Expert Systems With Applications 297 (2026) 129405

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

BUT-Net: Boundary-Aware U-Net structure with Two-Path Transformers for lesion segmentation in mCNV using OCT images

Hai Xie [a,e], Zhenquan Wu [b], Shaobin Chen [a,f], Guanghui Yue [a], Tianfu Wang [a], Chuan-Ming Liu [d], Lin Lu [c], Guoming Zhang [b,*], Baiying Lei [a,*]

[a] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
[b] Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, China
[c] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat sen University, Guangzhou 510060, China
[d] Department of Computer Science and Information Engineering, National Taipei University of Technology (Taipei Tech), Taipei 106344, Taiwan
[e] School of Cyberspace Security, Hainan University, Haikou 570228, China
[f] Faculty of Applied Sciences, Macao Polytechnic University, Gomes Street, Macao 999078, China

ARTICLE INFO

Keywords:
mCNV Lesion Segmentation
Boundary-aware U-Net
Two-path Transformers
Dual Attention Gate

ABSTRACT

Myopic choroidal neovascularization (mCNV) is a common complication of high myopia. This disease will spread to the macular region of the retina, which has a serious impact on vision. Accurate segmentation of mCNV in retinal optical coherence tomography (OCT) plays a significant role in evaluating the condition of the disease and guiding the formulation of the treatment. However, accurate segmentation is challenging due to the blurred boundary, low contrast, tiny lesion areas, and high interference in OCT images. To this end, we propose a boundary-aware U-Net structure using two-path Transformers (i.e., BUT-Net) to accomplish lesion segmentation in mCNV using OCT images. Specifically, the encoder is built by two-path feature extractors, in which a four-stage Transformer is used to extract global features with long-distance dependency so that the network can express the morphological structure and tiny lesions. Four mixed blocks are designed by ResNet-34 and Transformer modules to excavate contextual and deep semantic features. The two-path fusion (TPF) module is utilized to fuse the two-path features by channel and spatial attention mechanism, which retains the local and global features of the fused features. For the decoder, we devise a dual attention gate (DAG) module to enhance the more complex channel dependencies in the up-sampling process. In addition, the boundary information is employed to refine the generated masks by weak supervision. The extensive experiments conducted on an in-house OCT dataset and two public different modality datasets demonstrate that our proposed method achieves superior segmentation performance. Our code is available at https://github.com/shehare9517/BUT_Net.

1. Introduction

Myopia is one of the most common fundus diseases worldwide. The growth of the ocular axis in patients with high myopia easily leads to a series of pathological vascular changes for myopia, which can result in a progressive decline of visual quality and blindness in severe cases. Among them, myopic choroidal neovascularization (mCNV) is a common complication of high myopia. The lesion involves the macular region of the retina, which has a serious impact on vision. At present, the main treatment of mCNV is to inject anti-vascular endothelial growth factor (VEGF) into the eyes, which requires multiple injections to turn the lesion from active to inactive (Cheung et al., 2017). Optical coherence tomography (OCT) is a noninvasive examination, which observes the fine structures of the retina and choroid at different levels. Hence, OCT can be used to observe the size of the lesion to judge the effectiveness of each injection treatment or the possibility of recurrence (Ding et al., 2018) and evaluate the status of the illness, and formulate the appropriate treatment in the next stage. In the procedure, the segmentation of lesions in OCT is an effective way to automatically evaluate the change of the lesions. Five representative examples of the obtained OCT images are shown in Fig. 1. In clinical, the ophthalmologists decide whether the lesion change is inactive or not via observing the lesion's

* Corresponding authors.
E-mail addresses: xiehai@hainanu.edu.cn (H. Xie), wuzhenquan@sz-eyes.com (Z. Wu), p2314850@mpu.edu.mo (S. Chen), yueguanghui@szu.edu.cn (G. Yue), tfwang@szu.edu.cn (T. Wang), cmliu@ntut.edu.tw (C.-M. Liu), lulin@gzzoc.com (L. Lu), zhangguoming@sz-eyes.com (G. Zhang), leiby@szu.edu.cn (B. Lei).

https://doi.org/10.1016/j.eswa.2025.129405
Received 5 September 2023; Received in revised form 10 August 2025; Accepted 14 August 2025
Available online 15 August 2025
0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

size; the lesion changes from active to inactive when the area becomes smaller. However, this judgment process is time-consuming and requires experienced ophthalmologists. Furthermore, different ophthalmologists may make different judgments on the OCT images of the same patient, yielding subjective differences among doctors. Therefore, an automatic method of segmenting the lesion areas in OCT images is imperative to assess the condition of mCNV. We can see from Fig. 1 that it is rather challenging to segment the lesion regions, mainly due to tiny lesions, low contrast, and boundary blur. To this end, we need to design the network with the ability to extract high-level and boundary-aware features. In general, an accurate segmentation method of mCNV lesions in OCT is crucial for evaluating the developing condition of mCNV.

To address the challenges of the inherent characteristics of OCT images, promising algorithms for OCT image segmentation have been proposed (Chen et al., 2013; Lee et al., 2009; Wu et al., 2017). For example, Chen *et al.* devised a novel automated drusen segmentation algorithm using spectral-domain OCT (SD-OCT) images by integrating 3D spatial information in retinal structures and information in projection images of drusen (Chen et al., 2013). Lee *et al.* combined a multi-scale 3D graph search method *k*-NN classifier to segment retinal layers, optic disc cup, and neuroretinal rim using SD-OCT images (Lee et al., 2009). Xiang *et al.* presented a 3D supervised layer segmentation method for macular-centered OCT images with exudative age-related macular degeneration (AMD), in which a constrained graph search algorithm was proposed to accurately detect surfaces between retinal layers to perform layer and abnormal region segmentation with high accuracy (Xiang et al., 2018). However, these traditional methods have the limited ability to extract deep features.

Convolution neural network (CNN) is considered an excellent visual model due to its powerful feature-extracting ability, and this network is widely applied for segmentation tasks (Minaee et al., 2021; Tajbakhsh et al., 2020; Wu et al., 2022b). In 2015, Long *et al.* proposed a convolution network for semantic segmentation, which opens the prelude to image segmentation using CNN (Long et al., 2015). For medical image segmentation, Ronneberger *et al.* designed a U-shape model (i.e., U-Net) (Ronneberger et al., 2015), which consists of an encoder and decoder by exploring CNN structure to implement down-sampling and up-sampling operations during training in an end-to-end system. As one of the most popular CNN frameworks, the U-Net plays an important role in medical image segmentation. Based on U-Net architecture, a variety of U-Net variants have emerged for medical image segmentation (Isensee et al., 2021; Jha et al., 2019; Oktay et al., 2018; Zhou et al., 2018). For instance, Oktay *et al.* devised an attention U-Net by adding a simple yet effective attention gate block to filter the features skipped from the down-sampling to the up-sampling path for organ segmentation in CT images (Oktay et al., 2018). To reduce the semantic map between the encoder and decoder, Zhou *et al.* presented a U-Net ++ model with a series of nested and dense skip pathways for the medical image segmentation of different modalities (Zhou et al., 2018). When a network is designed deeply, the gradient of the loss function is likely to diminish in the training process, which results in the decline of segmentation performance. To solve this issue, residual U-Net is developed by adopting residual connection in each block in U-Net (Cheng et al., 2022; Pang et al., 2023), which is applied to multiple biomedical image segmentation tasks, such as liver cancer (Li et al., 2018), prostate cancer (Zhu et al., 2019), and melanoma (Wu et al., 2020). Although these CNNs extract rich semantic information and achieve promising segmentation performance, the obtained feature information is mostly local due to the unique convolution operation of CNN and ignores global information which might improve performance.

To extract global information, the Vision Transformer framework (Vaswani et al., 2017) has been explored in medical image segmentation (Gao et al., 2021; Reza et al., 2022). For example, Gao *et al.* proposed a simple yet efficient hybrid Transformer structure (i.e., UTNet) (Gao et al., 2021), which integrates the self-attention mechanism in Transformer into CNN structure to capture the long-term dependence of different scale features for both the encoder and decoder. And hence the performance of medical image segmentation can be enhanced. Azad *et al.* presented a context-related attention network, which used the global feature expression ability of Transformer to extract long-distance dependent feature relationships of context (Reza et al., 2022). This work uses CNN to extract local semantic feature information, and takes the boundary of salient objects as weak supervision information to enhance the segmentation performance. Although these Transformer networks can extract contextual information and global features, the boundary information are ignored, which results in the generated masks with incomplete boundary.

To address the above issues, this paper proposes a two-path Transformer-based boundary-aware U-Net model for mCNV lesion segmentation in OCT images (i.e., BUT-Net), which can extract rich global and semantic features. A two-path fusion (TPF) module is devised to fuse the extracted features from two paths. A dual attention-gate (DAG) mechanism is explored to focus on the generation of lesion features in the up-sampling process. Finally, the lesion boundary is regarded as weak supervision information to enhance the expression of the boundary information. The extensive experiments show that the proposed method can effectively segment the mCNV lesions in OCT images. Also, the comparative experiments with the existing methods on two public different modal datasets verify the effectiveness of the proposed method. The main contributions can be summarized below:

1) The joint network structure of the two-path Transformer and CNN is designed as an encoder to extract rich global and contextual semantic information. The boundary information is introduced to refine the segmentation results in a weak supervision manner after the decoder.
2) Based on the spatial and channel attention mechanism, a TPF module is designed, in which the channel attention module is used to process the features obtained by the Transformer to enhance the expression of its global features and the spatial attention module is utilized to handle the features obtained by the mixed block (a module composed of ResNet and Transformer). These modules enhance the expression of local information and suppress the features of irrelevant regions.
3) We devise a DAG unit based on the channel attention mechanism, which integrates the features generated by up-sampling and TPF in a dual-fusion mode. Hence, the decoded features capture more complex channel dependencies.

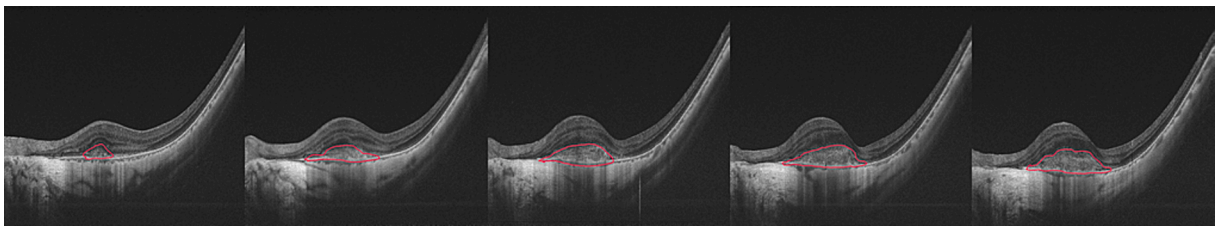


Fig. 1. Five representative examples of the collected OCT images from in-house hospital, with the red line locating the lesion in each example.

2. Related work

2.1. Lesion segmentation in OCT

Many methods were proposed to segment lesions in OCT since it is easily accessible and non-invasive (Fu et al., 2019; He et al., 2022; Mahapatra et al., 2020; Novosel et al., 2017; Wang et al., 2016). For example, Wang *et al.* applied the higher-order constraint to the conventional Markov random field (MRF) framework to keep the pixel consistency for the segmentation of the fluid-associated region in SD-OCT (Wang et al., 2016). To make full use of layer information in OCT, Novosel *et al.* proposed a method to jointly segment retinal layers and lesions in OCT images (Novosel et al., 2017). However, these conventional methods are limited to extracting deep semantic features. Hence, there are devised models based on deep learning. For instance, Xing *et al.* presented a model based on FCN to segment three types of pathological fluid lesions in OCT simultaneously, in which the attention and spatial pyramid pooling modules are employed to improve the segmentation performance (Xing et al., 2022). To extract rich local and global features, Wang *et al.* devised a novel multi-scale Transformer global attention network (i.e., MsTGANet) for drusen segmentation in retinal OCT images (M. Wang et al., 2021). However, most of the mentioned methods are proposed for lesion segmentation of diabetic retinopathy and age-related macular, and we focus our research on mCNV segmentation from OCT images.

2.2. Hybrid framework of CNNs and Transformer for medical segmentation

According to the characteristics of U-Net and Transformer, U-Net pays more attention to local context semantic information while global information can be more captured by Transformer. Hence, many researchers have studied the hybrid framework of U-Net and Transformer for medical image segmentation (Du et al., 2022; Li et al., 2022; Lin et al., 2022; Wu et al., 2023; Wu et al., 2022a,b; Yang & Yang, 2023; Wang et al., 2021; Zhang et al., 2021). For example, Zhang *et al.* presented a novel parallel architecture (i.e., TransFuse) for medical segmentation combining CNNs and Transformer in parallel mode to obtain both global dependency and low-level spatial details (Zhang et al., 2021). Wu *et al.* proposed a novel segmentation network (i.e., FAT-Net) with feature adaptive Transformers, which consists of a dual encoder with CNNs and Transformer to extract more discriminative local features and long-range dependencies for skin lesion segmentation (Wu et al., 2022a). To accomplish the accurate segmentation of tiny lesions in diabetic retinopathy (DR) images, Huang *et al.* employed U-Net as the backbone to extract features, which is implemented by two global Transformer blocks (GTBs) to generate more specific features of vessels and lesions. The obtained features are then handled at the pathological connection level by relation Transformer block (RTB) to realize multi-lesion segmentation of DR (Huang et al., 2022). Yang et al. combined a CNN and a Swin Transformer as a feature extraction backbone to build a pyramid structure network for feature encoding and decoding, and a boundary detection (BD) module was devised to pay additional attention to the boundary information of breast lesions for improving the segmentation performance of breast ultrasound images (Yang and Yang, 2023). In order to refine the boundary, Wu *et al.* presented a two-stage network for brain MRI lesion segmentation (called W-Net), which utilizes the CNN and Transformer-based architecture as the backbone network and introduces boundary deformation module (BDM) and boundary constraint module (BCM) to handle the tough condition of fuzzy boundary (Wu et al., 2023). Inspired by the above methods, we not only combine CNNs and Transformer to extract deep local semantic features and global features but also apply the boundary information as the supervised feature to produce fine segmentation mask.

3. Methodology

This paper proposes a boundary-aware based U-Net with a two-path Transformer model for the lesion segmentation of mCNV using OCT images. The devised model contains two paths to extract features. One path contains a four-stage Transformer which is utilized to extract global features with different scales, while the other path is implemented by residual CNN and Transformer structures (denoted by Mix_block, which is illustrated in Fig. 3) to acquire the contextual semantic and global features. Subsequently, the features from the above two paths are fused by the devised TPF module. For the decoder, a DAG unit is proposed to fuse the features from the encoder and up-sampling modules, which enforces the network to focus on the generation of lesions. The boundary information is introduced to refine the segmentation results by weak supervision. The whole flowchart of the proposed method is shown in Fig. 2 and we will describe the modules in detail as follows Fig. 3.

3.1. Transformer structure

As shown in Fig. 2, both feature extraction paths use Transformer. Before introducing Transformer, we give a brief explanation of its principle. The visual Transformer structure is mainly composed of a multi-head self-attention (MHSA) module, which allows the model to jointly infer salient objects in the image from different feature representation subspaces. The features processed by MHSA are cascaded together as the input of a feed-forward network, as shown in Fig. 4 (a). In this paper, the number of MHSA is determined by the number of channels in different stages, which is set to 1, 2, 4, and 8 from stage 1 to stage 4, respectively. Assuming that the input images $X \in \mathbb{R}^{H\times W\times 3}$, where H, W represent the height and width of images, respectively. The input X is first rearranged into query vector Q, key vector K and value V according to the three linear mappings, where $Q, K, V \in \mathbb{R}^{H\times W\times d}$, d represents the embedded dimension of each attention head. Q, K, V are flattened into three token sequences with the same size $n \times d$, where $n = HW$. Then the self-attention mechanism can be expressed by a scale dot product:

$$Atten(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where $\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \in \mathbb{R}^{n\times n}$ is an attention matrix, which is used to calculate the similarity of each token between the query and key vector, and further gain the context information of the value V. This attention mechanism is added to model the long dependence of features and integrate them to generate global features.

The four stage Transformer feature extractor selects the pyramid vision Transformer module (W. Wang et al., 2021), which first divides the input image into $\frac{HW}{4^2}$ blocks with a stride of 4, and each block size is 4 × 4 × 3. The scale size of feature extraction at each stage is set through the channel size of the embedding layer. Assuming that the block size of stage i is PS_i, we first divide the feature map $Fea_{i-1} \in \mathbb{R}^{H_{i-1}W_{i-1}C_{i-1}}$ into the patches having the size of $\frac{H_{i-1}W_{i-1}}{PS_i}$. Each patch is then flattened into the embedding layer with the dimension of C_i. After linear rearrangement, the embedding patch can be expressed as $\frac{H_{i-1}}{PS_i} \times \frac{W_{i-1}}{PS_i} \times C_i$. In this way, the feature scale of each stage can be flexibly adjusted to build a pyramid Transformer feature extractor.

At present, most Transformer feature extractors neglect spatial information. This results in insufficient extracted features and affects the overall performance. Inspired by (Gao et al., 2021), we introduce the relative position-coding into the Transformer structure to build t he hybrid module in Fig. 2. The 2D relative position coding is mainly realized through the relative height and width information between pixels. Supposing that there are two pixels: $A = (A_x, A_y)$ and $B = (B_x, B_y)$, which can be used to calculate the logic output before softmax by the following relative position-coding:

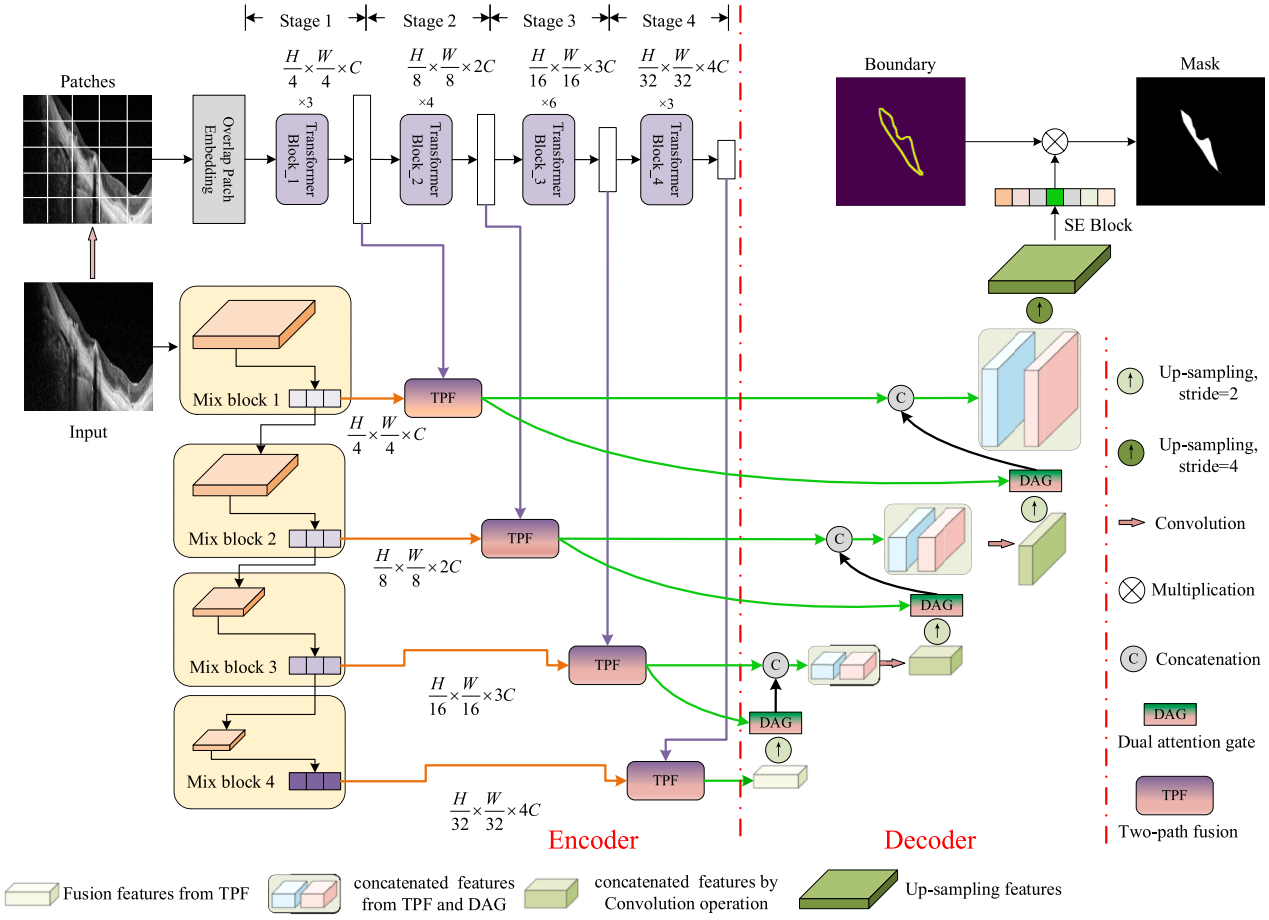


Fig. 2. Architecture of the proposed method, with a framework consisting of an encoder and decoder. The encoder is built by a two-path feature extractor, where one path is used to extract the global features of different scales by a four-stage Transformer structure and the other path is devised by the hybrid structure of residual CNN and Transformer (denoted by Mix_block) to obtain contextual local features. Subsequently, features with the same scale from two paths are fused by the devised TPF module. For the decoder, a DAG unit is proposed to fuse the features from the encoder and up-sampling modules, which enforces the network to focus on the generation of lesions. Boundary information is introduced to refine the segmentation results by weak supervision.

$$RE_{A,B} = \frac{q_A^T}{\sqrt{d}}\left(k_B + r^W_{B_x - A_x} + r^H_{B_y - A_y}\right), \tag{2}$$

where q_A is the query vector of pixel A and k_B is the key vector of pixel B. $r^W_{B_x-A_x}$ and $r^H_{B_y-A_y}$ are the embedding layer representations of relative width and height, respectively. The self-attention mechanism after the softmax layer can be expressed as

$$Atten_{RE}(Q,K,V) = softmax\left(\frac{QK^T + q_A^T r^W_{B_x-A_x} + q_A^T r^H_{B_y-A_y}}{\sqrt{d}}\right), \tag{3}$$

It can be seen from Fig. 2 that the downlink path in the proposed method adopts a mixed module, which comprises the relative position-coding Transformer and ResNet-34 model (He et al., 2016). This mixed module first utilizes the pre-training model of ResNet-34 on the ImageNet dataset to extract the high-level semantic features, and then the position-coding Transformer structure is employed to capture the long-distance dependencies.

3.2. TPF module

The features produced by the four-stage Transformer framework and the hybrid feature extraction contain rich semantic texture and global information. These features need to be further fused to form good representations of disease lesions. To address this issue, we employ the TPF module to fuse the features that are mainly composed of channel attention and position attention coding modules. We shall briefly describe these two modules in the TPF mechanism based on the definition of channel and location attention coding module in (J. Fu et al., 2019). Let $F_T \in \mathbb{R}^{C\times H\times W}$ and $F_M \in \mathbb{R}^{C\times H\times W}$ denote the extracted features by the four-stage pyramid Transformer and hybrid feature extractor, respectively. The extracted features by Transformer have long-distance dependency with less emphasis on spatial information, while the locally extracted features by CNN pay more attention to the semantic features. Hence, the extracted features by Transformer and hybrid feature extractor are processed by location and channel attention coding, respectively, as shown in Fig. 4 (b). The F_{TP} feature is obtained by TPF, which is expressed as:

$$\begin{aligned} F_{TP} = Point_CNN\big(&Sum\big(F_T, Reshape\big(M(Reshape(F_T),\, M\big(F_T, F_T^T\big)\big)\big)\big) \\ &+ Sum\big(F_M,\, Reshape\big(M\big(Reshape(F_M), M\big(F_M, F_M^T\big)\big)\big)\big)\big) \end{aligned} \tag{4}$$

where $Point_CNN(\bullet)$ is the point-wise convolution with the kernel of 1×1. $Sum(\bullet)$ and $M(\bullet)$ denote matrix addition and multiplication, respectively. $Reshape(\bullet)$ represents that the size of the feature dimension is reshaped. Through this TPF module, the features extracted by the four-stage Transformer path can be supplemented in spatial information, and the local semantics extracted by the hybrid module can be supplemented in the channel attention map. Through this, the features extracted by the two paths is effectively integrated and complemented in different feature dimensions, which improves the overall performance of the network.

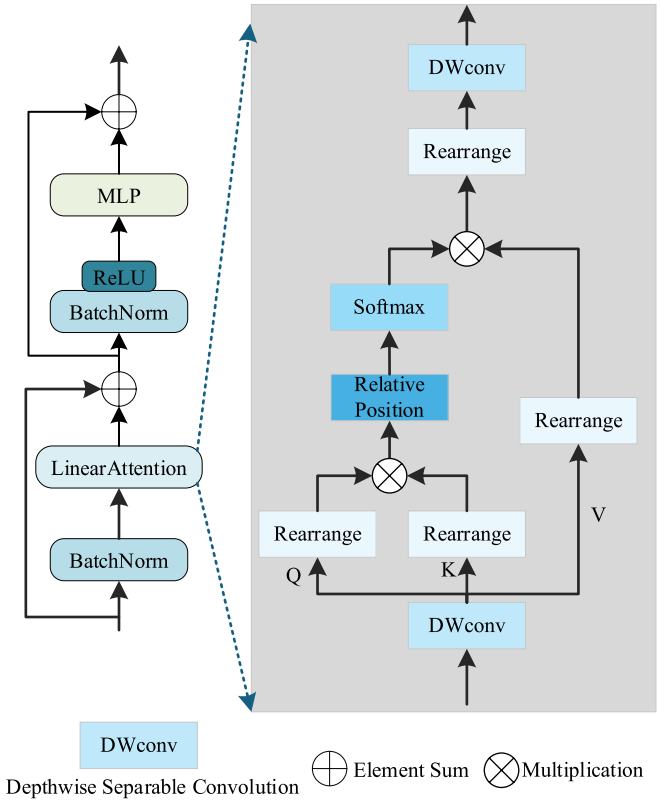


Fig. 3. The illustration of Mix_block module.

3.3. DAG module

To emphasize the generation of lesion features during the up-sampling, we use the channel attention mechanism to design a DAG unit to enhance the features F_{TP} and F_{up} from TPF and up-sampling modules, respectively, as shown in Fig. 4 (c). This approach was applied in (H. Wang et al., 2021). After obtaining features F_{TP} and F_{up}, these features are processed by the max and average pooling layers, and then activated by ReLu function. Through two fully connected layers, the number of channels of the feature is changed to $\frac{1}{16}\times\frac{1}{16}$. The final features obtained by max and average pooling are then weighted summed, which are processed with the input features by the point multiplication of the matrix to obtain the attention map. To maintain the integrity of features, the obtained attention maps are performed by the maximum and average pooling layers again, and then the features obtained from the two pooling processes will be cascaded, which is utilized to gain the single attention gating features by matrix multiplication with the above attention maps. Besides, we perform the same operation on the other input feature (such as up-sampling feature F_{up}). Finally, the weighted sum of the features obtained from the two attention modules is used to obtain the final dual attention gating feature map F_{DAG}. The formulations of the features are given as

$$F'_{TP} = Sum\left(FC(ReLU(Pool_{Max}(F_{TP}))), FC\left(ReLU\left(Pool_{Avg}(F_{TP})\right)\right)\right), \tag{5}$$

$$F''_{TP} = M(F'_{TP}, Concat(Pool_{Max}\left(M\left(F_{TP}, F'_{TP}\right)\right), Pool_{Avg}(M(F_{TP}, F'_{TP})))) \tag{6}$$

$$F'_{up} = Sum\left(FC\left(ReLU\left(Pool_{Max}\left(F_{up}\right)\right)\right), FC\left(ReLU\left(Pool_{Avg}\left(F_{up}\right)\right)\right)\right) \tag{7}$$

$$F''_{up} = M(F'_{up}, Concat(Pool_{Max}\left(M\left(F_{up}, F'_{up}\right)\right), Pool_{Avg}(M(F_{up}, F'_{up})))) \tag{8}$$

$$F_{DAG} = Sum\left(F''_{TP}, F''_{up}\right) \tag{9}$$

where $Pool_{Max}(\bullet)$ and $Pool_{Avg}(\bullet)$ denote max and average pooling, respectively. $FC(\bullet)$, $Sum(\bullet)$, $M(\bullet)$, and $Concat(\bullet)$ indicate fully-connected layer, matrix sum, matrix multiplication, and concatenation operation, respectively. Through this DAG module, the decoder pays more attention to the generation of the feature information of the lesion area, which has a positive effect on the prediction results.

3.4. Loss function

The mCNV lesion segmentation of OCT image can be seen as a binary classification task. Hence, the binary cross entropy is selected as the loss function. In addition, the boundary of OCT lesions is also used as weak supervision information, which enables the network to generate a segmentation mask with rich boundary information. This boundary information can be considered as a binary classification task, so the binary cross entropy function is also employed as the boundary loss.

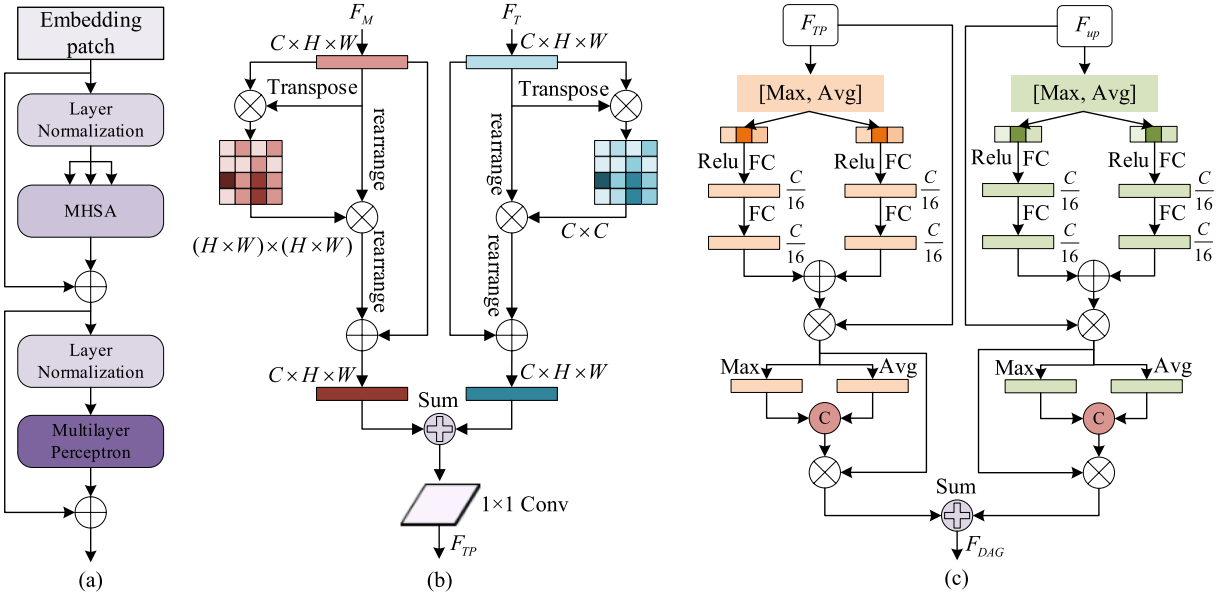


Fig. 4. Schematic diagram of each module in this method. (a)-(c) indicate the basic Transformer structure, TPF and DAG, respectively. ⊕and⊗ represent matrix summation and multiplication, respectively.

$$L_{bce_{object}} = -\sum_i [(1-GT_i)ln(1-PO_i) + GT_i lnPO_i], \tag{10}$$

$$L_{bce_{boud}} = -\sum_i \left[(1-GT_i^B)ln(1-PB_i) + GT_i^B lnPB_i\right], \tag{11}$$

where GT_i and GT_i^B are the i-th pixel belongs to the ground-truth of lesion and boundary, respectively. PO_i and PB_i are the i-th pixel belongs to the prediction of lesion and boundary, respectively.

To avoid over-fitting, the dice loss function is utilized to enable the network to accurately predict the lesion area. This loss function has been proven to have good performance and stability in medical image segmentation tasks, which is defined as

$$L_{Dice} = 1 - 2 \times \frac{\sum_i PO_i GT_i}{\sum_i PO_i + \sum_i GT_i}, \tag{12}$$

Overall, the final loss function can be expressed as

$$L_{total} = \alpha L_{bce_{object}} + \beta L_{bce_{boud}} + \gamma L_{Dice}. \tag{13}$$

where α, β, γ are the control weights of each used loss, which are set to 0.4, 0.2, 0.4 in our experiments, respectively.

4. Experiments

4.1. Experimental Materials

4.1.1. OCT images of mCNV

The OCT images of mCNV were collected from Zhongshan Ophthalmology Center of Sun Yat-Sen University, including 2402 original OCT images and their corresponding ground-truth. The annotation was jointly completed by three experts in high myopia fields. Images with poor quality are excluded from the 2402 images. All images are resized to 256 × 256 as the input of all models. The dataset is divided into training, validation, and test sets with a ratio of 8:1:1, and we have 1922 training images, 240 validation images, and 240 test images.

4.1.2. OCT images of DME

The segmentation of irregular occurrence areas of diabetic retinopathy macular edema (DME) is important for the diagnosis of patients with macular disease. The used OCT dataset of DME is publicly available from Duke University (Chiu et al., 2015), referred to as the Duke dataset. This dataset contains 10 patients with 110 OCT images having the size of 768 × 496, and are resized to 256 × 256 for our experiments. Two ophthalmologists annotated the regions of macular edema. By comparing the results of any one of the two expert labels as the gold standard in (Chiu et al., 2015), there is little difference between the two segmentation results. Hence, we select the annotation of expert 1 as the golden standard in our experiments. Due to the limited amount of this dataset, we augment this dataset by mirror, rotation, horizontal, and vertical flipping to generate 3520 OCT images.

4.1.3. ISIC 2018

To verify the generalization ability of our methods in other medical modalities, we conduct experiments on the ISIC2018 skin lesion segmentation dataset (Codella et al., 2019). This dataset contains 2594 images and ground-truth. All the original images are resized to 256 × 256. In order to keep consistent with the comparison methods, the division of data in our experiments is the same as (Reza et al., 2022), that is, the number of training, validation, and test sets are set as 1815, 259, and 520, respectively.

4.2. Experimental setup

Python is selected as the programming language and Pytorch library is the implementation framework for our development. A NVIDIA TITAN XP GPU was used to speed up the training and testing process. Adam is selected as the training optimizer, and its parameters are set as beta=(0.9, 0.997), eps = 10^{-8}. The initial learning rate of the network is set to 0.001, which is dynamically adjusted according to the change of the loss function. That is, if the loss function does not decline after 10 epochs, the learning rate is reduced by 2 times. The maximum epoch is set to 100. Considering the computing power of the hardware, the batch size is set to 4. Several commonly used segmentation evaluation indicators are employed to evaluate and analyze the experimental results, which include Accuracy (ACC), Sensitivity (SE), Specificity (SP), Intersection over Union (IoU), and Dice coefficient (Dice).

4.3. Experimental results

4.3.1. Results of in-house OCT images

To demonstrate the effectiveness of the proposed method, we selected several popular models in the field of medical image segmentation in recent years as the comparison including U-Net(Ronneberger et al., 2015), Atten U-Net(Oktay et al., 2018), Swin-UNet(Cao et al., 2021), UTNet(Gao et al., 2021), TMUNet(Reza et al., 2022), Res_UTNet (Gao et al., 2021). The quantification results for the lesion segmentation of mCNV using OCT images of the mentioned models are summarized in Table 1.

The models in this comparison are designed with U-Net as the baseline. The Swin-UNet, UTNet, TMUNet, and Res_UTNet are the classical hybrid networks of CNN and Transformer. Compared with the other models, the proposed method achieves the highest evaluation value in terms of Acc, SP, Dice and IoU, and works better than most of the compared methods in terms of SE. In particular, the Dice and IoU, the two most commonly used and objective indicators for evaluating segmentation performance, have been increased by nearly 3 % and the IoU score has been increased by more than 4.5 %, as compared with the baseline U-Net. By comparing U-Net and UTNet, it can be seen that the segmentation performance has been greatly improved by the joint Transformer network, in which the Dice and IoU have both been increased by more than 1 %. These results demonstrate that the Transformer framework can strengthen the ability to extract global features to improve the representation of lesion features. Compared with the second best-performing network in Table 1, the proposed method has a positive difference of 1.14 % and 2.65 % for the Dice and IoU, respectively. This observation suggests that our proposed model can fully integrate the local semantic features extracted by the CNN network and the global features with long-distance dependency extracted by the Transformer network, which enhances the overall segmentation performance. In addition, the qualitative analysis of the segmentation results of the comparison models is conducted and presented in Fig. 5.

From Fig. 5, we observe that poor segmentation results are obtained with the U-Net and Atten U-Net models. The prediction of the lesion area is incomplete, and some of them are undetected. It demonstrates that the network pays more attention to the semantic feature information when only CNN model is used and the learning ability of spatial and global feature information is limited. In addition, the OCT images of mCNV have a lot of noise and blurry boundary information, which easily leads to the CNN learning a lot of interference information so that the network learns the wrong information and cannot completely express the

Table 1
Quantitative comparison results of different methods on mCNV Lesion segmentation task (%).

Method	Year	Acc	SE	SP	Dice	IoU
U-Net	2015	98.94	88.19	99.42	87.70	78.09
Atten U-Net	2018	98.97	86.71	99.52	87.87	78.36
Swin-UNet	2021	98.85	89.75	99.22	86.92	75.39
UTNet	2021	99.07	89.04	99.52	89.15	80.42
Res_UTNet	2021	99.08	90.38	99.47	89.44	80.89
TMUNet	2022	99.06	**90.26**	99.49	89.32	80.56
Proposed	–	**99.20**	89.67	**99.63**	**90.58**	**82.78**

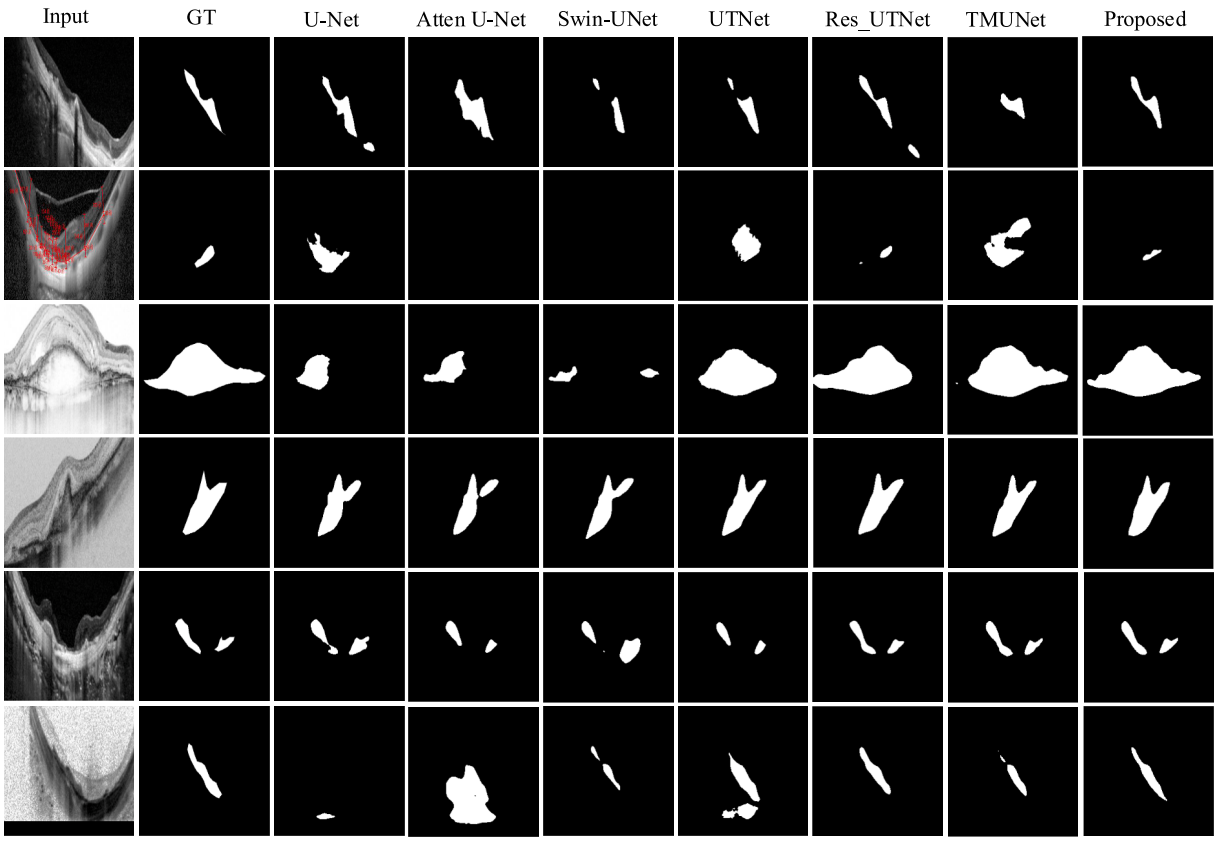


Fig. 5. Qualitative results of the comparison methods. Each column denotes the segmentation results for one method.

features of the lesions. For example, the segmentation results are presented in the second row of Fig. 5, where the lesion areas are not even detected for Atten U-Net and Swin-UNet, and only the background is shown. By observing the sixth to eighth columns, with the Transformer structure (e.g., UTNet, TMUNet, Res_UTNet), the lesion feature information can be fully expressed as the long-distance dependent global features are well learned. Compared with these models, the proposed model can not only better learn the local semantic features and long-distance dependent global features, but also learn the boundary information of the lesion and reduce fuzzy boundaries of the lesions in OCT images.

4.3.2. *Ablation results of in-house OCT image*

As shown in Fig. 2, our framework contains multiple modules. To verify the effectiveness of these modules, we perform several ablation experiments to prove the improvement for each module (i.e. Boundary (BD), DAG, and TPF). In addition, the overall loss function consists of binary cross-entropy, Dice, and boundary loss, and hence the ablation experiments are also conducted to validate the improvement of each loss function. Note that the boundary loss and boundary module are integrated, and thus, only a set of relevant experiments can be done. The segmentation results are summarized in Table 2.

It can be seen that each module has a positive effect on improving the segmentation performance (i.e., the segmentation performance will decline when one of the modules is removed). In particular, the performance of the network declines sharply when the TPF module is abolished. Specifically, the Dice and IoU declined by 1.19 % and 1.97 %, respectively. This is attributed to the fact that the network only contains a single path feature extraction module to build the encoder when the TPF module is removed. The lack of Transformer framework results in the loss of some global feature information with long-distance dependency and thus reduces the segmentation performance.

To further illustrate the effectiveness of each module, we also conduct a qualitative analysis of the segmentation results for each module, which are illustrated in Fig. 6. The red boxes indicate the segmentation difference of different modules and the green boxes represent the error segmentation result that the module is not used. It can be seen from the third column NoTPF that the prediction results without using the TPF module are inadequsate. Also, the boundary cannot be completely expressed without the weak supervision information from the boundary. These results indicate that the TPF module and the weak supervision of the boundary are quite effective, which can enforce the network to learn rich global and boundary information to improve the segmentation performance for mCNV lesions. It is observed from the image in the third column of the third row that many important features are missed without the TPF module, resulting in incomplete

Table 2
Quantitative comparison results of different methods on mCNV lesion segmentation task (%).

BD	DAG	TPF	BCE	Dice	Acc	SE	SP	Dice	IoU
	√	√	√	√	99.18	89.32	99.62	90.37	82.42
√		√	√	√	99.18	90.14	99.59	90.45	82.56
√	√		√	√	99.08	90.02	99.49	89.39	80.81
√	√	√		√	99.17	89.63	99.59	90.26	82.25
√	√	√	√		99.16	**90.48**	99.55	90.31	82.33
√	√	√	√	√	**99.20**	89.67	**99.63**	**90.58**	**82.78**

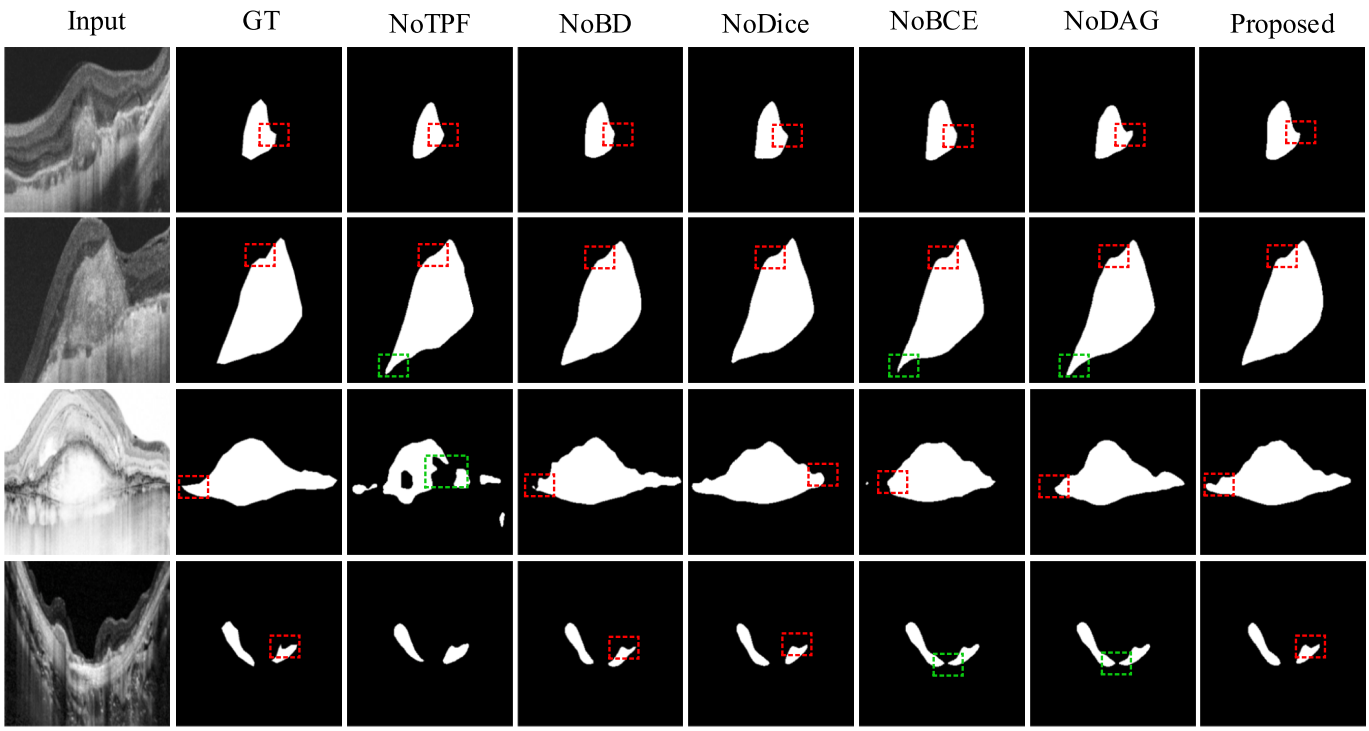


Fig. 6. Qualitative results analysis of ablation experiments with different modules and loss functions. NoTPF, NoBD, NoDice, NoBCE, and NoDAG represent segmentation results without using TPF, boundary, Dice loss, cross-entropy loss, and DAG, respectively.

segmentation results. In addition, it can be seen from the second and fourth rows of the penultimate column that the network has learned the wrong lesion areas, which demonstrates that the DAG module is effective and enforces the network to pay more attention to the generation of information related to the lesion characteristics. Therefore, the segmentation performance can be boosted.

As mentioned in Section 3.4, we set α, β, γ to 0.4, 0.2, 0.4 as default for the final loss function, such setting is to ensure that the weights sum of coefficients is 1. We here empirically demonstrate why we choose such values by conducting experiments that set them to different values. Table 3 presents the results, from which we can see that the proposed method yields the best results on average when α, β, γ are set to 0.4, 0.2, 0.4 for the final loss function.

Furthermore, we compute the attention maps obtained by the DAG module to demonstrate the effectiveness of the used attention mechanism in our framework. The results are provided in Fig. 7, from which we can observe that the attention module assists the whole framework in identifying the location of the lesion, thereby helping the segmentation network to produce a more accurate result of lesion areas.

4.4. Validation of other datasets

4.4.1. Results of the public OCT dataset

The macular degeneration is a major cause of blindness. The lesion of the macular retina is an important indicator for discovering this disease. If not found and treated timely, then the patients are more likely to be blind consequently. OCT images can also capture the micro-structure in the retina and are widely used for examining and diagnosing fundus diseases. Hence, the segmentation of lesions in the macular retina is significant for diagnosing and tracking macular degeneration using OCT images. Similar to (Rashno et al., 2017), we select the OCT dataset of DME from Duke University as the validation dataset to prove the effectiveness of our method. The experimental results are reported in Table 4. We only report the results obtained by the annotation of expert 1 in (Rashno et al., 2017), and the other results are gained by U-Net (Ronneberger et al., 2015), AttU-Net (Oktay et al., 2018), TransUNet (Chen et al., 2021), UTNet (Gao et al., 2021), TMUNet (Reza et al., 2022) and our proposed method using the augmented dataset.

We can see that our method achieves the highest Dice and IoU with 80.04 % and 66.73 %, respectively. Due to the hybrid framework of CNN and Transformer of the proposed method, deep and contextual features are extracted from the Duke dataset with tiny lesions. Furthermore, the augmented data can effectively avoid over-fitting in the process of training. The main reason is that the lesion is small, and some images are out of focus, which leads to high Acc and SP values in the prediction results for all the models.

We perform the segmentation visualization of different methods, which is shown in Fig. 8. We observe that compared with other models, the segmentation masks obtained by our proposed model are closest to the ground-truth. From the third row, we discover that the results from UNet and Att-UNet are better than those of TransUNet, which demonstrates that the contextual semantic information obtained by CNN is significant for the prediction of tiny lesions. Hence, the combination of CNN and Transformer in the proposed method is effective and feasible.

Table 3
The results of different weight parameters for final loss function (%).

α	β	γ	Acc	SE	SP	Dice	IoU
0.3	0.5	0.2	99.18	88.93	99.64	90.36	82.41
0.3	0.3	0.4	99.19	89.06	99.62	90.41	82.51
0.4	0.4	0.2	99.19	88.88	**99.66**	90.45	82.57
0.4	0.2	0.4	**99.20**	**89.67**	99.63	**90.58**	**82.78**
0.5	0.3	0.2	99.18	89.78	99.60	90.39	82.47
0.5	0.2	0.3	99.15	89.31	99.60	90.12	82.02

4.4.2. Results of ISIC2018 dataset

The skin lesion segmentation is considered to be an important step in the computer-aided diagnosis of melanoma. Recently, many studies have been performed to propose new methods for skin lesion segmentation. From CNN to Transformer, as well as the joint framework of CNN and Transformer, the technology and performance of depth feature extraction is constantly enhanced. As a representative dataset of the

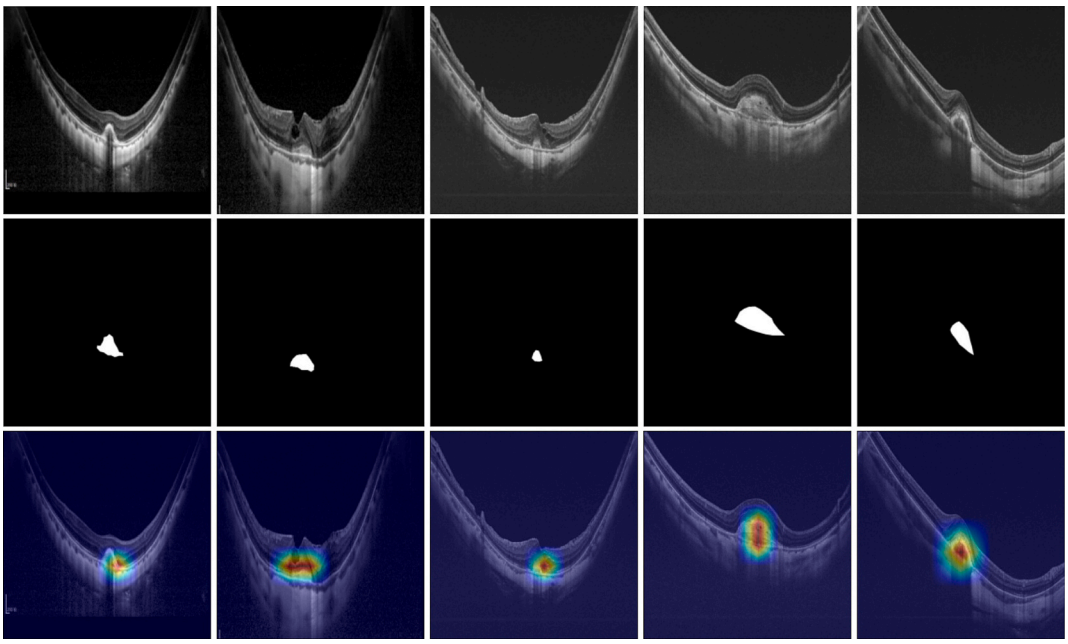


Fig. 7. Attention maps obtained by the used attention module in our method. From top to bottom rows: original images, ground-truth, and attention maps, respectively.

Table 4
Quantitative comparison results of different methods on DME lesion segmentation task from OCT images (%).

Method	Year	Acc	SE	SP	Dice	IoU
(Chiu et al., 2015)	2015	–	73.19	–	62.36	–
(Rashno et al., 2017)	2018	–	73.26	–	53.33	–
U-Net	2015	99.77	64.92	99.98	76.89	62.46
AttU-Net	2019	99.76	64.40	99.97	76.14	61.48
TransUNet	2021	99.77	72.54	99.93	78.73	64.92
UTNet	2021	99.77	67.03	99.96	77.17	62.83
TMUNet	2022	99.80	68.05	99.96	78.16	63.59
Proposed	–	99.79	72.38	99.95	**80.04**	**66.73**

challenge of skin segmentation, the ISIC2018 dataset is selected as the validation dataset to verify the effectiveness of the lesion segmentation task of dermoscopic images using the proposed method. Referencing (Reza et al., 2022; Wu et al., 2022a), we have selected the following eight methods as comparison models, namely, U-Net (Ronneberger et al., 2015), AttU-Net (Oktay et al., 2018), DAGAN (Lei et al., 2020), CKDNet (Jin et al., 2021), TransUNet (Chen et al., 2021), UTNet (Gao et al., 2021), FAT-Net (Wu et al., 2022a), TMUNet (Reza et al., 2022), AS-Net (Hu et al., 2022). To keep consistent with the experimental results in (Reza et al., 2022; Wu et al., 2022a), we directly report their results in Table 5 to make a quantitative comparison with the proposed method.

From Table 5, we see that the compared methods contain CNN, Transformer and the combination of CNN and Transformer. Compared with other methods, the proposed method achieves the best segmentation performance with 96.12 %, 97.73 %, 90.65 %, and 82.90 % in terms of the segmentation metrics of Acc, SP, Dice, and IoU, respectively. Comparing TransUNet with U-Net, it can be seen that the segmentation performance only by Transformer framework is not as good as that by CNN alone. There are two reasons: First, Transformer pays more attention to the global features and has limited expression of local semantic features. Second, the network is prone to over-fitting due to the limited training data so that the extraction of features is insufficient. However, the segmentation performance will be significantly improved when these two structures are combined, TMUNet is a good model combining these two structures, which obtains Dice and IoU of 90.59 % and 82.65 %, respectively. However, TMUNet only uses a Transformer structure. On this basis, we add another Transformer path to build a two-path feature extraction module with Transformer, which effectively extracts and fuses the global and local features from the two paths. From Table 5, we can see that our designed model is effective. In addition, compared with traditional CNN (such as U-Net), the proposed method has improved the Dice and IoU by more than 5 %, which also proves that the effective combination of CNN and Transformer structures is very helpful to the improvement of segmentation performance.

We also conducted the qualitative analysis of segmentation results of different methods on the ISIC2018 dataset, which are shown in Fig. 9. By intuitively comparing with the ground-truth, we can observe that our proposed method obtains the best segmentation results, in which the lesion areas are segmented well and the boundary information is approximately expressed in its entirety. For example, for the segmentation result of the first row, even if the contrast between the lesion and the background area is low, the proposed method can completely segment the lesion area. However, some lesion areas cannot be segmented correctly by other methods, and even the segmentation result of the joint Transformer framework is not as good as that of the pure CNN networks (e.g., U-Net and AttU-Net). This is due to the low contrast of some images, and the semantic features extracted by CNN and the global features extracted by Transformer are not effectively integrated, leading to insufficient complementation and utilization of the features. In addition, from the third row, we can see that all methods achieve poor segmentation results, but the proposed method still achieves good segmentation results since it benefits from the effective integration of CNN and Transformer structures.

5. Discussion

5.1. The effect of boundary loss

As mentioned in Section 3.4, we employ the BCE function as the BD loss to refine the boundary information. We here empirically demonstrate its effectiveness by comparing the proposed method with its variant that replaces the BCE function with the MSE function. On all three datasets, the BCE function beats MSE functions for almost all

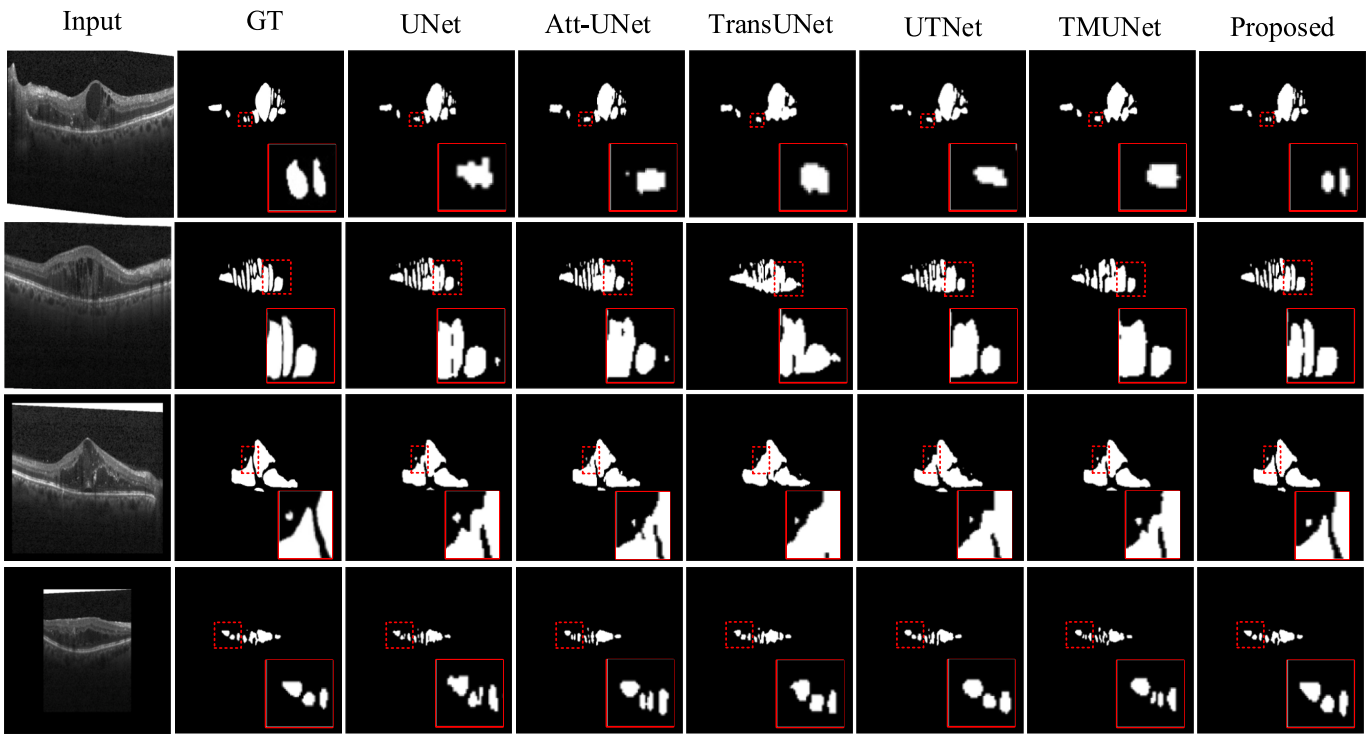


Fig. 8. Comparison results of different methods on the task of DME lesion segmentation of the Duke dataset, where the red frames represent the enlarged lesion areas with the small target. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Quantitative comparison results of different methods on ISIC 2018 skin lesion segmentation task (%).

Method	Year	Acc	SE	SP	Dice	IoU
U-Net	2015	94.04	88.00	96.97	85.45	77.33
AttU-Net	2019	93.76	86.00	98.26	85.66	77.64
DAGAN	2020	93.24	90.72	95.88	88.07	81.13
CKDNet	2021	94.92	90.55	97.01	87.79	80.41
TransUNet	2021	94.52	85.78	96.53	84.99	76.65
UTNet	2021	95.06	88.39	96.82	88.20	78.89
Res_UTNet	2021	95.64	88.69	97.47	89.47	80.94
FAT-Net	2022	95.78	91.00	96.99	89.03	82.02
TMUNet	2022	96.03	90.38	97.46	90.59	82.65
AS-Net	2022	95.68	**93.06**	94.69	89.55	–
Proposed	–	**96.12**	90.03	**97.73**	**90.65**	**82.90**

metrics in all scenarios, except for the SE metric in OCT of mCNV and OCT of DME datasets, as shown in Table 6.

5.2. The analysis of computational complexity

In order to fully evaluate the proposed method, we here analyze the computational complexity of our method against existing methods by comparing the total training time, GPU memory footprint, and the number of model parameters. As described in Section 4.2, all the experiments are conducted on one NVIDIA TITAN XP GPU card (Section 4.2), and we here report the analysis results under the same conditions on the OCT of the mCNV dataset. The analysis data is presented in Table 7. Compared to networks with only CNN or Transformer, our network has higher training time, GPU consumption, and parameters. This is because the hybrid network we designed includes both CNN and Transformer networks, resulting in better experimental results. When compared to TMUNet, which is also a hybrid network, our network performs better in various metrics.

5.3. Limitations and future work

The proposed method has achieved superior segmentation performance on multiple modal data as demonstrated above. However, it is not always as excellent as expected when the images are imaged with significant interference and low contrast (e.g., OCT images of mCNV); see Fig. 10 for such failure cases. Here note that our method indeed fails to yield the complete lesion segmentation mask in such a scenario but it still works best than other methods as we show in the comparative analysis parts. This limitation, however, may be handled by designing modules with stronger feature expression ability to force the network to pay more attention to the spatial, semantic, and boundary features of the lesion areas. Another one is that our method can be designed more computationally effective. The current version leverages multiple paths and employs two Transformer frameworks, causing the computational efficiency of the whole framework cannot reach its theoretical bound. In future work, we shall develop the whole framework with a lightweight network while reserving its strong ability for high-level feature extraction, especially for data with low imaging contrast and blurry boundaries. We shall also investigate the network fusion mechanism of multimodal data that can capture the correlation among different modal data and thus improve the segmentation performance.

6. Conclusions

In this paper, a boundary-aware U-Net segmentation model based on a two-path Transformer is proposed to achieve fine lesion segmentation of mCNV using OCT images. Specifically, a Transformer framework with a four-stage feature extractor is used to extract the global features of OCT images, and a hybrid module of CNN and Transformer is employed to extract local and global features. A TPF module is designed by making full use of the channel and position attention modules to fuse the local and global features from the above two paths. Hence, the extracted features have rich semantics and global features with long-distance dependencies. For the decoder, a dual attention gating mechanism is designed to force the decoder to pay more attention to the generation of

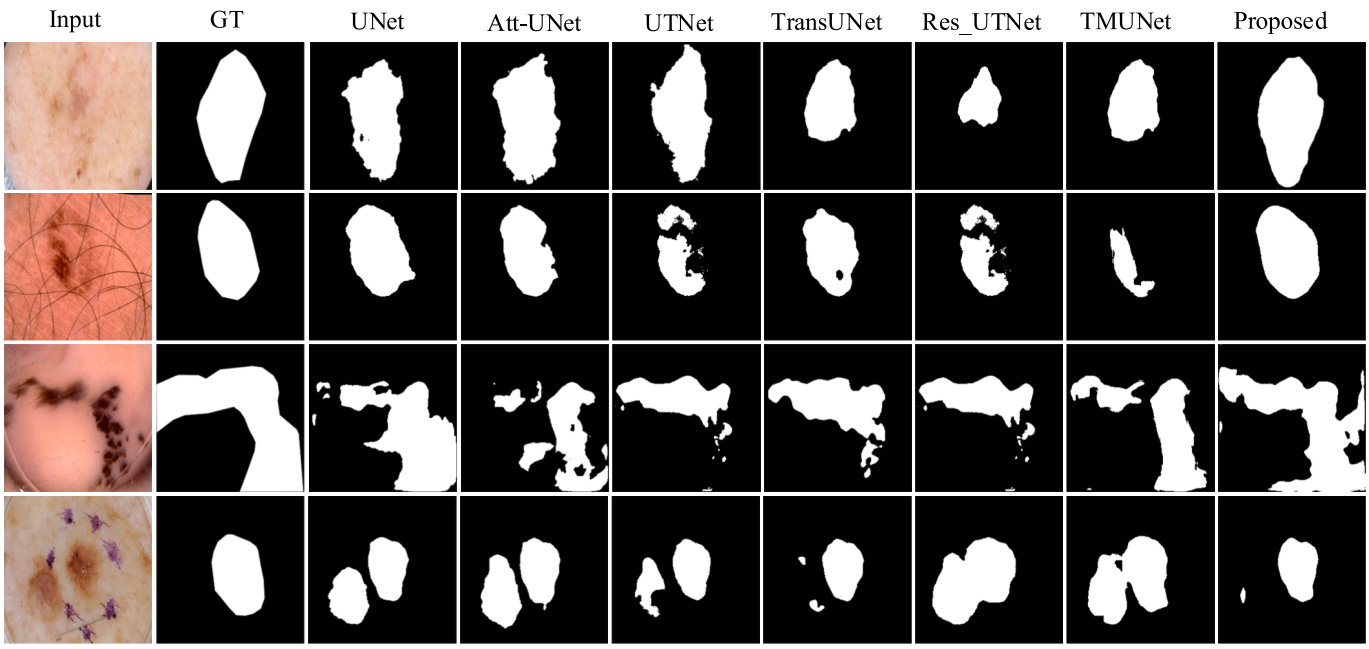


Fig. 9. Comparison results of different methods on the task of melanoma lesion segmentation of the ISIC 2018 dataset.

Table 6
Quantitative comparison results on the used three datasets when the BD loss is set to MSE and BCE (%).

BD	OCT of mCNV					OCT of DME					ISIC2018				
Loss	Acc	SE	SP	Dice	IoU	Acc	SE	SP	Dice	IoU	Acc	SE	SP	Dice	IoU
MSE	99.15	**90.18**	99.55	90.16	82.09	99.03	**74.43**	97.60	78.91	65.16	95.94	89.30	97.69	90.18	82.13
BCE	**99.20**	89.67	**99.63**	**90.58**	**82.78**	**99.79**	72.38	**99.95**	**80.04**	**66.73**	**96.12**	**90.03**	**97.73**	**90.65**	**82.90**

Table 7
Comparison of the total training time, required GPU memory, and total parameters of the network by different methods on the mCNV dataset.

Method	Training time	GPU memory	Param(M)
U-Net	3.3 h	3.26G	34.5
Atten U-Net	3.5 h	3.56G	34.88
Swin-UNet	6.5 h	7.2G	68.89
UTNet	3.2 h	3.29G	10.02
Res_UTNet	5 h	4.3G	46.19
TMUNet	12.5 h	9.8G	165.2
Proposed	8.7 h	6.7G	54.1

key lesion features. In addition, boundary information is also adopted as weak supervision to ensure that the predictive segmentation mask has a complete boundary structure. The extensive experiments have proved that our proposed method is effective and feasible, and can achieve fine lesion segmentation of mCNV from OCT images. At the same time, we compare our method with the current optimal segmentation algorithms to verify the effectiveness and universality of the algorithm using three public datasets with different modalities. The experimental results also demonstrate that the proposed method achieves superior segmentation performance. Therefore, in clinical practice, our method can provide clinicians with accurate lesion segmentation results to assist ophthalmologists in observing the size of lesions of mCNV and to timely adjust treatment and prevention programs.

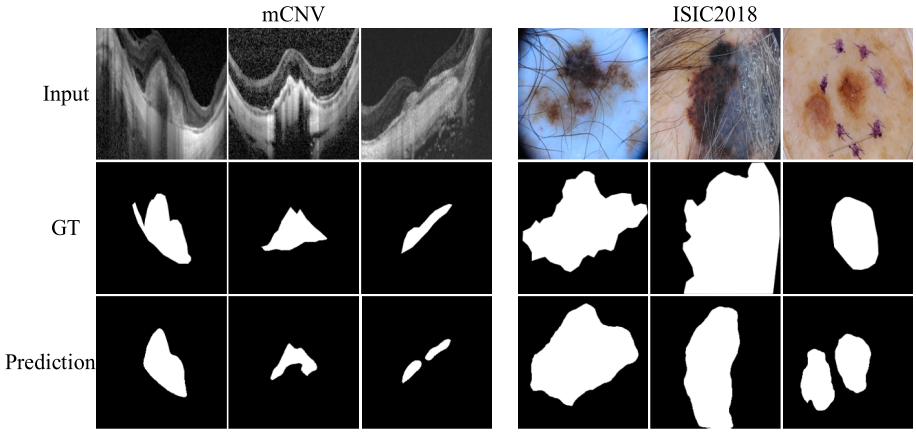


Fig. 10. Failure cases of our mCNV data and ISIC2018 dataset.

CRediT authorship contribution statement

Hai Xie: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Zhenquan Wu:** Conceptualization, Validation. **Shaobin Chen:** Methodology, Validation, Software. **Guanghui Yue:** Validation. **Tianfu Wang:** Supervision, Project administration. **Chuan-Ming Liu:** Writing – review & editing. **Lin Lu:** Supervision, Project administration, Data curation. **Guoming Zhang:** Supervision, Project administration. **Baiying Lei:** Conceptualization, Methodology, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported partly by the National Natural Science Foundation of China (Nos. 62376164,62106153, 62271328, U22A2024, and 62171312), Shenzhen Medical Research Funds (No.C2301005), Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515011676, 2024A1515011950, 2022A1515012326 and 2021A1515011348), National Natural Science Foundation of Hainan Province (No. 725MS068), Shenzhen Science and Technology Program (Grant Nos. JCYJ20241202124202004, JCYJ20240813142111015 and JCYJ20220818095809021), and Medical-Engineering Interdisciplinary Research Foundation of Shenzhen University (Project no.2023YG003), Shenzhen Key Medical Discipline Construction Fund (No. SZXK038), Shenzhen Fund for Guangdong Provincial High-level Clinical Key Specialties (No. SZGSP014), Postdoctoral Innovation Achievement Funding Program of Health Science Center of Shenzhen University, NTUT-SZU Joint Research Program.

Data availability

Data will be made available on request.

References

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv: 2105.05537*.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Q., Leng, T., Zheng, L., Kutzscher, L., Ma, J., de Sisternes, L., & Rubin, D. L. (2013). Automated drusen segmentation and quantification in SD-OCT images. *Med. Image Anal., 17*(8), 1058–1072.

Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., Wu, W., Liu, S., & Lu, H. (2022). ResGANet: Residual group attention network for medical image classification and segmentation. *Med. Image Anal., 76*, Article 102313.

Cheung, C. M. G., Arnold, J. J., Holz, F. G., Park, K. H., Lai, T. Y., Larsen, M., Mitchell, P., Ohno-Matsui, K., Chen, S.-J., & Wolf, S. (2017). Myopic choroidal neovascularization: Review, guidance, and consensus statement on management. *Ophthalmology, 124*(11), 1690–1711.

Chiu, S. J., Allingham, M. J., Mettu, P. S., Cousins, S. W., Izatt, J. A., & Farsiu, S. (2015). Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomed. Opt. Express, 6*(4), 1172–1194.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., & Marchetti, M. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Ding, X., Zhan, Z., Sun, L., Yang, Y., Li, S., Zhang, A., Luo, X., & Lu, L. (2018). Retinal pigmental epithelium elevation and external limiting membrane interruption in myopic choroidal neovascularization: Correlation with activity. *Graefes Arch. Clin. Exp. Ophthalmol., 256*(10), 1831–1837.

Du, H., Wang, J., Liu, M., Wang, Y., & Meijering, E. (2022). SwinPA-Net: Swin Transformer-based Multiscale Feature Pyramid Aggregation Network for Medical image Segmentation. *IEEE Trans. Neural Networks Learn. Syst.*. https://doi.org/10.1109/TNNLS.2022.3204090

Fu, H., Xu, Y., Lin, S., Wong, D. W. K., Baskaran, M., Mahesh, M., Aung, T., & Liu, J. (2019). Angle-closure detection in anterior segment OCT based on multilevel deep network. *IEEE Trans. Cybern., 50*(7), 3358–3366.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3146–3154).

Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: A hybrid transformer architecture for medical image segmentation. In *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 61–71).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

He, X., Fang, L., Tan, M., & Chen, X. (2022). Intra-and inter-slice contrastive learning for point supervised OCT fluid segmentation. *IEEE Trans. Image Process., 31*.

Hu, K., Lu, J., Lee, D., Xiong, D., & Chen, Z. (2022). AS-Net: Attention Synergy Network for skin lesion segmentation. *Expert Syst. Appl., 201*, Article 117112.

Huang, S., Li, J., Xiao, Y., Shen, N., & Xu, T. (2022). RTNet: Relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Trans. Med. Imaging, 41*(6), 1596–1607.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods, 18*(2), 203–211.

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., & Johansen, H. D. (2019). Resunet++: An advanced architecture for medical image segmentation. In *In 2019 IEEE International Symposium on Multimedia (ISM)* (pp. 225–2255).

Jin, Q., Cui, H., Sun, C., Meng, Z., & Su, R. (2021). Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. *Appl. Soft Comput., 99*, Article 106881.

Lee, K., Niemeijer, M., Garvin, M. K., Kwon, Y. H., Sonka, M., & Abramoff, M. D. (2009). Segmentation of the optic disc in 3-D OCT scans of the optic nerve head. *IEEE Trans. Med. Imaging, 29*(1), 159–168.

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., & Wang, S. (2020). Skin lesion segmentation via generative adversarial networks with dual discriminators. *Med. Image Anal., 64*, Article 101716.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., & Heng, P.-A. (2018). H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging, 37*(12), 2663–2674.

Li, Y., Zhang, Y., Liu, J.-Y., Wang, K., Zhang, K., Zhang, G.-S., Liao, X.-F., & Yang, G. (2022). Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation. *IEEE Trans. Cybern., 53*(9), 5826–5839.

Lin, A., Xu, J., Li, J., & Lu, G. (2022). ConTrans: Improving Transformer with Convolutional attention for Medical image Segmentation. In *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 297–307).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Mahapatra, D., Bozorgtabar, B., & Shao, L. (2020). Pathological retinal region segmentation from oct images using geometric relation based augmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9611–9620).

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell., 44*(7), 3523–3542.

Novosel, J., Vermeer, K. A., De Jong, J. H., Wang, Z., & Van Vliet, L. J. (2017). Joint segmentation of retinal layers and focal lesions in 3-D OCT data of topologically disrupted retinas. *IEEE Trans. Med. Imaging, 36*(6), 1276–1286.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., & Kainz, B. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Pang, S., Du, A., Orgun, M. A., Wang, Y., Sheng, Q. Z., Wang, S., Huang, X., & Yu, Z. (2023). Beyond CNNs: Exploiting further inherent symmetries in medical image segmentation. *IEEE Trans. Cybern., 53*(11), 6776–6787.

Rashno, A., Koozekanani, D. D., Drayna, P. M., Nazari, B., Sadri, S., Rabbani, H., & Parhi, K. K. (2017). Fully automated segmentation of fluid/cyst regions in optical coherence tomography images with diabetic macular edema using neutrosophic sets and graph algorithms. *IEEE Trans. Biomed. Eng., 65*(5), 989–1001.

Reza, A., Moein, H., Yuli, W., & Dorit, M. (2022). Contextual attention network: Transformer meets u-net. *arXiv preprint arXiv:2203.01932*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241).

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., & Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal., 63*, Article 101693.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (p. 30).

Wang, H., Cao, P., Wang, J., & Zaiane, O. R. (2021). UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. *arXiv preprint arXiv:2109.04335*.

Wang, M., Zhu, W., Shi, F., Su, J., Chen, H., Yu, K., Zhou, Y., Peng, Y., Chen, Z., & Chen, X. (2021). MsTGANet: Automatic Drusen Segmentation from Retinal OCT Images. *IEEE Trans. Med. Imaging, 41*(2), 394–406.

Wang, T., Ji, Z., Sun, Q., Chen, Q., Yu, S., Fan, W., Yuan, S., & Liu, Q. (2016). Label propagation and higher-order constraint-based segmentation of fluid-associated regions in retinal SD-OCT images. *Inf. Sci., 358*, 92–111.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568–578).

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., & Wen, Z. (2022a). FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal., 76*, Article 102327.

Wu, H., Pan, J., Li, Z., Wen, Z., & Qin, J. (2020). Automated skin lesion segmentation via an adaptive dual attention module. *IEEE Trans. Med. Imaging, 40*(1), 357–370.

Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., & Qin, J. (2022b). PolypSeg+: A Lightweight Context-Aware Network for Real-Time Polyp Segmentation. *IEEE Trans. Cybern., 53*(4), 2610–2621.

Wu, M., Chen, Q., He, X., Li, P., Fan, W., Yuan, S., & Park, H. (2017). Automatic subretinal fluid segmentation of retinal SD-OCT images with neurosensory retinal detachment guided by enface fundus imaging. *IEEE Trans. Biomed. Eng., 65*(1), 87–95.

Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., & Li, J. (2023). W-Net: A boundary-enhanced segmentation network for stroke lesions. *Expert Syst. Appl., 230*, Article 120637.

Xiang, D., Tian, H., Yang, X., Shi, F., Zhu, W., Chen, H., & Chen, X. (2018). Automatic segmentation of retinal layer in OCT images with choroidal neovascularization. *IEEE Trans. Image Process., 27*(12), 5880–5891.

Xing, G., Chen, L., Wang, H., Zhang, J., Sun, D., Xu, F., Lei, J., & Xu, X. (2022). Multi-scale pathological fluid segmentation in OCT with a novel curvature loss in convolutional neural network. *IEEE Trans. Med. Imaging, 41*(6), 1547–1559.

Yang, H., & Yang, D. (2023). CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Syst. Appl., 213*, Article 119024.

Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. In *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 14–24).

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3–11).

Zhu, Q., Du, B., & Yan, P. (2019). Boundary-weighted domain adaptive neural network for prostate MR image segmentation. *IEEE Trans. Med. Imaging, 39*(3), 753–763.