the global and local features of the image adaptively. Moreover, we employ a dual self-integrated teacher-student framework, effectively utilizing large amounts of unlabeled fundus images through semi-supervised learning, thereby refining OD and OC segmentation results. Finally, we use a boundary difference over union loss (BDoU-loss) to optimize boundary prediction further. We implemented the comparative experiments on the publicly available dataset RIGA+. The OD and OC Dice values of the proposed DST-Net reached 95.12 ± 0.14 and 85.69 ± 0.27, respectively, outperforming other state-of-the-art (SOTA) methods. In addition, DST-Net shows strong generalization on the DRISHTI-GS1 and RIM-ONE-v3 datasets, proving its promising prospect in OD and OC segmentation.

Keywords: OD and OC; segmentation; CNN; transformer; semi-supervised

1. Introduction

Glaucoma is one of the most severe causes of blindness worldwide. It is anticipated that by 2040, the number of individuals diagnosed with glaucoma will reach 110 million [1]. The conventional method for localizing and segmenting retinal images relies on manual expertise and is susceptible to subjective and objective variables. This process is also time-consuming and prone to errors, particularly in detecting subtle lesions. The challenge is further compounded by variations in image quality, illumination, and anatomical differences across patients, which hinder the robustness of traditional segmentation methods. Consequently, developing efficient automated glaucoma detection technology is crucial for large-scale screening [2]. Clinically, the cup-to-disc ratio (CDR) is a pivotal diagnostic and monitoring indicator for glaucoma [3]. Accurate segmenting of the OD and OC in the retinal image is critical. Under normal conditions, the CDR is small. However, once the CDR increases, the risk of developing glaucoma increases, requiring further examination and treatment. Deep learning methods, especially convolutional neural networks (CNNs), have been widely used in several image segmentation tasks and have demonstrated satisfying performance [4,5]. For example, Guo et al. [6] proposed the CAFR-CNN framework for cross-domain joint OD and OC segmentation, which consists of faster R-CNN detectors, a spatial attention-based region alignment module, a pyramid ROI alignment module, and a prototype-based semantic alignment module. Luo et al. [7] proposed a new segmentation architecture, BGA-Net, which introduces auxiliary boundary branching and adversarial learning to segment OD and OC in a multi-labeled manner jointly. Yin et al. [8] proposed a deep learning-based level set approach for OD and OC segmentation in an automated retinal diagnostic system. Nevertheless, the CNN-based methods still face challenges in modeling long-range relationships due to the inherent limitations of convolutional operations [9].

With the breakthrough of the Transformer architecture, semantic segmentation has ushered in a new development opportunity [10]. Dosovitskiy et al. [11] proposed a Vision Transformer (ViT) in 2020, which successfully extends the Transformer model to image processing tasks, and realizes the integration of natural language processing and computer vision. ViTs break through the limitation of traditional CNNs on the input image size by serializing image pixels and capturing the global features of an image using the Transformer's self-attention mechanism. It also demonstrates excellent performance on image classification tasks. Wu et al. [12] proposed MedSegDiff-V2, a diffusion-based medical image segmentation framework incorporating Transformer modules to enhance feature representation and improve segmentation accuracy. Chen et al. [13] proposed the Laplacian-guided Hierarchical Transformer, a network for medical image segmentation that enhances segmentation accuracy by leveraging a Laplacian-guided multi-level feature extraction mechanism to better preserve structural details. Despite achieving promising segmentation performance on single datasets, these methods often do not perform prominently enough when tested on images from different datasets. Recent studies have explored advanced data augmentation, unsupervised learning, and zero-shot segmentation techniques to enhance medical image segmentation models' generalization ability and robustness. Goceri [14] proposed various medical image data augmentation techniques, including generative adversarial networks (GANs), contrastive learning, and deformation-based methods, to enhance data diversity and improve model robustness. Wang et al. [15] proposed Fourier Visual Prompting (FVP). This source-free, unsupervised domain adaptation method leverages Fourier-based perturbations to generate domain-invariant prompts, effectively enhancing medical image segmentation across domains. Shi et al. [16] adapted the segment anything model (SAM) for zero-shot medical image segmentation by integrating prompt-based fine-tuning and domain-specific adaptations.

It is important to note that ViTs alone does not yield good semantic segmentation results because it emphasizes global information and lacks local detail extraction. In contrast, CNNs focus on extracting local information, and thus, combining ViTs and CNNs enables the model to leverage local and global information in images, improving its performance and making it suitable for OD and OC segmentation tasks. However, ViTs typically require a large amount of data for practical training. Still, fundus image samples are sparse, with many unlabeled data and few labeled data, and the boundary segmentation of the optic cup and the optic disc is not comprehensive and complete enough. Although data augmentation can be beneficial, it alone may not be sufficient to address these challenges fully.

To address the above limitations, we developed a dual self-integrated transformer network, DST-Net, which combines a CNN and a ViT for joint segmentation of the OD and OC using a semi-supervised approach. The main contributions of this paper are as follows:

- A dual self-integrated teacher-student network is constructed using a CNN and a ViT, with weight updates performed via the exponential moving average (EMA), which gives higher weights to recent data and improves the stability and robustness of the model. Through a dual-view co-training mechanism, the network fully leverages the mutual learning capabilities of CNNs and ViTs to learn both global and local image information adaptively.
- Aiming at the problem of unclear boundaries of the OD and OC, the boundary difference loss function BDoU-loss is employed to enhance the network's attention to the boundaries, improving the accuracy of boundary segmentation.
- A semi-supervised learning approach utilizes labeled and unlabeled data for training, in which the teacher network generates pseudo-labels to guide the student network, enhancing the model's learning process.

2. Related works

This section discusses related work in CNN-based and Transformer-based methods.

2.1. CNN-based methods

With the rapid development of deep learning technology, CNN-based models have been thoroughly studied in medical image processing. For example, Zilly et al. [17] enhanced the performance of CNNs using an integrated learning technique, which employs an entropy sampling technique to identify the key information points in the image. Wang et al. [18] proposed an asymmetric segmentation network based on the U-Net model to segment the OD region, which combines the classical U-Net architecture with a unique cross-linked subnetwork to accurately localize the OD and improve sensitivity to morphological variations in ROI. Tulsani et al. [19] proposed an improved UNET++ architecture for automatically segmenting the OD and OC in glaucoma assessment, in which a clinical feature-based classification model with preprocessing and customized loss functions was introduced to address the problem of vessel and category imbalance. Pachade et al. [20] proposed a novel segmentation network called NENet, which combines EfficientNetB4 as an encoder and a series of pre-activated residual blocks, atrous space pyramid pooling (ASPP) blocks, and attention gates (AGs). Meanwhile, a modified patch-based discriminator was designed for NENet to improve local segmentation details. Guo et al. [21] proposed a joint segmentation method for OD and OC based on an improved U-Net architecture. The feature fusion module in U-Net is first added to reduce the information loss in feature extraction. Then, the channel and spatial attention mechanisms are combined to highlight the essential features related to segmentation and suppress the expression of features in irrelevant regions. Finally, multilabel loss is used to generate joint segmentation results. Experimental results show the method performs well on OD and OC segmentation tasks. Fu et al. [22] proposed the M-Net based on a multilabel deep network with polar coordinate transformation to segment OD and OC jointly. The model uses U-Net as the backbone network, the input layer as a pyramid model for receiving feature mappings at different scales, and the side output layer outputs the corresponding local prediction results. Zhu et al. [23] proposed an ultrasound image segmentation network called DBUNet with a dual-branch structure to address noise, artifacts, and feature fusion issues in ultrasound image segmentation. Moreover, in brain image segmentation, they further proposed a brain tumor segmentation method based on the fusion of deep semantic and edge information from multimodal MRI. The segmentation performance was effectively improved by designing multiple functional modules called ESAB [24]. Fu et al. [25] proposed an automatic optic disc segmentation method that combines U-Net with the probabilistic bubble model, effectively solving the problem of interference in optic disc segmentation in abnormal fundus images.

Currently, most CNN-based methods, such as U-Net and its variants, rely on extracting and fusing multiscale features to improve segmentation accuracy. However, the fused features still have a small "effective receptive field" and are mainly concentrated in localized regions of the image, thus limiting their performance.

2.2. Transformer-based methods

With the emergence of the Transformer, semantic segmentation entered a brand-new era. Chen et al. [9] proposed a TransU-Net for medical image segmentation by combining Transformer and U-Net. Cao et al. [26] designed a pure Transformer similar to U-Net based on the translation window mechanism, Swin-U-Net, for medical image segmentation. Li et al. [27] proposed a Transformer-based medical image segmentation model, Segtran. The model combines the Transformer's unrestricted sensory field at high feature resolution and the advantages of multiscale feature extraction using a compressed attentional block specification of the Transformer's self-attention mechanism. It also learns diverse representations using extended blocks. In addition, the method employed a new positional coding scheme that imposes a continuum induction bias on the image. Experimental results show that, compared to existing representative methods, Segtran achieves higher segmentation accuracy in OD and OC segmentation tasks and demonstrates good cross-domain generalization capabilities. Zhu et al. [28] proposed a method for multimodal spatial information enhancement and boundary shape correction, which consists of the Modal Information Extraction (MIE), Spatial Information Enhancement (SIE), and Boundary Shape Correction (BSC) modules. The MIE module processes the input data, the SIE module is embedded in the backbone network, and the BSC module is used for boundary correction. Together, they form a 3D brain tumor segmentation model. This method effectively solves the challenges in brain tumor segmentation and improves the segmentation accuracy. Moreover, Zhu et al. [29] proposed a lightweight medical image segmentation network (LMIS). By integrating grouped convolution and MobileViT, this network reduces the number of model parameters, effectively addressing the problem of limited computational resources in medical image segmentation. As a result, the model can maintain segmentation accuracy while reducing its parameter count. Zhu et al. [30] proposed SDV-TUNet, which combines the sparse dynamic encoder-decoder module and multi-level edge feature fusion (MEFF) module to effectively integrate global spatial information with local edge details, thereby improving the accuracy of MRI brain tumor segmentation. Fu et al. [31] used modules, such as the TCA module and SCFF to calculate the spatial relative positions and capture relative attention, which enhances the spatial information and improves the segmentation ability. Yi et al. [32] proposed a deep learning model for joint OD and OC segmentation, C2FTFNet (Coarse-to-Fine transformer network). The model employs a coarse-to-fine strategy to automatically segment OD and OC in fundus images by gradually increasing the segmentation accuracy. Hussain and Basak [33] proposed a new segmentation method called UT-Net, which exploits the advantages of the UNet and the Transformer in the coding layer and employs an attention-gated bilinear fusion scheme. In addition, multi-head contextual attention is introduced to augment the self-attention used in the traditional visual Transformer. Wu et al. [34] proposed a novel Transformer-based model, SeATrans, to transfer segmentation knowledge to disease diagnosis networks. Specifically, an asymmetric multiscale interaction strategy is first proposed to associate each low-level diagnostic feature with a multiscale segmentation feature. Then, an efficient strategy called SeA-block is employed to activate the diagnostic features through the associated segmentation features. In addition, some studies [35,36] have also explored lightweight Transformers, such as Swin-Transformer v2 and the Lite Vision Transformer to reduce computational complexity.

Although the Transformer architecture performs well in capturing global contextual information, it is inadequate in capturing local spatial information. Besides, transformer-based models usually require much data to train their self-attention mechanism for optimal performance. In addition, there is often a lack of high-quality labeled datasets in medical image segmentation, which limits the model's training effectiveness and generalization ability. Therefore, optimizing the Transformer model under limited data conditions to balance the ability to capture global and local features remains a challenge in current research.

3. Methods

3.1. Proposed DST-Net

The proposed DST-Net combines the strengths of CNNs and Transformers through dual-view co-training, and adaptively learns an image's global and local features to achieve high-accuracy segmentation (The code is publicly available at https://github.com/ky120/DST-Net). The network structure of DST-Net is depicted in Figure 1. It is a semi-supervised, self-integrated segmentation network that consists of two self-integrated networks. However, the two networks have different purposes and are trained using different loss functions. Each self-integrated network contains a student network, as shown in the upper half of Figure 1, and a teacher network, as shown in the lower half of Figure 1. The teacher network has the same structure as the student network. Still, its weights are updated as the EMA of the student network's weights [37–39]. Structurally, DST-Net builds a dual self-integration framework to learn more discriminative features. The first self-integration network consists of a U-shaped network composed of CNNs for obtaining the initial segmentation results of the OD and OC, and the second self-integration network consists of a U-shaped structured network composed of two ViTs for more accurate segmentation results. The parameters of the student network in each self-integration framework are passed to the teacher network through the EMA, which in turn updates the weights of the parameters in the teacher network.

As shown in Figure 1, the network training process is divided into two parts. First, the labeled source domain data is utilized for training, and the student network performs feature extraction by the CNN and ViT to generate the predicted feature map. Then, it is compared with the ground truth, and the loss is calculated. Since the teacher network is pre-trained with a small amount of data to provide a roughly correct segmentation trend, the Softmax function is used as the loss function, which is noted as *L1*. The teacher network generates pseudo-labels using labeled data, compares them to the ground truth, and calculates the loss. At the same time, the student network also uses the same data. Both networks are pre-trained using the *L2* mean squared error (MSE) as the loss function, which is *L2*. Finally, the predicted feature map generated by the student network is compared with the pseudo-label generated by the teacher network. The loss function named BdoU-Loss, which has the best performance in boundary extraction, is used, and the calculated loss is denoted as *L3*. Since the parameter update of the student network is ahead and has a high degree of uncertainty, the parameter update of the teacher network lags slightly behind that of the student network, and its output results are more stable. During the fusion process, the proportion of the teacher network should be marginally more significant than that of the student network. In the final process of determining the parameter update of the student network by comparing the student network's output with the ground truth, the fused loss function is shown in Eq (1),

$$L_{fusion} = \alpha L_1 + \beta L_2 + \gamma L_3. \tag{1}$$

During the experiment, it is finally determined that setting $\alpha = 0.2$, $\beta = 0.5$, and $\gamma = 0.3$ can achieve the ideal effect.

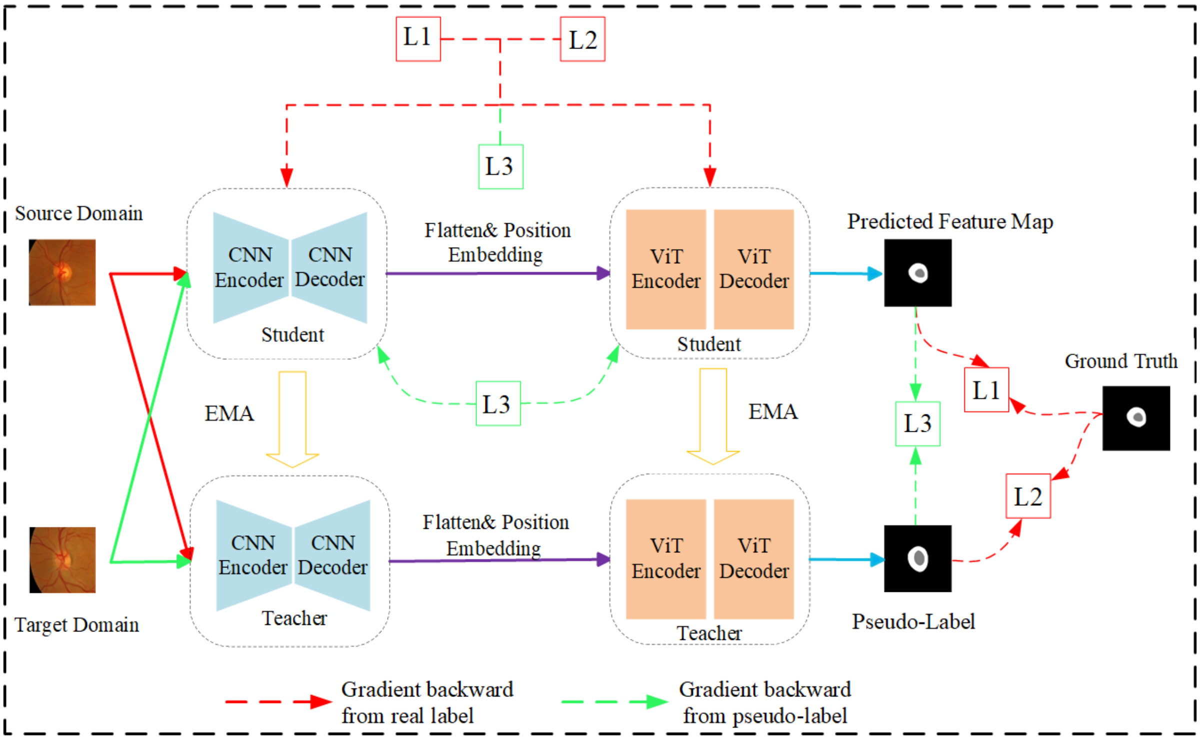
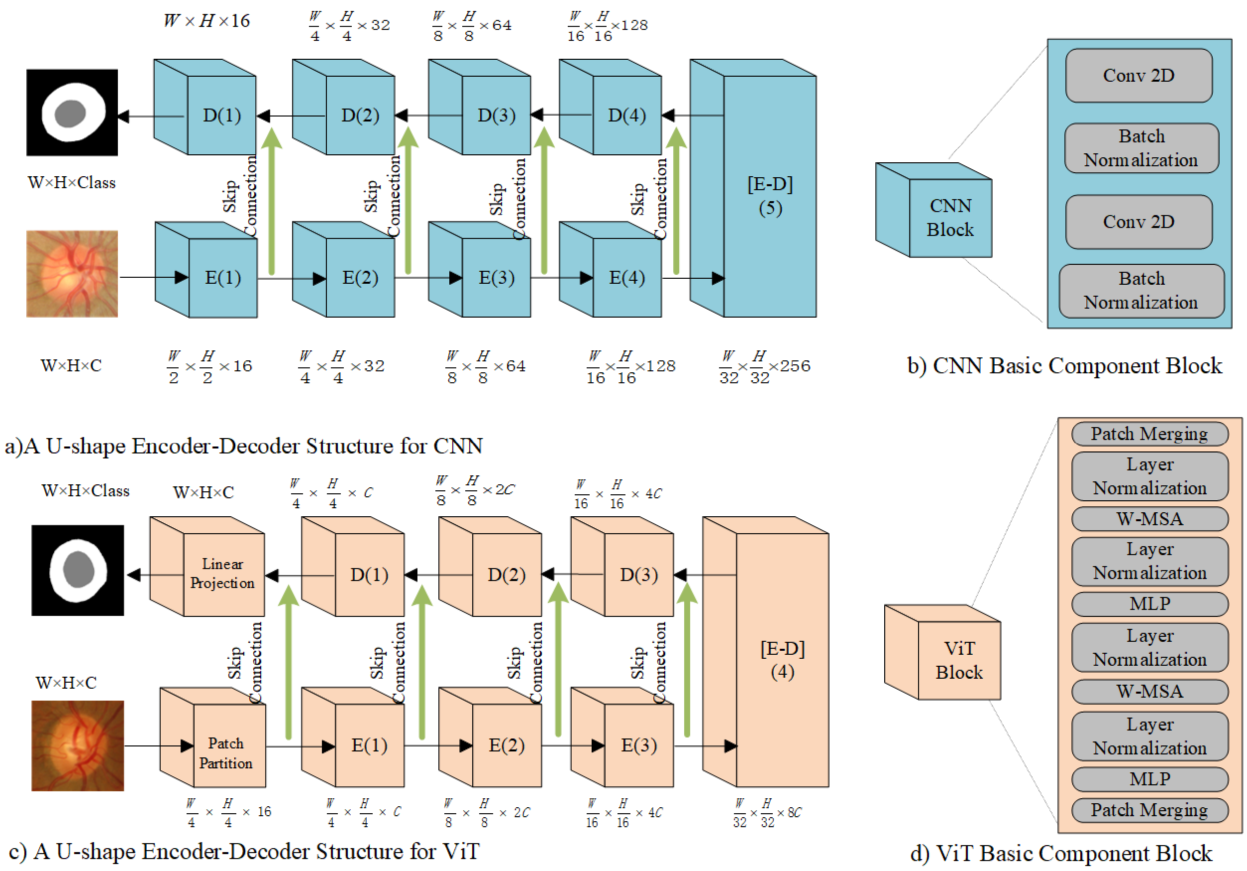


Figure 1. DST-Net network architecture.

The weights are passed through the EMA to update the entire network, and the loss functions are all chosen as the BDoU-loss. Second, after the training using labeled data is completed, the unlabeled data is fed into the teacher network to generate pseudo-labels. The loss is computed with the predicted feature maps generated by the student network, and the loss is passed to the student network for weight updating and then passed to the teacher network via EMA. To ensure that the teacher network can effectively guide the student network, the teacher network is pre-trained using ImageNet [40].

3.2. CNN and ViT-based encoder-decoder structure

To fully utilize the feature learning capabilities of CNN and ViT, this section constructs the encoder and decoder of the self-integrated network based on the CNN and the network blocks of ViT, respectively. The four encoders and decoders are connected using U-Net skip connections (shown in Figure 2).



$W \times H \times 16$
$\frac{W}{4} \times \frac{H}{4} \times 32$
$\frac{W}{8} \times \frac{H}{8} \times 64$
$\frac{W}{16} \times \frac{H}{16} \times 128$
D(1)
D(2)
D(3)
D(4)
[E-D] (5)
W×H×Class
Skip Connection
E(1)
E(2)
E(3)
E(4)
W×H×C
$\frac{W}{2} \times \frac{H}{2} \times 16$
$\frac{W}{4} \times \frac{H}{4} \times 32$
$\frac{W}{8} \times \frac{H}{8} \times 64$
$\frac{W}{16} \times \frac{H}{16} \times 128$
$\frac{W}{32} \times \frac{H}{32} \times 256$
CNN Block
Conv 2D
Batch Normalization
Conv 2D
Batch Normalization
W×H×Class
W×H×C
$\frac{W}{4} \times \frac{H}{4} \times C$
$\frac{W}{8} \times \frac{H}{8} \times 2C$
$\frac{W}{16} \times \frac{H}{16} \times 4C$
Linear Projection
D(1)
D(2)
D(3)
[E-D] (4)
W×H×C
Patch Partition
E(1)
E(2)
E(3)
$\frac{W}{4} \times \frac{H}{4} \times 16$
$\frac{W}{4} \times \frac{H}{4} \times C$
$\frac{W}{8} \times \frac{H}{8} \times 2C$
$\frac{W}{16} \times \frac{H}{16} \times 4C$
$\frac{W}{32} \times \frac{H}{32} \times 8C$
ViT Block
Patch Merging
Layer Normalization
W-MSA
Layer Normalization
MLP
Layer Normalization
W-MSA
Layer Normalization
MLP
Patch Merging

a)A U-shape Encoder-Decoder Structure for CNN

b) CNN Basic Component Block

c) A U-shape Encoder-Decoder Structure for ViT

d) ViT Basic Component Block

Figure 2. CNN and ViT-based encoding and decoding structure.

Figure 2a) shows the structure of the CNN-based network, and Figure 2c) shows the structure of the ViT-based network. In each CNN-based block, two 3 × 3 convolutional layers and two batch normalizations are used to build the structure, as shown in Figure 2b). The ViT-based network blocks are constructed using Swin-Transformer, as shown in Figure 2d). Unlike conventional Transformer blocks, Swin-Transformer employs layer normalization with offset windows, multi-head self-attention, residual connectivity, and multilayer perceptrons (MLPs) with Gaussian error linear units (GELUs) with a moving window design. This design forms the windows multi-head self-attention (WMSA) and shifted window-based multi-head self-attention (SWMSA) mechanisms. WMSA is a windowed multi-head self-attention mechanism that handles the computation of self-attention within each window. WMSA reduces computational complexity by dividing the input features into multiple fixed-size windows and performing local self-attention calculations within the windows. SWMSA is an improved version of WMSA. It adds a shifted windowing mechanism based on window division to compensate for the lack of information interaction between windows in WMSA. SWMSA integrates layer normalization, multi-head self-attention, and residual connection to better capture the local features of an image. Through the size of the window and the sliding step, the Swin-Transformer moves over the data sequence and performs the computation. Precisely, a sliding window moves across the data sequence, selecting a continuous data segment as the window at each step. It analyzes or computes the data within the window. The window is then slid forward in fixed steps, and the process continues until the entire data sequence is covered. WMSA and SWMSA are applied in two consecutive Transformer blocks, respectively. In this way, the Transformer block maps the position of the input sequence $Z_0 = [Z_{0,1} \dots, Z_{0,N}]$ to $Z_L = [Z_{L,1} \dots, Z_{L,N}]$. The detailed process of feature learning of the data through the ViT network based on WMSA, SWMSA, and MLP is summarized in Eqs (2)–(6), where *i* = [1, L] and L is the number of blocks.

$$Z_i = WMSA\big(LN(Z_{i-1})\big) + Z_{i-1}, \tag{2}$$

$$Z_{i+1} = MLP\big(LN(Z_i)\big) + Z_i, \tag{3}$$

$$Z_{i+2} = SWMSA\ \big(LN\ (Z_{i+1})\big) + Z_{i+1}, \tag{4}$$

$$Z_{i+3} = MLP\ \big(LN(Z_{i+2})\big) + Z_{i+2}, \tag{5}$$

$$WMSA(Z^*) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V, \tag{6}$$

where $Q, K, V \in R^{M^2 \times d}$, M^2 denotes the number of image blocks in a window, and d represents the dimension of the query and key.

The self-attention mechanism consists of three point-wise linear layers that map the input tokens into intermediate representations. It then computes the significance of each element in the sequence relative to every other element. In this way, a query Q, key K, and value V are generated for each component. Then, a weighted summation is performed to obtain a self-attention representation for each element. This self-attention representation determines how much attention each element pays to the other elements. Subsequently, these self-attention representations undergo operations such as normalization and residual connections before being passed into a feedforward neural network. In the feedforward network, each element undergoes point-wise linear transformations and is enhanced through nonlinear transformations to improve its features. Finally, the output from the feedforward network is combined with the initial input through another residual connection, forming a transformed sequence that represents the output of the Transformer module. This process is repeated across multiple Transformer modules to capture hierarchical representations and dependencies within the input sequence.

Unlike the basic blocks of CNNs, ViT is designed with merge and expansion layers between the encoder or decoder of each base block instead of the traditional subsampling and upsampling steps. This design allows information to flow more freely between layers, preserving more details and facilitating the better capture of complex relationships in the input sequence. The Merge layer is designed to halve the number of tokens and double the feature dimensions. It divides the input patch into four parts, then joins them together, and finally applies a linear layer to unify the dimensions by a factor of two. The expansion layer is designed to resize the input feature mapping to twice its original size and reduce the feature dimension to half the dimension of the input feature mapping. It uses a linear layer to increase the feature dimension, then employs a rearrangement operation to enlarge the size and reduce the feature dimension to a quarter of the input dimension. The size of the feature mapping in each step is briefly illustrated in Figure 2c), where W, H, and C represent the width, height, and channel dimensions of the feature map, respectively. The patch size is set to 4, the input channel is 3, the embedding dimension is 96, and the number of self-attention heads is 3, 6, 12, and 24, respectively, with a window size of 7 for each encoder and decoder. This design enables the ViT to capture the features and relationships within the input sequence more effectively while avoiding the subsampling and upsampling steps in traditional CNNs, thus improving model performance and efficiency.

3.3. Exponential moving average (EMA)

EMA is commonly used to smooth time series data. The principle is to perform a weighted average of the series data, giving more weight to recent data points and less to earlier data points. It smoothes the time series data effectively, making it more continuous and stable. In deep learning, EMA is commonly used to smooth the update of model parameters. Specifically, EMA is applied to the model parameters each time they are updated, thus reducing the fluctuation of each update and making the model more stable. The formula for EMA is shown in Eq (7):

$$EMA[t] = \alpha^{*}x[t] + (1-\alpha)^{*}EMA[t-1], \tag{7}$$

where t represents the time step, $x(t)$ denotes the original data at the t-th time point, α is the smoothing factor, which usually takes a value between 0 and 1, indicating the weight of the current sample, $(1-\alpha)$ represents the weight of the historical data, and $EMA[t-1]$ is the EMA value of the previous time point.

The EMA method is employed in the teacher-student network to transfer weights and enhance the model's generalization capability. The core idea of this method is to introduce an EMA to integrate the model's predictions at different time steps. EMA assigns greater weight to recent predictions, allowing the model to focus on the most current information. During the weight transfer process, the student network performs forward propagation to compute the EMA weights, which are then applied to the corresponding layers of the teacher network. In this way, the parameters of the teacher network are updated by exponentially weighted moving averages during training, making the teacher network more robust and guiding the student network to use pseudo-labeling for learning in the context of consistency concerns.

3.4. Boundary difference over union loss (BDoU-Loss)

Most of the current image segmentation algorithms uniquely label the image according to specific rules, which are affected by the geometric constraints of the view. Especially in the fundus image of a lesion, since the captured fundus image is a two-dimensional (2D) planar view while the OC is a three-dimensional (3D) anatomical structure, the OD and OC region presents a circular 2D planar projection in the fundus image, which prevents the accurate representation of its true 3D shape. This leads to an uneven distribution of pixels between the OC and the background region, which reduces the boundary segmentation accuracy of the OC. The comparison between the 3D anatomical structure and the 2D planar of the OD and OC in the fundus image is shown in Figure 3.

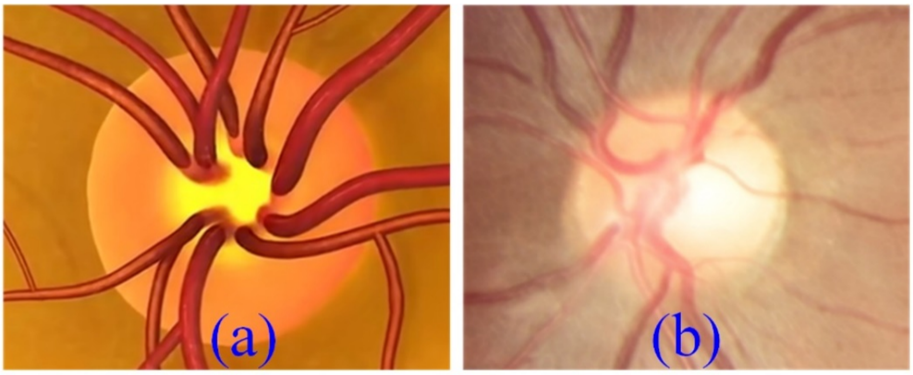


Figure 3. Stereoscopic and planar comparison of OD and OC: (a) 3D anatomical structure, (b) 2D planar structure.

Current loss functions for medical image segmentation primarily focus on the overall segmentation results, and less loss is proposed for the guidance of boundary segmentation. We employ the BDoU-Loss [41], abbreviated as L_{BD}, to address this issue to enhance boundary segmentation. L_{BD} calculates the error region near the boundary by computing the set difference between the ground truth and prediction. The error region is then reduced by minimizing the difference ratio set to the partial intersection or union. The task of joint segmentation of OD and OC is a multilabel problem, where a specific pixel can belong to multiple categories. L_{BD} improves the attention to the region near the boundary and effectively addresses the challenges of insufficient attention to boundary segmentation and uneven categorization. The schematic diagram of the L_{BD} calculation is shown in Figure 4.

Figure 4 shows the structure of the principle of BDoU-loss, and the green line region on the right represents the union area minus the intersection of the prediction and ground truth. Below, a hyperparameter α controls this union minus the intersection area. In this way, the boundary of different images can be adjusted to better guide the boundary segmentation, thus increasing the accuracy of OD and OC segmentation. The principle is illustrated in Eq (8).

$$L_{BD} = \frac{G \cup P - G \cap P}{G \cup P - \alpha^{*} G \cap P}, \tag{8}$$

where α is a hyperparameter that controls the influence of the partial joint region. α is calculated as shown in Eq (9).

$$\alpha = 1 - 2 \times \frac{C}{S}, \alpha \in [0,1), \tag{9}$$

where *C* is the boundary length, and *S* is the target size.

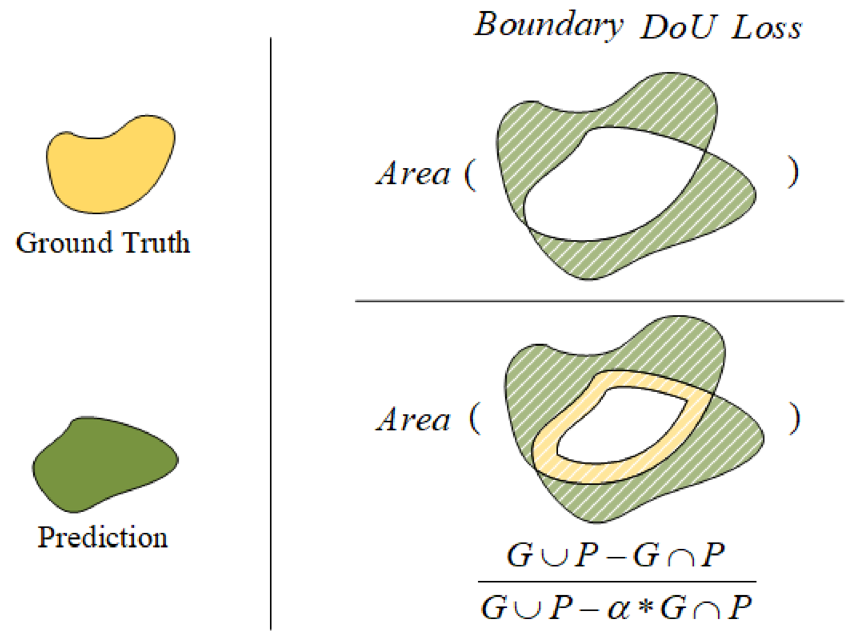


Figure 4. Schematic diagram of Boundary-DoU loss calculation.

4. Experiments

4.1. Dataset

The proposed network is evaluated using the RIGA+, DRISHTI–GS1, and REFUGE datasets. RIGA+ [42] is a composite dataset comprising five subsets: Binrush, Magrabia, BASE1, BASE2, and BASE3. Magrabia and Binrush serve as the source domain, while BASE1, BASE2, BASE3, DRISHTI–GS1, and RIM-ONE-v3 are used as the target domain 1, target domain 2, target domain 3, target domain 4, and target domain 5. Binrush and Magrabia have 195 and 95 labeled fundus retinal images for semi-supervised training, respectively. Labeled and unlabeled retinal fundus images are present in BASE1, BASE2, and BASE3, where unlabeled images are used for semi-supervised training, and labeled images are used to test the segmentation performance of the model. BASE1 consists of 227 unlabeled images and 35 labeled images, BASE2 includes 238 unlabeled images and 30 labeled images, and BASE3 comprises 252 unlabeled images and 27 labeled images. In addition, the DRISHTI-GS1 dataset contains 101 retinal images with a size of 2896 × 1944 pixels, and the RIM-ONE-v3 dataset contains 159 labeled fundus image data. We selected 50 test images from each DRISHTI-GS1 and RIM-ONE-v3 dataset as Targets 4 and 5 to verify the model's generalization performance and crop all the datasets to a size of 800 × 800. Table 1 details the information on the RIGA+, DRISHTI-GS1, and RIM-ONE-v3 datasets, and Figure 5 shows the fundus images of the source and target domains.

Table 1. Introduction to the RIGA+, DRISHTI-GS1, and RIM-ONE-v3 datasets.

Domain	Dataset Names	Numbers (Training + Test)	Image Size
Source	BinRushed	195 (195 + 0)	800 × 800
Source	Magrabia	95 (95 + 0)	800 × 800
Target 1	MESSIDOR-BASE1	173 (138 + 35)	800 × 800
Target 2	MESSIDOR-BASE2	148 (118 + 30)	800 × 800
Target 3	MESSIDOR-BASE3	133 (106 + 27)	800 × 800
Target 4	DRISHTI-GS1	50 (0 + 50)	800 × 800
Target 5	RIM-ONE-v3	50 (0 + 50)	800 × 800

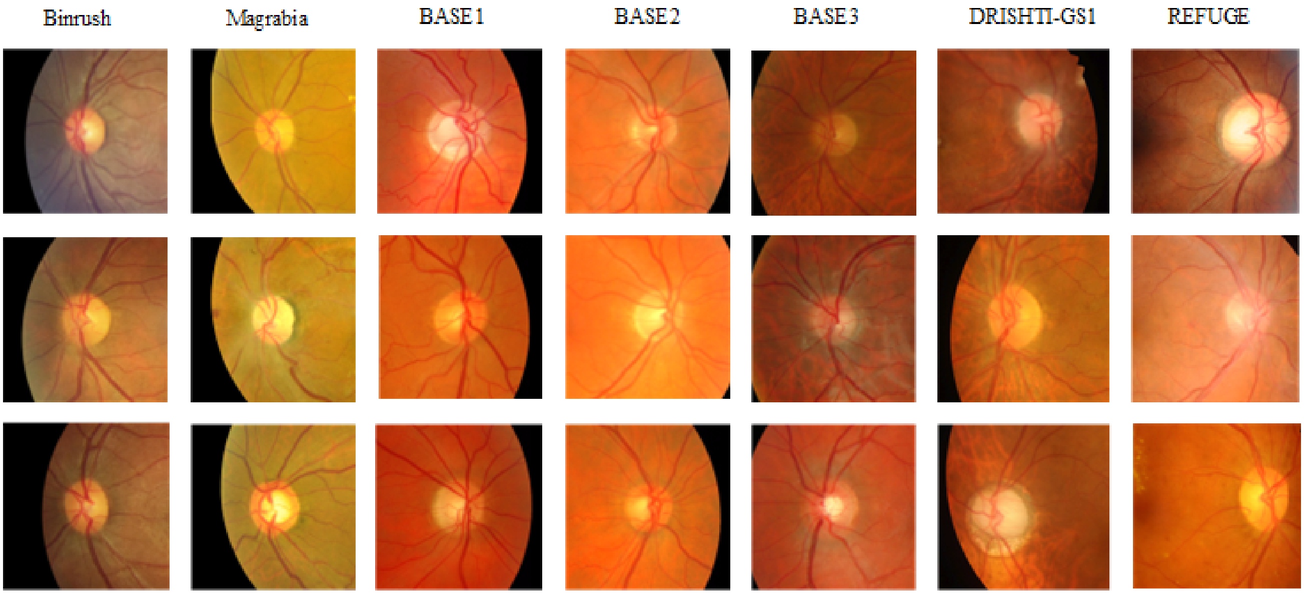


Figure 5. Fundus images of the source and target domains of RIGA+, DRISHTI-GS1, and RIM-ONR-v3 datasets.

4.2. Evaluation metrics

To evaluate the performance of the proposed method and other models, ACC (accuracy), IoU (Intersection over Union), Dice, and Hausdorff distance are used as OD and OC segmentation evaluation metrics.

ACC measures the proportion of correctly classified pixels. It is a representative evaluation indicator in segmentation tasks. The calculation formula is shown in Eq (10).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Here, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

IoU measures the overlap between the predicted result and the actual label. It is an essential indicator for measuring the performance of segmentation tasks, especially for dealing with class imbalance problems. The calculation formula is shown in Eq (11).

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

The dice coefficient is a set similarity measurement function usually used to calculate the similarity between two samples. Higher coefficient values indicate better segmentation results for both OD and OC. The expression of Dice is shown in Eq (12). In most datasets, classic methods yield Dice values for OD segmentation in the range of 70% to 80%, and for OC segmentation in the range of 65% to 70%. Improving the Dice values for segmentation is crucial for future OD and OC segmentation research.

$$Dice = \frac{2\times TP}{2\times TP+FP+FN} \tag{12}$$

The Hausdorff distance is a metric used to determine the degree of resemblance between two point sets. It is beneficial in comparing images or evaluating the similarity between image segmentation results. Given two point sets $A = \{a1, \ldots, ap\}, B = \{b1, \ldots, bq\}$, the Hausdorff distance $H\,(A, B)$ between these two sets is defined as shown in Eqs (13)–(15):

$$H(A, B) = max(h(A, B), h(B, A)), \tag{13}$$

$$h(A, B) = \max_{a\in A}\left\{\min_{b\in B} \parallel a - b \parallel\right\}, \tag{14}$$

$$h(B, A) = \max_{b\in B}\left\{\min_{a\in A} \parallel b - a \parallel\right\}, \tag{15}$$

where $h(A, B)$ and $h(B, A)$ are the directed Hausdorff distance from A to B and B to A, respectively.

In image segmentation, a variant of the Hausdorff distance, Hausdorff distance at the 95th percentile (HD95), is often used to assess the quality of the segmentation results, measuring the distance between two sets. It measures the discrepancy between the segmentation result and the ground truth by calculating the distances between all point pairs in the two sets and finding the 95th percentile of these distances. HD95 is commonly used to assess the accuracy of segmentation results, where a smaller value indicates a higher degree of agreement with the ground truth and, consequently, better segmentation quality. The Hausdorff distance is highly sensitive to outliers, but HD95 mitigates this sensitivity by using the percentile of the distances, providing a more robust metric.

4.3. Implementation details

4.3.1. Experimental setup

We implemented the experiment using Ubuntu 18.04, Pytorch version 1.7, Cuda version 11.4, and Python version 3.7 throughout the OD and OC segmentation comparison, ablation, and generalization experiments. The network training hyperparameters are listed in Table 2.

Table 2. Network training hyperparameters.

Parameter Name	Value
Epoch	200
Batch Size	16
Learning Rate	0.005
Optimizer	Adam
Weight Decay	0

4.3.2. Image preprocessing

The overall preprocessing process is divided into two stages. The first stage operates on a single fundus image, including grayscale conversion, standardization, contrast limited adaptive histogram equalization (CLAHE), and gamma correction. The objective is to enhance the image's contrast, reduce the network training burden, and accelerate convergence. The second stage focuses on data augmentation, including random cropping, image rigid transformation, and elastic transformation, as shown in Figure 6.

In the first stage, to speed up the convergence of the training network and reduce the network training burden, three-channel color images are converted into single-channel grayscale images. The grayscale images are standardized to improve the model's generalization ability and achieve the unity of the metrics. The CLAHE algorithm is used to enhance the local contrast. Finally, a nonlinear Gamma correction is applied to adjust the light intensity of the input retinal image, performing nonlinear operations on the intensity values to establish an exponential relationship between the input and output image intensities.

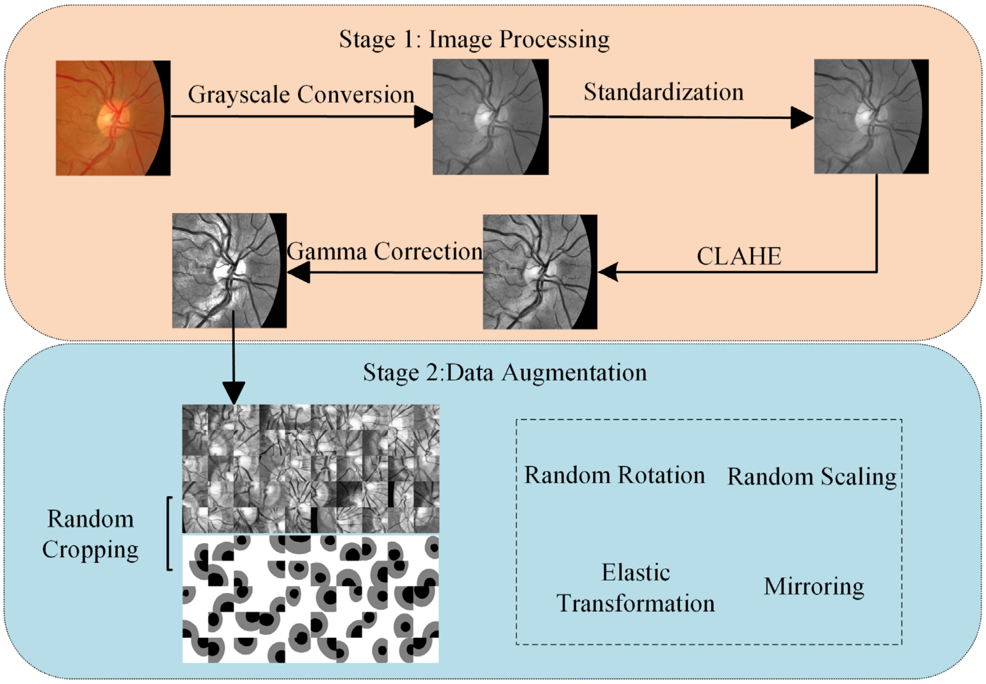


Figure 6. Image preprocessing process.

After the preprocessing stage, the OD is localized and cropped to enlarge the area occupied by the OD in preparation for subsequent data augmentation and OD and OC segmentation. To enhance the limited medical image dataset, data augmentation is an effective strategy to mitigate insufficient training data and reduce overfitting. Various rigid and elastic transformations are applied to the images in the dataset, including scaling, rotation, mirroring, and B-spline elastic deformation. Finally, the dataset images are randomly cropped, and the cropped image blocks of 256 × 256 are then used for network training.

4.4. Ablation analysis

We conducted ablation experiments to validate the effectiveness and contribution of each module within the proposed network for OD and OC segmentation. For the teacher-student network structure, various combinations of CNN and ViT were tested to evaluate the effectiveness of their integration in the model, with the ablation results presented in Table 3.

Table 3. Results of ablation experiments on the RIGA+BASE1 dataset.

Model		$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)	ACC_{OD}	ACC_{OC}	IoU_{OD}	IoU_{OC}
Student	Teacher						
CNN	CNN	76.32 ± 0.98	61.35 ± 1.01	0.8234	0.8156	0.8123	0.6876
ViT	ViT	76.21 ± 1.41	63.81 ± 1.56	0.8198	0.8342	0.8145	0.6842
CNN+CNN	CNN+CNN	82.65 ± 0.38	73.07 ± 0.27	0.8642	0.8578	0.8435	0.7289
ViT+ViT	ViT+ViT	80.76 ± 0.65	70.58 ± 0.49	0.8578	0.8463	0.8382	0.7198
ViT+CNN	ViT+CNN	92.60 ± 0.12	82.62 ± 0.19	0.9465	0.8985	0.8983	0.7891
CNN+ViT	CNN+ViT	**95.12 ± 0.14**	**85.69 ± 0.27**	**0.9723**	**0.9224**	**0.9091**	**0.8149**

Table 3 shows that the combination of CNN and ViT achieves the best OD and OC segmentation performance, with Dice scores of 95.12 ± 0.14 for the OD and 85.69 ± 0.27 for OC, as well as ACC_{OD} of 0.9723, ACC_{OC} of 0.9224, IoU_{OD} of 0.9091, and IoU_{OC} of 0.8149. In contrast, the combinations of ViT+ViT and CNN+CNN result in the poorest segmentation performance. The superior performance of CNN+ViT can be attributed to the strengths of both components: the CNN's spatial locality allows for the effective capture of local features, and its translation invariance helps the model handle variations in the positions of the OD and OC within the image. Meanwhile, the ViT's global awareness facilitates the establishment of connections across the entire image, making it well-suited for capturing the global relationship and overall structural features between OD and OC. The ViT+ViT and CNN+CNN combinations perform poorly due to information redundancy. Using two CNNs can result in redundant extraction of similar features, while two ViTs may lead to redundancy in global perception. The ViT+CNN combination may underperform if the feature extraction in the ViT stage is not robust enough, limiting the CNN stage's ability to classify with rich features, thus constraining the model's performance. Also, the transformer network may lose some spatial information when processing images because it does not explicitly consider spatial localization as CNN does, which may impact segmentation tasks such as OD and OC that require spatial information.

To evaluate the results of ablation experiments more intuitively, we depict the ROC curves of different combinations (as shown in Figure 7). It can be seen from the figure that the AUC of Student: Vit and Teacher: Vit (green line) and Student: CNN and Teacher: CNN (black line) are only 0.9257 and 0.9287, respectively. Then, when the two models are cross-combined, their performance is continuously improved (0.9303, 0.9425). Finally, while S: CNN+ViT combined with T: CNN+Vit is used, the AUC reaches the maximum value of 0.9523, thus proving the rationality of the proposed method.

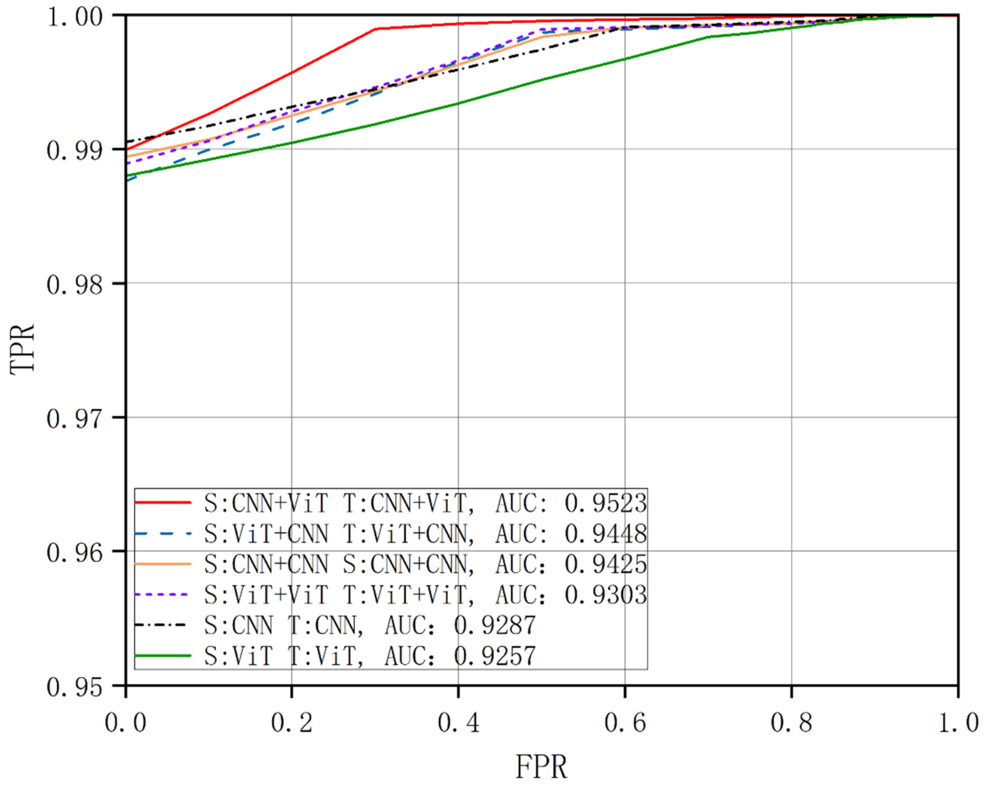


Figure 7. Evaluate ablation experiments on the RIGA+ BASE1 dataset by ROC curve.

In summary, the combination of using a CNN to perform local feature extraction first and then establishing global association through a ViT allows the model to maintain spatial information while being able to globally associate and understand semantic features in different regions of the image, improving the flexibility and versatility of the model and maximizing the advantages of both. The visualization results of the ablation experiments are shown in Figure 8, and the visualization comparison map can more intuitively demonstrate the segmentation differences between different combinations. To better illustrate the segmentation results of varying module combinations, the segmentation of the OD and OC is extracted and shown in green. In contrast, the ground truth is shown in red. By overlaying these two colors, the resulting visualization indicates segmentation quality. In this way, the segmentation effect can be visualized.

From Figure 8, it can be seen that the CNN+CNN combination has obvious over-segmentation and under-segmentation when processing the OC region. For the segmentation of the OD, the over-segmentation is particularly prominent. This is because the combination pays too much attention to local feature extraction, resulting in strong segmentation ability in the local region and many over-segmentation phenomena due to the weak performance in the segmentation judgment of the boundary part. This indicates that relying solely on local feature extraction can lead to over-extraction of features, thereby degrading segmentation performance. The combination of ViT+ViT performs well in overall control, but is deficient in detailed feature extraction. As a result, this combination tends to experience under-segmentation issues in OD and OC segmentation. The significant presence of red areas in the figure indicates that under-segmentation is a prevalent problem. Secondly, the ViT+CNN combination has certain deficiencies in feature extraction. Due to the poor effect of feature extraction in the ViT stage, the impact of feature extraction in the CNN stage is also affected. As can be seen from the figure, although there is an improvement compared to the first two combinations, there are still more under-segmentation cases. Finally, the CNN+ViT combination demonstrates excellent performance. After feature extraction by the CNN, the model obtains rich local feature information, while ViT excels in overall detail management. This combination effectively combines the local and global information, significantly improving the model's over-segmentation and under-segmentation. As can be seen from the figure, the red and green regions in the OD and OC regions are the least in all comparisons.

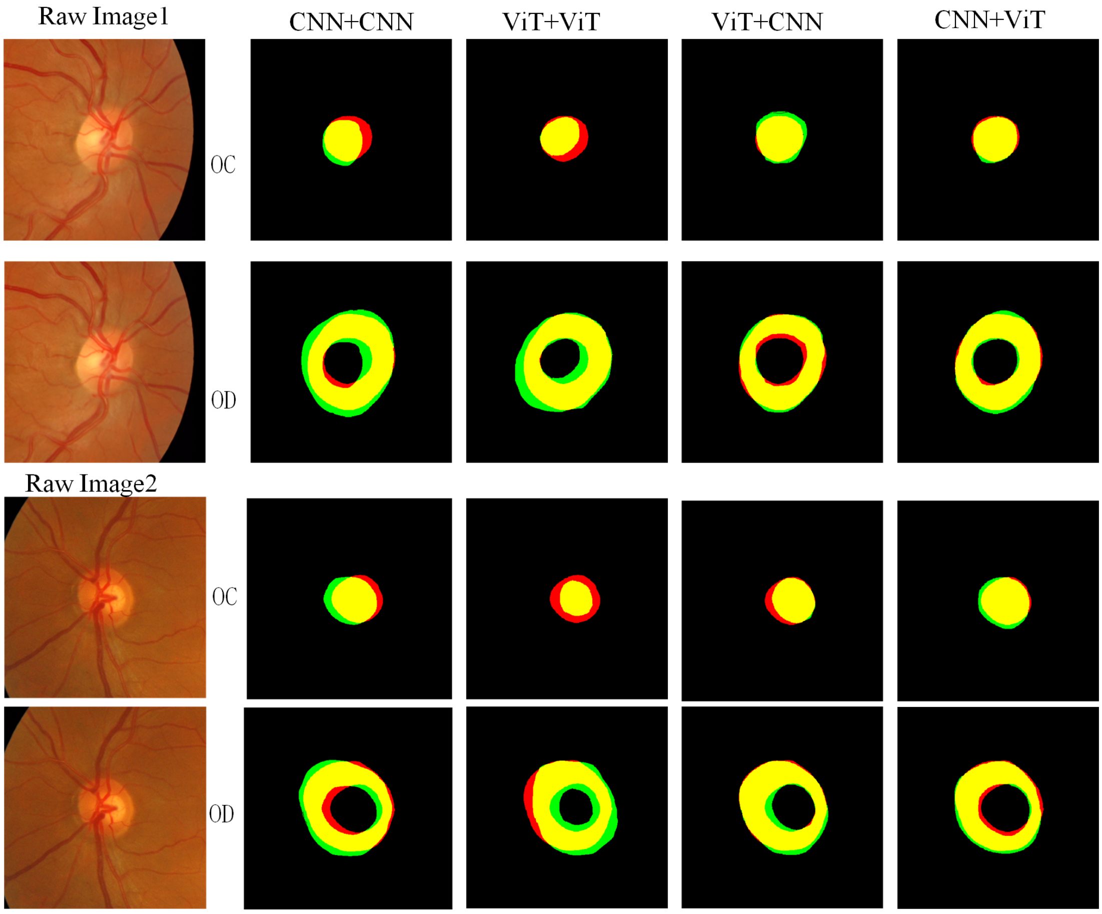


Figure 8. Visualization of ablation experiment on the RIGA+BASE1 dataset (green is the prediction result, red denotes the ground truth, and yellow represents the overlaying region).

4.5. Quantitative comparison with SOTA methods

In this section, the proposed network is compared in detail with SOTA methods. Table 4 details the experimental results of the various techniques on the RIGA+BASE1, BASE2, and BASE3 datasets. The evaluation metric used is the mean Dice score, with the standard deviation indicated and the best results highlighted in bold.

Table 4. Performance of methods on the RIGA+BASE1, BASE2, and BASE3 datasets.

Methods	BASE1		BASE2		BASE3	
	$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)	$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)	$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)
AdaEnt [43]	94.46 ± 0.03	82.71 ± 0.06	92.77 ± 0.02	77.79 ± 0.03	93.72 ± 0.03	81.87 ± 0.04
AdaMI [44]	94.50 ± 0.06	82.80 ± 0.19	92.72 ± 0.02	78.86 ± 0.19	93.65 ± 0.06	82.71 ± 0.11
FSM [45]	94.96 ± 0.63	84.30 ± 1.47	93.10 ± 0.32	81.39 ± 0.91	94.41 ± 0.98	83.21 ± 1.92
pOSAL [46]	94.01 ± 0.23	83.37 ± 0.09	95.09 ± 0.12	84.28 ± 0.17	94.77 ± 0.12	84.06 ± 0.15
BEAL [47]	**95.31** ± 0.35	85.44 ± 0.21	95.57 ± 0.34	83.18 ± 0.25	94.79 ± 0.19	83.98 ± 0.22
ProFSDA [48]	95.29 ± 0.12	85.61 ± 0.24	94.71 ± 0.01	85.33 ± 0.08	95.47 ± 0.01	85.53 ± 0.15
HPFG [49]	93.11 ± 0.16	84.43 ± 0.21	93.12 ± 0.43	83.88 ± 0.31	93.23 ± 0.23	84.01 ± 0.21
MBU-Net [50]	95.00 ± 0.15	85.50 ± 0.25	95.80 ± 0.30	85.20 ± 0.20	95.30 ± 0.16	85.55 ± 0.12
RFAUCNxt [51]	95.05 ± 0.13	85.60 ± 0.23	95.90 ± 0.28	85.35 ± 0.18	95.40 ± 0.14	85.60 ± 0.10
DST-Net	95.12 ± 0.14	**85.69** ± 0.27	**95.97** ± 0.31	**85.37** ± 0.23	**95.49** ± 0.17	**85.62** ± 0.11

As shown in Table 4, DST-Net performs best segmentation on the BASE2 and BASE3 datasets. However, in the BASE1 dataset, the OD Dice score is slightly lower than that of the BEAL network due to the presence of images with lesions. The BEAL network generates more precise boundaries and reduces high-uncertainty predictions in the segmentation of the OD and OC by encouraging similarity between boundary predictions and mask probability entropy maps (uncertainty maps) in the target and source domains. This approach provides an advantage in OD boundary segmentation and enhances accuracy in identifying pathological regions.

Figure 9 displays the segmentation results of several networks on the BASE1, BASE2, and BASE3 datasets. It can be observed that DST-Net performs well in segmenting both OD and OC. BEAL performs better than pOSAL for OD segmentation, showing its advantage in boundary segmentation, especially in cases where the OD has a more regular boundary. However, for the case of irregular boundaries such as the OC, the BEAL model has more obvious under-segmentation. In contrast, DST-Net achieves better segmentation of irregular boundaries by leveraging transfer learning through the pre-training of the teacher network, allowing it to better adapt to the OD and OC segmentation tasks. With the cooperation of CNN and ViT, DST-Net has significantly improved its segmentation effect in dealing with irregular boundaries.

To more intuitively demonstrate the proposed network's focus on different regions of fundus images, heatmaps were generated to illustrate the network's attention across various areas, as shown in Figure 10.

It can be observed from Figure 10 that the blue region indicates that the network pays less attention to the region, while the red region indicates that the network pays more attention to the region. As each layer's weights accumulate, the attention to the OD and OC regions gradually increases, indicating that the segmentation network is more focused on these areas, resulting in improved performance. In the heatmap of the BASE2 dataset, the OD region is predominantly covered by red, showing relatively good segmentation. However, some black shadowed areas are at the bottom, probably due to the interference of blood vessels in the early stages of training, attracting the model's attention. In the heatmap of the BASE3 dataset, the focus is mainly on the OD and OC regions, with other areas showing a deep blue, indicating that the segmentation of the OD and OC has been successful.

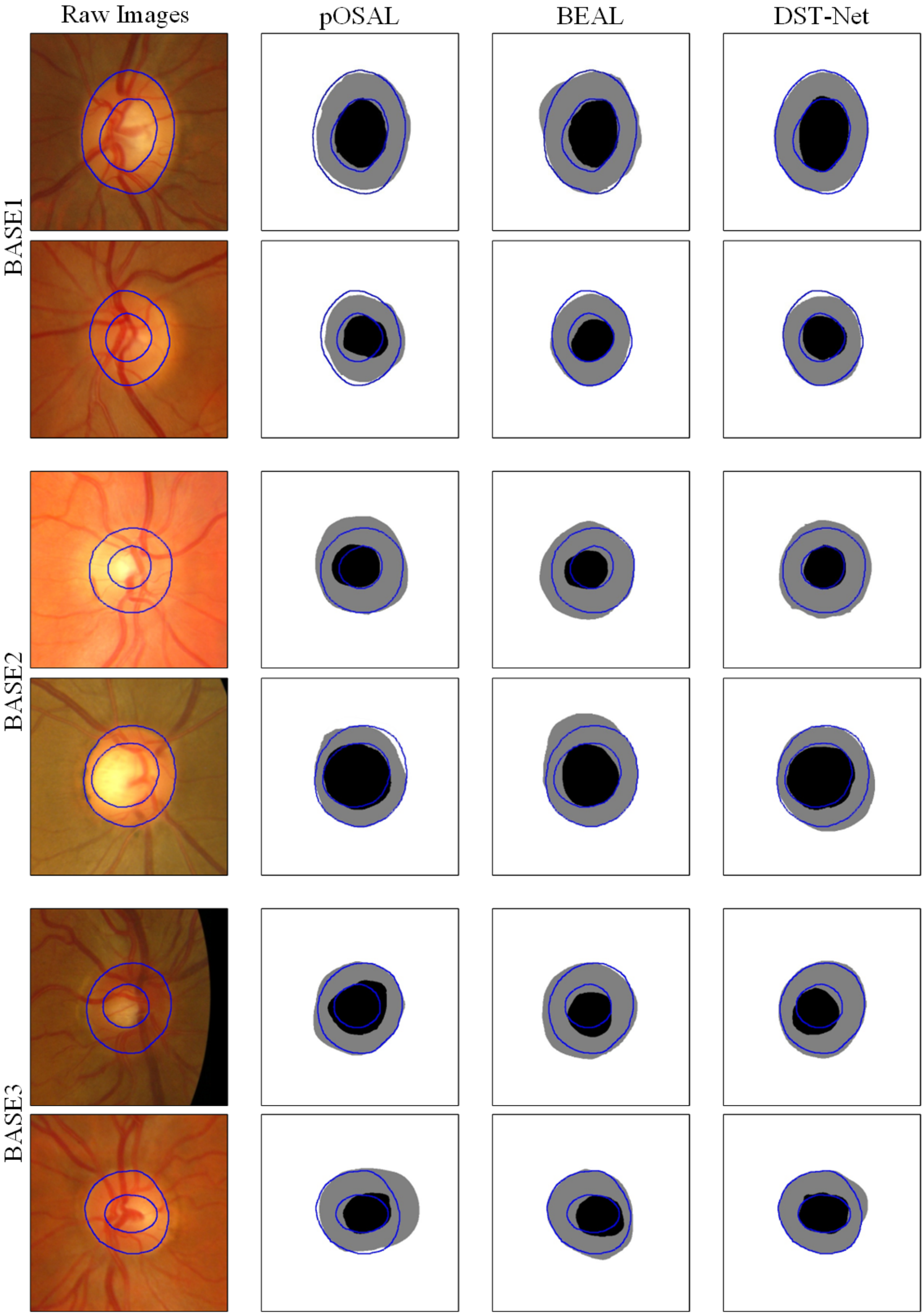


Figure 9. Comparison of visualization of different network segmentation results on the RIGA+BASE1, BASE2, and BASE3 datasets (blue line represents ground truth).

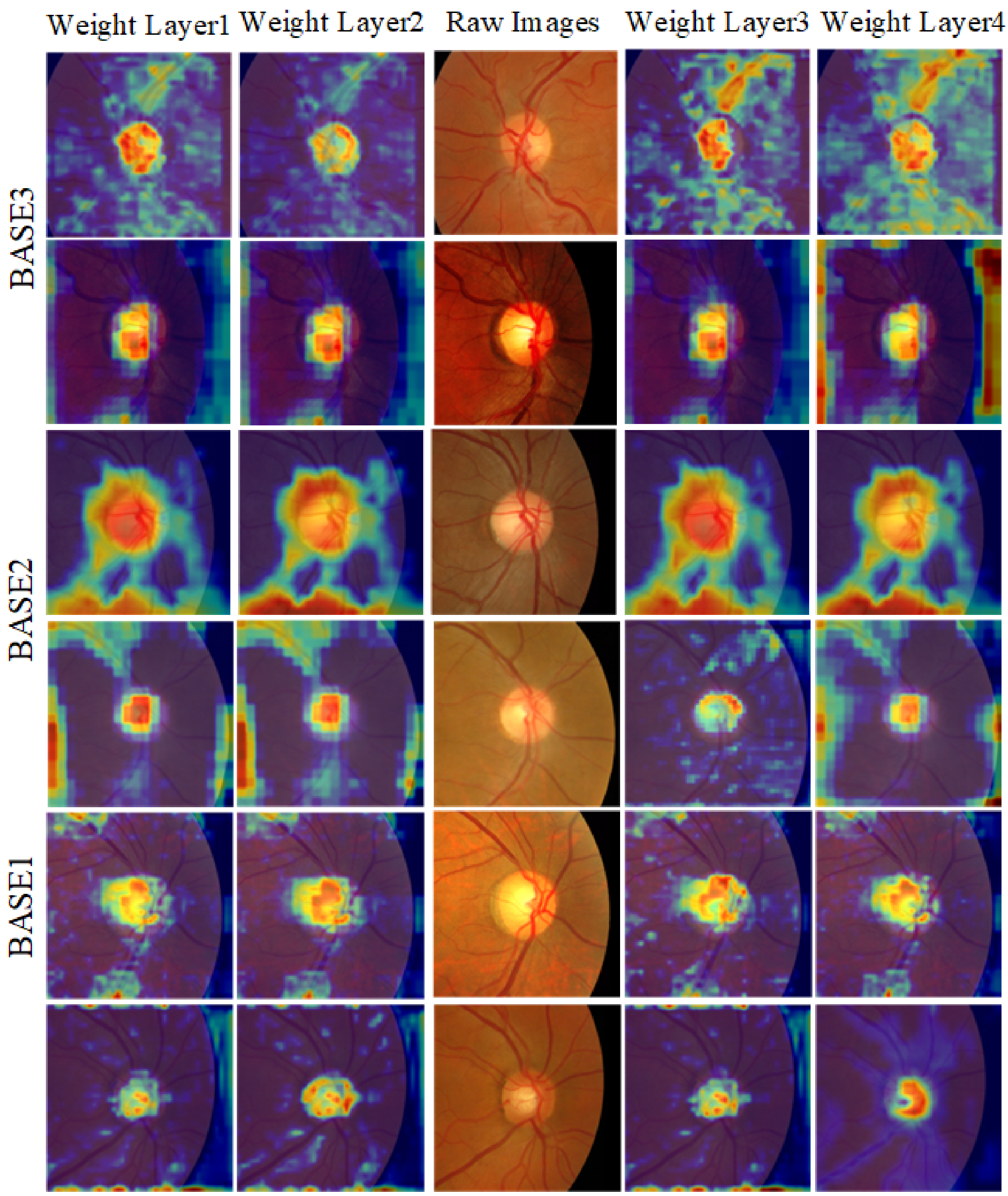


Figure 10. Heat map displaying the attention of different weight layers of the network on the RIGA+BASE1, BASE2, and BASE3 datasets.

Despite some black shadow regions affecting the segmentation performance, the model shows great competitiveness overall. This finding provides direction for subsequent optimization and improvement for challenging regions. Meanwhile, the combined analysis with the heatmap offers a clearer understanding of the regions of interest of the network for the fundus image, which will help to further optimize the network structure and parameters for more accurate OD and OC segmentation.

4.6. Loss function analysis on the RIGA+Dataset

To verify the effectiveness of the boundary difference loss function selected in this paper. This section will use four loss functions, IoU-loss, Dice-loss, MSE-loss, and BDoU-loss, to conduct comparative experiments on the BASE1 dataset of RIGA+. The Dice similarity coefficient is used to measure the effectiveness of the boundary differential loss function, and the HD95 is used to evaluate boundary attention. The experimental results are shown in Table 5.

Table 5. Comparative experiments of different loss functions.

Loss Function	$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)	HD95
IoU-loss	94.76 ± 0.78	83.59 ± 0.66	14
Dice-loss	**95.44 ± 0.55**	83.94 ± 0.63	17
MSE-loss	87.14 ± 0.25	76.90 ± 0.23	19
BDoU-loss	95.12 ± 0.14	**85.69 ± 0.27**	**11**

As can be seen in Table 5 in the comparison of the metrics, the MSE-loss performs the worst across various metrics. Its principle involves penalizing model errors by calculating the squared error between prediction and the ground truth, making it highly sensitive to outliers. Furthermore, the MSE loss is less effective in dealing with classification problems, contributing to poor segmentation. Compared with IoU-loss, BDoU-loss achieves higher Dice values for both OD and Dice. The main reason is that for triple classification problems like OD and OC segmentation, the BDoU-loss pays more attention to capturing the detailed information of the segmentation boundary, especially for the OD boundary and the OD and OC boundary, thereby improving the model's segmentation accuracy. In terms of OD segmentation, BDoU-loss performs slightly worse than Dice-loss. However, it achieves a higher average Dice value overall. Dice-Loss usually has better sensitivity to boundary prediction, and thus has higher accuracy in segmentation tasks with regular boundaries, which is better reflected in OD segmentation. Nevertheless, due to the low contrast of the OD and OC boundary and the relatively irregular boundary of the OC, boundary incompleteness might occur, leading to slightly inferior segmentation performance.

In contrast, BDoU-loss is made more tolerant to some slight boundary incompleteness by introducing a limiting factor, which mitigates the negative impact of these incompleteness and irregular shapes on the segmentation of the OD and OC, resulting in better segmentation of the OC region. Overall, BDoU-loss demonstrates the best comprehensive performance among the compared loss functions, yielding the most optimal segmentation results. The superiority of BDoU-loss is further illustrated by the variation in loss values during the training process with different loss functions, as shown in Figure 11.

As shown in Figure 11, the loss curve of BDoU-loss during the training process is smoother, with more minor training fluctuations. This indicates that BDoU-loss provides superior stability against outliers or noise during training. In datasets with noise or anomalies, BDoU-loss helps the model better adapt to these conditions. A comparison of training speeds reveals that BDoU-loss tends to converge more quickly to local minima during training due to the introduction of a constraint factor. This means it has a faster convergence speed, reducing training time and improving efficiency.

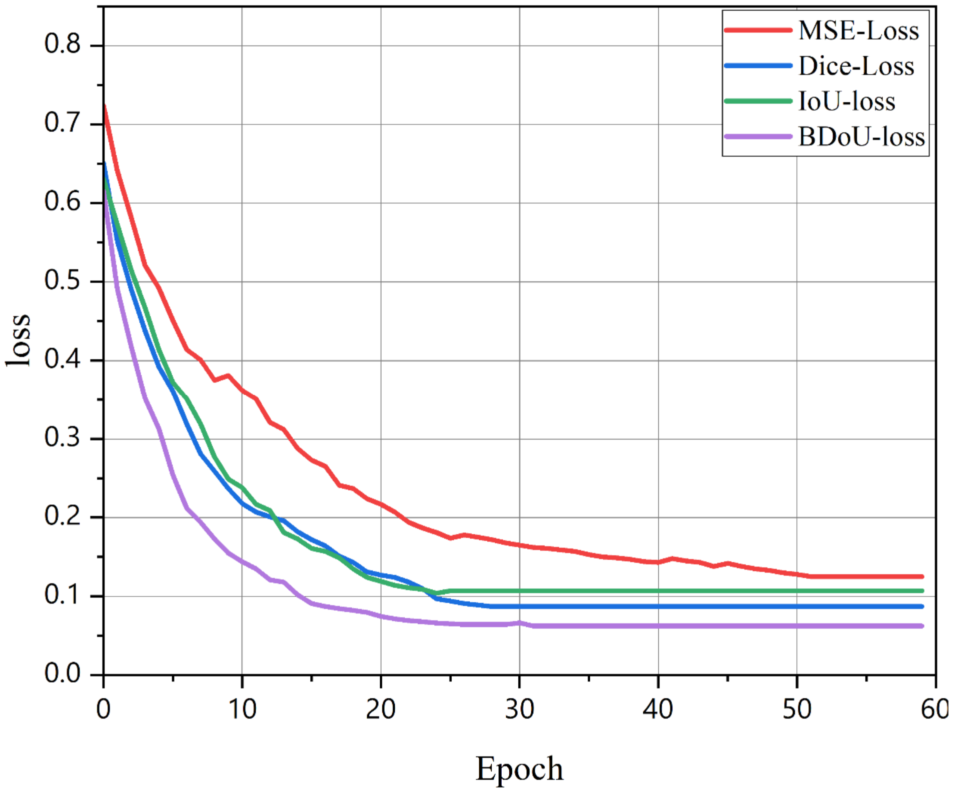


Figure 11. Comparison of training loss curves with different loss functions.

To show the segmentation differences between different loss functions more intuitively, the segmentation results of the ground truth and different loss functions are drawn on the original graph with blue lines for visualization, as shown in Figure 12.

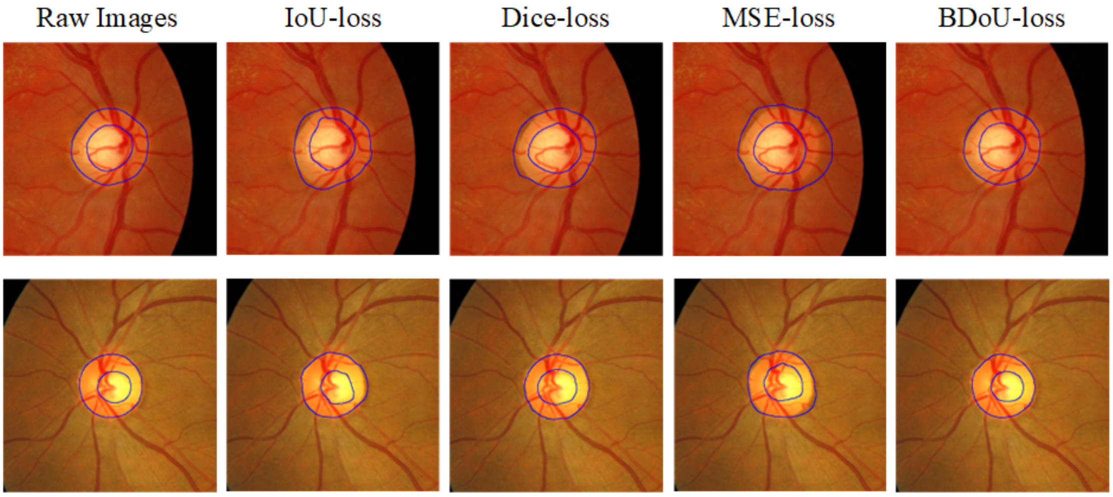


Figure 12. Comparison of different loss function training visualizations.

Since the boundaries of the vast majority of the OD and OC segmentation results are continuous, the segmentation results can be observed more intuitively by extracting the boundaries of the segmentation results of different loss functions and overlaying them into the original image, as shown in Figure 12. From the visualization comparison, it can be seen that the Dice-loss and BDoU-loss yield almost identical segmentation results for the OD. However, BDoU-loss shows better performance for OC segmentation, which is closer to the ground truth. IoU-loss performs poorly for the OC segmentation and introduces many irregular areas in the OD segmentation. This is because IoU-loss is more sensitive to minor deviations in boundary predictions, and such sensitivity might penalize the model when slight irregularities appear in predicted boundaries, resulting in inferior performance compared to Dice-loss and BDoU-loss, which shows the importance of boundary segmentation for OD and OC segmentation.

MSE-loss is less effective for both OD and OC segmentation. The reason is that it does not consider the relative positional relationship between pixels and thus may not be able to handle the boundary ambiguity in the segmentation task effectively. If there is an imbalance in pixel distribution between the OD and OC, such as when one category has significantly more pixels than the other, the MSE loss function may guide the model to favor the category with a higher number, resulting in the model being more inclined to learn the category with a higher number while ignoring the one with a lower number.

4.7. Parameters and FLOPs.

We compare our DST-Net with SOTA. Table 6 lists the giga floating-point operations per second (GFLOPs) and the Dice scores for OD and OC for the different methods on the RIGA+BASE2 dataset. In general, the number of parameters of a network is proportional to its computational complexity, and smaller parameters tend to degrade the network's performance. Table 6 shows that the proposed DST-Net performs best regarding OD and OC segmentation accuracy compared to the listed methods. The main reason is that the teacher model is trained by semi-supervised learning, and the generated pseudo-labels are used to train the student model, which improves the model's performance and reduces the need for expensive expert labeling.

Table 6. Training results for each model trained on the RIGA+BASE2 dataset.

Method	Parameters	GFLOPs	BASE2	
			Dice$_{\text{Disc}}$ (%)	Dice$_{\text{Cup}}$ (%)
AdaEnt [43]	40,774,656	**3.621**	92.77 ± 0.02	77.79 ± 0.03
FSM [45]	**34,273,776**	3.736	93.10 ± 0.32	81.39 ± 0.91
pOSAL [46]	53,118,960	4.608	95.09 ± 0.12	84.28 ± 0.17
BEAL [49]	63,247,099	6.410	95.57 ± 0.34	83.18 ± 0.25
ProFSDA [46]	42,513,479	4.531	94.71 ± 0.01	85.33 ± 0.08
HPFG [47]	140,641,904	12.528	93.12 ± 0.43	83.88 ± 0.31
MBU-Net [50]	58,234,128	6.21	94.12 ± 0.15	84.78 ± 0.21
RFAUCNxt [51]	72,651,984	7.89	93.95 ± 0.22	84.23 ± 0.19
Proposed DST-Net	81,672,536	8.26	**95.97** ± 0.31	**85.37** ± 0.23

In addition, the proposed method is the second to last regarding the number of parameters and GLOPs. This is mainly because Transformer methods typically have more parameters than multiscale CNN methods regarding parameter comparison. However, the proposed DST-Net comprises both CNN and Transformer modules, only using high-level features extracted by CNNs for global feature modeling in the Transformer, which can significantly reduce computational load to a certain extent.

4.8. Generalization performance evaluation

We conducted additional segmentation experiments on two external datasets, DRISHTI-GS1 and RIM-ONE-v3, to assess our proposed method's robustness and generalization ability. The qualitative and quantitative results are shown in Figure 13 and Table 7, respectively.

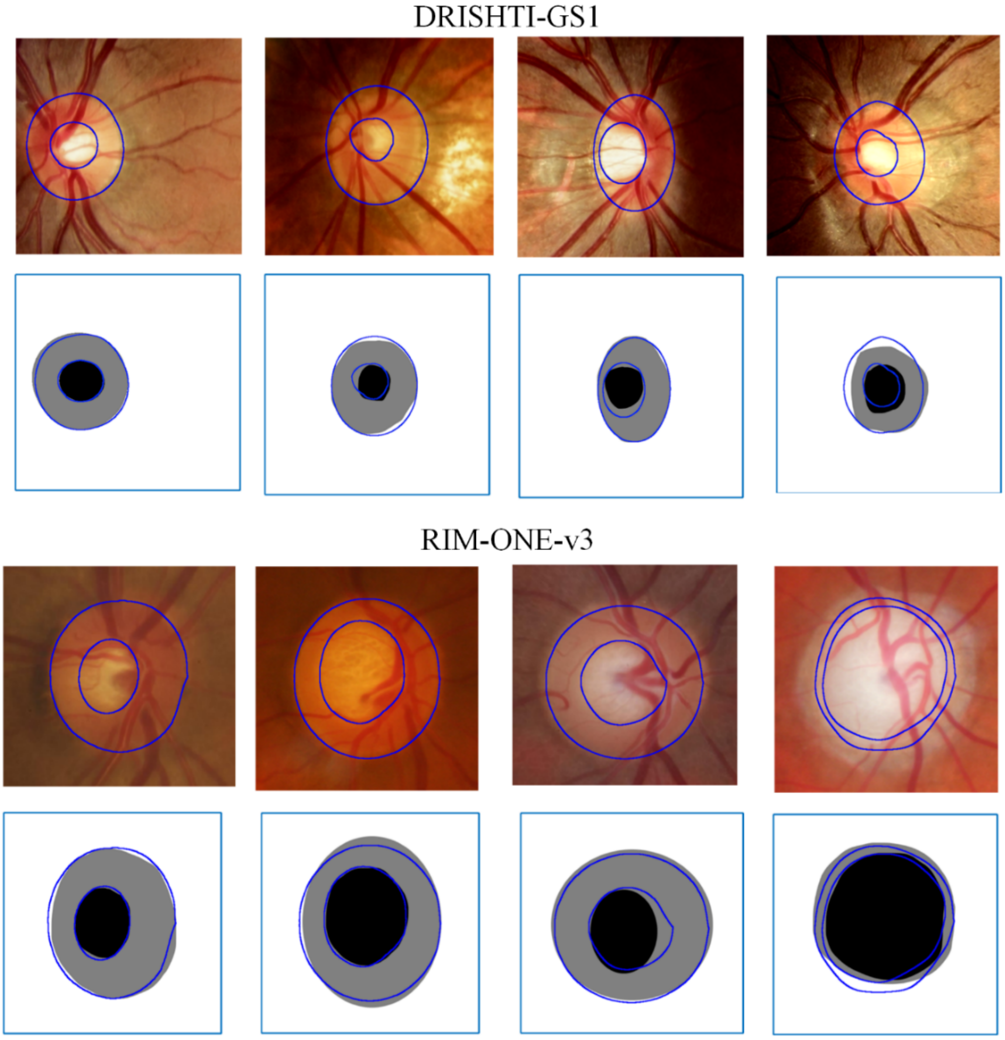


Figure 13. Visualization of the segmentation results of the proposed method on datasets DRISHTI-GS1 and RIM-ONE-v3.

Figure 13 shows the segmentation results on the DRISHTI-GS1 and RIM-ONE-v3 datasets. Our method can still accurately delineate the target region even in challenging situations, such as significant variations in image illumination, contrast, and pathology conditions in DRISHTI-GS1 and differences in acquisition conditions across clinical centers in RIM-ONE-v3. Table 7 summarizes the quantitative performance of our method on these external datasets. The results show that our model maintains high segmentation accuracy, highlighting its robustness in different imaging distributions and strong generalization ability.

Table 7. Generalization of experimental results on DRISHTI-GS and RIM-ONE-v3 datasets.

Dataset	$\text{Dice}_{\text{Disc}}$ (%)	Dice_{Cup} (%)	ACC_{OD}	ACC_{OC}	IoU_{OD}	IoU_{OC}
DRISHTI-GS1	94.78 ± 0.21	85.34 ± 0.35	0.9682	0.9197	0.9045	0.8061
RIM-ONE-v3	95.35 ± 0.42	84.78 ± 0.39	0.9756	0.9221	0.8912	0.8167

5. Discussion

While the proposed CNN-ViT hybrid model demonstrates competitive performance in OD and OC segmentation, its computational complexity and parameter size may constrain deployment in resource-limited environments, such as mobile ophthalmology screening devices. Although the ViT component excels in global context modeling, it incurs an $O(N^2)$ computational cost for self-attention operations on high-resolution fundus images, resulting in substantial memory consumption and increased inference latency.

To mitigate these limitations, recent advancements in lightweight vision Transformers offer promising alternatives. Swin-Transformer V2 leverages hierarchical feature learning through shifted windows and a cosine attention mechanism, reducing computational complexity from quadratic to linear scaling within local windows. Implementing such a shifted window attention mechanism can preserve global context modeling while reducing GPU memory usage by over 40%. Meanwhile, Lite Vision Transformers (LViTs) employ structural re-parameterization and dynamic sparse attention to eliminate redundant token interactions. Their token-slimming module adaptively prunes less informative patches in fundus images, leading to a potential speedup of 1.5 to 2 times without compromising segmentation accuracy. TinyViT [52] further optimizes efficiency by introducing a multi-stage architecture with weight-sharing mechanisms and low-rank decomposition, significantly reducing parameter size while maintaining strong feature representation. This enables TinyViT to achieve high segmentation accuracy on resource-constrained devices.

Additionally, MobileViTv3 [53] enhances efficiency through a hybrid design that fuses convolutional and self-attention mechanisms, improving local-global feature interactions. By leveraging lightweight attention layers and feature fusion techniques, MobileViTv3 achieves competitive performance with minimal computational overhead, making it well-suited for real-time applications. These advancements collectively contribute to more efficient and scalable OD and OC segmentation solutions in diverse clinical settings.

6. Conclusions

In this paper, we proposed a dual self-integrated transformer network (DST-Net) for semi-supervised segmentation of OD and OC. The proposed DST-Net fully leverages the characteristics of both ViTs and CNNs to balance local and global information. A semi-supervised approach effectively utilizes large amounts of unlabeled data for network training. Aiming at the problem that the real 3D morphology of the OD and OC cannot be displayed in the 2D fundus images, which leads to uneven pixel distribution and unclear boundaries, we introduce BDoU-loss. This loss function focuses on boundary segmentation, improving accuracy at the irregular boundaries between the OD and OC. Ablation experiments verify the effectiveness of the CNN and ViT combination.

In comparison experiments, DST-Net demonstrated superior performance on the BASE1, BASE2, and BASE3 datasets, with OD Dice scores of 0.9512, 0.9597, and 0.9549, and OC Dice scores of 0.8569, 0.8537, and 0.8562, respectively. The analysis of loss functions confirms the effectiveness of the BDoU-loss selected in this paper. Future work will benchmark these lightweight alternatives and systematically evaluate their trade-offs between computational efficiency and segmentation accuracy, enhancing their applicability in real-world scenarios.

Further, the network structure of DST-Net could be optimized to reduce the number of parameters while maintaining the segmentation accuracy to enhance model performance and optimize computational efficiency.

Data availability

The RIGA+ datasets used in this paper are publicly available as follows: https://deepblue.lib.umich.edu/data/concern/data_sets/3b591905z, and https://www.adcis.net/en/third-party/messidor2/.

The DRISHTI-GS1 datasets used in this paper are publicly available as follows: Drishti-GS - RETINA DATASET FOR ONH SEGMENTATION.

The RIM-ONE-v3 datasets used in this paper are publicly available as follows: Demos and Resources |Image Processing Group.

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 supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant PL2024F010.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, C. Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis, *Ophthalmology*, **121** (2014), 2081–2090. https://doi.org/10.1016/j.ophtha.2014.05.013
2. A. C. Thompson, A. A. Jammal, F. A. Medeiros, A review of deep learning for screening, diagnosis, and detection of glaucoma progression, *Transl. Vision Sci. Technol.*, **9** (2020), 42. https://doi.org/10.1167/tvst.9.2.42

3. R. C. Zhao, X. L. Chen, X. Y. Liu, Z. L. Chen, F. Guo, S. Li, Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning, *IEEE J. Biomed. Health Inf.*, **24** (2019), 1104–1113. https://doi.org/10.1109/JBHI.2019.2934477
4. C. Jia, F. Shi, M. Zhao, Y. Zhang, X. Cheng, M. Z. Wang, et al., Semantic segmentation with light field imaging and convolutional neural networks, *IEEE Trans. Instrum. Meas.*, **70** (2021), 5017214. https://doi.org/10.1109/TIM.2021.3115204
5. T. Hassan, B. Hassan, M. U. Akram, S. Hashimi, A. H. Taguri, N. Werghi, Incremental cross-domain adaptation for robust retinopathy screening via Bayesian deep learning, *IEEE Trans. Instrum. Meas.*, **70** (2021), 2516414. https://doi.org/10.1109/TIM.2021.3122172
6. Y. F. Guo, Y. J. Peng, B. Zhang, CAFR-CNN: Coarse-to-fine adaptive faster R-CNN for cross-domain joint optic disc and cup segmentation, *Appl. Intell.*, **51** (2021), 5701–5725. https://doi.org/10.1007/s10489-020-02145-w
7. L. Luo, D. Y. Xue, F. Pan, X. L. Feng, Joint optic disc and optic cup segmentation based on boundary prior and adversarial learning, *Int. J. Comput. Assisted Radiol. Surg.*, **16** (2021), 905–914. https://doi.org/10.1007/s11548-021-02373-6
8. P. S. Yin, Y. W. Xu, J. H. Zhu, J. Liu, C. A. Yi, H. C. Huang, et al., Deep level set learning for optic disc and cup segmentation, *Neurocomputing*, **464** (2021), 330–341. https://doi.org/10.1016/j.neucom.2021.08.102
9. J. N. Chen, Y. Y. Lu, Q. H. Yu, X. D. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306.
10. A .Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates, Inc., **30** (2017), 1–11.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
12. J. Wu, W. Ji, H. Z. Fu, M. Xu, Y. M. Jin, Y. W. Xu, Medsegdiff-v2: Diffusion-based medical image segmentation with transformer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, **38** (2024), 6030–6038. https://doi.org/10.1609/aaai.v38i6.28418
13. Y. Chen, D. Su, J. Luo, Laplacian-guided hierarchical transformer: A network for medical image segmentation, *Comput. Methods Programs Biomed.*, **260** (2025), 108526. https://doi.org/10.1016/j.cmpb.2024.108526
14. E. Goceri, Medical image data augmentation: Techniques, comparisons and interpretations, *Artif. Intell. Rev.*, **56** (2023), 12561–12605. https://doi.org/10.1007/s10462-023-10453-z
15. Y. Wang, J. Cheng, Y. Chen, S. Shao, L. Y. Zhu, Z. Z. Wu, et al., Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation, *IEEE Trans. Med. Imaging*, **42** (2023), 3738–3751. https://doi.org/10.1109/TMI.2023.3306105
16. P. L. Shi, J. N. Qiu, S. M. D. Abaxi, H. Wei, F. P. W. Lo, W. Yuan, Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation, *Diagnostics*, **13** (2023), 1947. https://doi.org/10.3390/diagnostics13111947
17. J. Zilly, J. M. Buhmann, D. Mahapatra, Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation, *Comput. Med. Imaging Graphics*, **55** (2017), 28–41. https://doi.org/10.1016/j.compmedimag.2016.07.012

18. L. Wang, J. Gu, Y. Z. Chen, Y. B. Liang, W. J. Zhang, J. T. Pu, et al., Automated segmentation of the optic disc from fundus images using an asymmetric deep learning network, *Pattern Recognit.* **112** (2021), 107810. https://doi.org/10.1016/j.patcog.2020.107810
19. A. Tulsani, P. Kumar, S. Pathan, Automated segmentation of optic disc and optic cup for glaucoma assessment using improved UNET++ architecture, *Biocybern. Biomed. Eng.*, **41** (2021), 819–832. https://doi.org/10.1016/j.bbe.2021.05.011
20. S. Pachade, P. Porwal, M. Kokare, L. Giancardo, F. Meriaudeau, NENet: Nested EfficientNet and adversarial learning for joint optic disc and cup segmentation, *Med. Image Anal.*, **74** (2021) 102253. https://doi.org/10.1016/j.media.2021.102253
21. X. X. Guo, J. H. Li, Q. F. Lin, Z. H. Tu, X. Y. Hu, S. T. Che, Joint optic disc and cup segmentation using feature fusion and attention, *Comput. Biol. Med.*, **150** (2022), 106094 https://doi.org/10.1016/j.compbiomed.2022.106094
22. H. Z. Fu, J. Cheng, Y. W. Xu, D. W. K. Wong, J. Liu, X. C. Cao, Joint optic disc and cup segmentation based on multilabel deep network and polar transformation, *IEEE Trans. Med. Imaging*, **37** (2018), 1597–1605. https://doi.org/10.1109/TMI.2018.2791488
23. Z. Q. Zhu, Z. M. Zhang, G. Q. Qi, Y. Y. Li, Y. Z. Li, L. Mu, A dual-branch network for ultrasound image segmentation, *Biomed. Signal Process. Control*, **103** (2025), 107368 https://doi.org/10.1016/j.bspc.2024.107368
24. Z. Q. Zhu, X. Y. He, G. Q. Qi, Y. Y. Li, B. S. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, *Inf. Fusion*, **91** (2023) 376–387. https://doi.org/10.1016/j.inffus.2022.10.022
25. Y. H. Fu, J. Chen, J. Li, D. Y. Pan, X. Z. Yue, Y. M. Zhu, Optic disc segmentation by U-net and probability bubble in abnormal fundus images, *Pattern Recognit.*, **117** (2021), 107971. https://doi.org/10.1016/j.patcog.2021.107971
26. H. Cao, Y. Y. Wang, J. Chen, D. S. Jiang, X. P. Zhang, Q. Tian, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in *European Conference on Computer Vision–ECCV 2022 Workshops*, Springer, (2022), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
27. S. H. Li, X. C. Sui, X. D. Luo, X. X. Xu, Y. Liu, R. Goh, Medical image segmentation using squeeze-and-expansion transformers, in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, (2021), 807–815. https://doi.org/10.24963/ijcai.2021/112
28. Z. Q. Zhu, Z. Y. Wang, G. Q. Qi, N. Mazur, P. Yang, Y. Liu, Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction, *Pattern Recognit.*, **153** (2024), 110553. https://doi.org/10.1016/j.patcog.2024.110553
29. Z. Q. Zhu, K. Yu, G. Q. Qi, B. S. Cong, Y. Y. Li, Z. X. Li, et al., Lightweight medical image segmentation network with multi-scale feature-guided fusion, *Comput. Biol. Med.*, **182** (2024), 109204. https://doi.org/10.1016/j.compbiomed.2024.109204
30. Z. Q. Zhu, M. W. Sun, G. Q. Qi, Y. Y. Li, X. B. Gao, Y. Liu, Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation, *Comput. Biol. Med.*, **172** (2024), 108284. https://doi.org/10.1016/j.compbiomed.2024.108284
31. Y. H. Fu, J. F. Liu, J. Shi, TSCA-Net: Transformer based spatial-channel attention segmentation network for medical images, *Comput. Biol. Med,*. **170** (2024), 107938. https://doi.org/10.1016/j.compbiomed.2024.107938

32. Y. G. Yi, Y. Jiang, B. Zhou, N. Y. Zhang, J. Y. 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
33. R. Hussain, H. Basak, Ut-net: Combining u-net and transformer for joint optic disc and cup segmentation and glaucoma detection, preprint, arXiv:2303.04939.
34. J. D. Wu, H. H. Fang, F. X. Shang, D. L. Yang, Z. W. Wang, J. Gao, et al., SeATrans: Learning segmentation-assisted diagnosis model via transformer, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, Springer, **13432** (2022), 677–687. https://doi.org/10.1007/978-3-031-16434-7_65
35. Z. Liu, H. Hu, Y. T. Lin, Z. L. Yao, Z. D. Xie, Y. X. Wei, et al., Swin transformer v2: Scaling up capacity and resolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2022), 12009–12019. https://doi.org/10.48550/arXiv.2111.09883
36. C. L. Yang, Y. L. Wang, J. M. Zhang, H. Zhang, Z. J. Wei, Z. Lin, et al., Lite vision transformer with enhanced self-attention, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2022), 11998–12008.
37. B. Han, Q. M. Yao, X. R. Yu, G. Niu, M. Xu,W. H. Hu, et al., Co-teaching: Robust training of deep neural networks with extremely noisy labels, in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Curran Associates, Inc., **31** (2018), 1–11.
38. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates, Inc., **30** (2017), 1–10.
39. S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, preprint, arXiv:1610.02242.
40. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. F. Li, Imagenet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
41. F. Sun, Z. M. Luo, S. Z. Li, Boundary difference over union loss for medical image segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023*, Springer, **14223** (2023), 292–301. https://doi.org/10.1007/978-3-031-43901-8_28
42. A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, et al., Retinal fundus images for glaucoma analysis: The RIGA dataset, in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, SPIE, (2018), 55–62. https://doi.org/10.1117/12.2293584
43. M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, I. B. Ayed, Source-relaxed domain adaptation for image segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, Springer, **12261** (2020), 490–499. https://doi.org/10.1007/978-3-030-59710-8_48
44. M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, I. B. Ayed, Source-free domain adaptation for image segmentation, *Med. Image Anal.*, **82** (2022), 102617. https://doi.org/10.1016/j.media.2022.102617
45. C. Yang, X. Guo, Z. Chen, Y. Yuan, Source free domain adaptation for medical image segmentation with fourier style mining, *Med. Image Anal.*, **79** (2022), 102457. https://doi.org/10.1016/j.media.2022.102457
46. S. J. Wang, L. Q. Yu, X. Yang, C. W. Fu, P. A. Heng, Patch-based output space adversarial learning for joint optic disc and cup segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2485–2495. https://doi.org/10.1109/TMI.2019.2899910

47. S. Wang, L. Yu, K. Li, X. Yang, C. W. Fu, P. A. Heng, Boundary and entropy-driven adversarial learning for fundus image segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, Springer, **11764** (2019), 102–110. https://doi.org/10.1007/978-3-030-32239-7_12
48. S. Hu, Z. Liao, Y. Xia, ProSFDA: Prompt learning based source-free domain adaptation for medical image segmentation, preprint, arXiv:2211.11514.
49. F. Li, A. Jiang, M. Li, C. Xiao, W. Ji, HPFG: Semi-supervised medical image segmentation framework based on hybrid pseudo-label and feature-guiding, *Med. Biol. Eng. Comput.*, **62** (2024), 405–421. https://doi.org/10.1007/s11517-023-02946-4
50. Y. L. He, J. Kong, D. Liu, J. Li, C. Zheng, Self-ensembling with mask-boundary domain adaptation for optic disc and cup segmentation, *Eng. Appl. Artif. Intell.*, **129** (2024), 107635. https://doi.org/10.1016/j.engappai.2023.107635
51. S. Mallick, J. Paul, J. Sil, Response fusion attention U-ConvNext for accurate segmentation of optic disc and optic cup, *Neurocomputing*, **559** (2023), 126798. https://doi.org/10.1016/j.neucom.2023.126798
52. K. Wu, J. Zhang, H. Peng, M. C. Liu, B. Xiao, J. L. Fu, et al., Tinyvit: Fast pretraining distillation for small vision transformers, in *European Conference on Computer Vision* – *ECCV 2022*, Springer, **13681** (2022), 68–85. https://doi.org/10.1007/978-3-031-19803-8_5
53. S. N. Wadekar, A. Chaurasia, Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features, preprint, arXiv:2209.15159.