Neurocomputing 538 (2023) 126183

Contents lists available at ScienceDirect

Neurocomputing

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

PKRT-Net: Prior knowledge-based relation transformer network for optic cup and disc segmentation

Shuai Lu [a], He Zhao [a], Hanruo Liu [b,c], Huiqi Li [a,b,*], Ningli Wang [c,*]

[a] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
[b] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[c] Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology & Visual Science Key Lab, Beijing China

ARTICLE INFO

Article history:
Received 25 April 2022
Revised 25 February 2023
Accepted 28 March 2023
Available online 31 March 2023

Keywords:
Optic cup segmentation
Optic disc segmentation
Medical image processing
Deep learning

ABSTRACT

Glaucoma causes irreversible vision loss, and early detection of glaucoma is essential to protect the vision of patients. The optic cup (OC) and optic disc (OD) are two critical anatomical structures for glaucoma diagnosis. Methods based on convolutional neural networks (CNNs) have been proposed to extract OC and OD, in which OC extraction is very challenging. However, the clinical prior knowledge is not fully utilized in existing CNN methods, which limits the performance of extracting OC and OD. Besides, CNN methods cannot learn long-range semantic information interaction well due to the intrinsic locality of convolution operations. In this paper, we propose a **P**rior **K**nowledge-based **R**elation **T**ransformer **N**etwork (PKRT-Net), which employs the clinical prior knowledge to assist OC segmentation and model efficient long-range relation of spatial features by the transformer. PKRT-Net consists of a dual-branch module, a relation transformer fusion module, and a decoder with weighted fusion. Dual-branch module decouples the fundus image into the vessel feature space and general local feature space; the relation transformer fusion module fuses the clinical prior information with local features to obtain more representative features; the weighted fusion module fuses the multi-scale side-outputs from the decoder with the representation of relation transformer module to improve the segmentation performance. We evaluate our proposed PKRT-Net on three public available OC and OD segmentation datasets (*i.e.*, Drishti-GS, RIM-ONE(r3), and REFUGE). The experimental results demonstrate that our proposed PKRT-Net framework achieves state-of-the-art OC and OD segmentation results on these three public datasets.

1. Introduction

Glaucoma is one of the most severe diseases that cause blindness worldwide. By 2040, the number of glaucoma patients will reach 110 Million [1]. Glaucoma is highly concerned in ophthalmology as the visual impairment of glaucoma patients is irreversible [2]. When glaucoma patients perceive their vision loss, severe and irreversible degeneration has already occurred in their optic nerve fiber layer [3]. If early-stage glaucoma is detected in the screening, reasonable treatment can effectively control the vision deterioration of the early-stage glaucoma patients [2]. Thus, large-scale glaucoma screening is essential for the prevention of vision loss [4]. However, early diagnosis of glaucoma usually requires experienced ophthalmologists, and the number of available specialists cannot meet the huge demand for glaucoma screening. Therefore, efficient automatic glaucoma detection technology is very important to achieving large-scale glaucoma screening.

Fundus image is a standard imaging technique for glaucoma detection [5–8]. Glaucoma specialists use fundus images to diagnose glaucoma based on two anatomical structures in the fundus images: optic cup (OC) and optic disc (OD). As shown in Fig. 1(a), the closed curves in green and blue represent the OC and OD boundaries, respectively. OD is the bright oval area in the fundus image, and it is the place where the optic nerve fibers enter the eye [9,10]. The OC is a bright cup-shaped area in OD. OC's boundary is unclear due to information loss from the 3D retina to the 2D image projection [11]. The information loss makes it very difficult to determine the OC's boundary, as shown in Fig. 1 and 2. Senior glaucoma specialists use vessel kinks to aid in the extraction of OC from 2D fundus images [12]. Vessel kink has a well-defined geometric definition as expert knowledge, it is the bending of the blood vessel at the OC's boundary. Fig. 1(b) shows the process of vessel kink from 3D projection to the 2D image. The OC boundaries

* Corresponding authors.
E-mail address: huiqili@bit.edu.cn (H. Li).

https://doi.org/10.1016/j.neucom.2023.03.044
0925-2312/

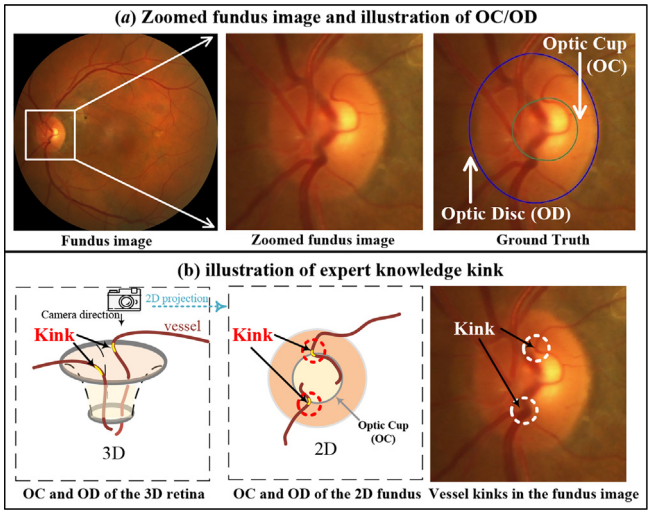

Fig. 1. The structure of the optic cup (OC) and optic disc (OD) in the color fundus image. (a) Illustration of the OC and OD in a zoomed fundus image. (b) Illustration of kink. **Kinks (*i.e.*, expert knowledge)** are bending of the blood vessel at the OC's boundary.

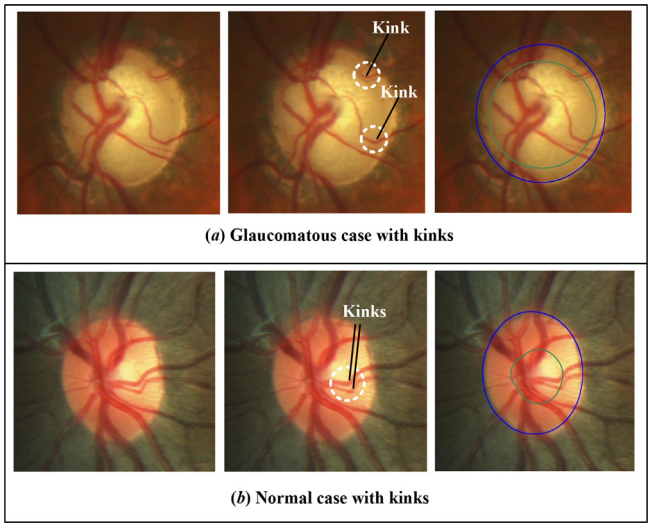

Fig. 2. Correlation between vessel kinks and OC boundaries. The closed curves in green and blue represent the optic cup and optic disc boundaries, respectively. **Kinks** are bending of the blood vessel at the OC's boundary. (a) Glaucomatous case with unclear boundaries requires kinks to determine the OC boundaries. (b) Normal case with unclear boundaries requires kinks to determine the OC boundaries.

and kinks in the 2D fundus image are highly correlated [13], as can be observed in Fig. 1 and 2.

Recently a number of works have been proposed for OC and OD segmentation. These methods can be divided into traditional methods based on hand-crafted features and deep learning based on CNNs. Traditional methods mainly include edge detection methods, thresholding methods, color difference methods, and superpixel methods [14–27]. These methods mainly perform image segmentation based on hand-crafted features, which are easily affected by the image quality and noisy lesions. Compared with traditional methods, convolutional neural networks (CNNs) can automatically extract features from images. Many CNN-based variants have been proposed to segment OC and OD [28–37]. These CNN-based methods achieve better performance than hand-crafted feature based methods.

However, the OC and OD segmentation methods based on CNNs are more inclined to focus on the image's local edge features and lack the ability to model long-range relations. When CNNs encounter fundus images without obvious boundaries, the performance of CNN-based methods will be greatly degraded. The OC boundaries of the cases in Fig. 2 (a and b) are not obvious. The OC boundaries of Fig. 2 (a and b) need to be determined with the clinical prior knowledge (*i.e.*, kinks). Kinks not only have a decisive effect on the adjacent OC boundaries, but also have a constraining effect on the distant OC boundaries. CNN-based methods generally exhibit limitations for modeling explicit long-range relations between OC and the clinical prior knowledge due to the properties of the local operation. Therefore, it is necessary to effectively model the long-range dependency between the clinical prior knowledge and OC boundary to improve the performance of OC segmentation. To our best knowledge, there is no work incorporating a clinical prior knowledge (*i.e.*, vessel information) into deep learning for OC and OD segmentation. To alleviate the above mentioned limitations, in this paper, we propose a novel Prior Knowledge-based Relation Transformer Network (PKRT-Net), which models long-range relations between features guided by the clinical prior knowledge to segment OC and OD. We introduce a vessel space containing prior knowledge (*i.e.*, kinks), which provides feature information related to the OC and OD from the perspective of vessel features to assist OC and OD segmentation. To overcome the intrinsic locality limitation of convolution operations, we design a novel relation transformer to model the long-range dependencies between OC/OD features and prior knowledge. Specifically, our proposed PKRT-Net consists of three parts: dual-branch module (DBM), relation transformer fusion module (RTFM), and decoder with weighted fusion module (WFM). Different from existing methods, DBM incorporates clinical knowledge into our framework. In DBM, the fundus image is first sent to two independent branch networks for feature extraction from the two aspects of the clinical knowledge and local information. The clinical prior knowledge branch is restricted to extracting OC information from the perspective of vessel features. Then the local edge features and vessel features from the two branch networks are further fed into RTFM. RTFM facilitates the fusion of local features and vessel features based on expert knowledge, and efficiently models long-range relations in the spatial space. Finally, the weighted fusion module enhances the features from the decoder by attention blocks, and simultaneously fuses the features of RTFM to high-level layers to improve segmentation performance. We evaluate our PKRT-Net framework on three public available datasets (*i.e.*, Drishti-GS [3], RIM-ONE(r3) [2], and REFUGE [38]). Our framework outperforms the state-of-the-art approaches, bringing significant improvements by incorporating the clinical knowledge. Our contributions are summarized as follows:

1. To our best knowledge, our proposed method is the first attempt to incorporate the clinical vessel knowledge in deep learning methods to segment OC and OD. When the edge information of OC is unclear, our proposed dual-branch module based on the clinical prior knowledge can provide the blood vessel information related to OC to assist the OC segmentation.
2. We propose a relation transformer fusion module (RTFM), which is able to exploit not only the intra-branch relationship in each branch, but also the inter-branch relationship between a local edge feature branch and a prior knowledge branch. In addition, RTFM can effectively model the long-range dependencies of features to achieve accurate segmentation of OC and OD.
3. The weighted fusion module is designed to directly fuse features from the RTFM with the decoder's features to improve the final OC segmentation performance while enhancing the multi-scale output with the attention block.

The remainder of this article will be organized as follows. We review related works in Section 2. Our proposed PKRT-Net framework is introduced in Section 3. The experimental setup and experimental results are presented in Section 4. Conclusions are put in Section 5.

2. Related works

In this section, related works about OC and OD segmentation are discussed. OC and OD segmentation methods are classified into two categories including hand-crafted feature based methods and deep learning based methods. Furthermore, we briefly review clinicians' knowledge related to OC and OD and transformer mechanism based methods.

2.1. OC and OD segmentation

Many hand-crafted feature based methods were designed to segment OC and OD. Earlier, template based models were proposed to obtain OD boundaries. The segmentation problem was transformed into a problem of minimizing energies related to intensity, texture, and boundary smoothness. The active contour model was utilized to detect the contour based on image gradient [18]. OD was modeled as elliptical objects using the Hough transform [19,26,27]. Deformable-based models were used to segment OC and OD [18,14]. For example, OD was firstly located and then OC was segmented using a deformable contour model [18]. However, these methods are susceptible to image quality, contrast changes, and blood vessels. Recently, several methods have been proposed to transform the boundary detection problem into a pixel classification problem. For example, Cheng et al. [15] segmented OC and OD using a superpixel classifier, which exploits various hand-crafted features. All of the above methods are highly dependent on hand-crafted features.

Many methods based on deep learning have been proposed to segment OC and OD. Zilly et al. [29] extracted OC and OD by an ensemble learning method based on CNN framework, which obtained informative points by an entropy sampling technique. Sevastopolsky proposed a lightweight and efficient Modified U-Net [30] to segment OC and OD. In [31], a cascade network was designed to extract OC and OD based on the U-Net networks. Unlike the previous CNNs, M-Net [32] proposed a method to jointly segment the OC and OD based on the U-shaped network. Gu et al. [39] proposed a context encoder network (CE-Net), which can maintain rich spatial information while extracting high-level features. Xu et al. [40] proposed a multi-scale and multi-kernel U-shaped network that can adaptively adjust the sampling of the sample space. A spatial-aware joint segmentation method [33] was proposed by considering the multi-scale spatially dense features. A level set based deep learning method [34] was proposed for optic disc and cup segmentation. To enhance the robustness of the model on datasets from different sources, some adversarial learning network-based methods were proposed to segment OC and OD. TAU [41] proposed a transferable attention U-Net model for OC and OD segmentation tasks with two discriminators and attention modules. In [35], a patch-based output space adversarial learning framework was designed to segment OC and OD. A WGAN domain adaptation framework was proposed in [36] for segmenting OC and OD in fundus images. In [37], a domain adaptation framework was proposed to segment OC and OD on fundus images based on image synthesis and feature alignment method.

Compared to hand-crafted feature-based methods, deep learning methods are able to extract features for OC and OD segmentation automatically. Many CNN-based network variants were designed to segment OC and OD. However, these deep learning methods neglect to combine the clinical prior information to obtain more accurate OC boundaries. Prior knowledge is essential for clinicians to determine OC boundaries.In our method, we explore incorporating the clinical vessel knowledge in an independent feature extraction branch. Then a transformer-based fusion module is proposed to fuse the clinical prior knowledge to assist OC segmentation.

2.2. Vessel kinks

When there is no color change at the OC boundary in a 2D retinal image, the method based on a color change to extract the boundary will fail. Several hand-crafted methods used vessel kink points as a clinical prior knowledge to extract OC boundaries. In [13], blood kinks were exploited to locate OC boundaries. In [14], the same vessel bend concept was proposed for OC segmentation. By using the information of vessel kink, meaningful information of 3D is obtained in a 2D retinal image. clinical prior knowledge is currently only used in hand-crafted feature based methods for OC segmentation. Deep learning methods lack the clinical prior knowledge to accomplish the target task. In this paper, we combine the clinical prior knowledge with deep learning for automatic OC segmentation.

2.3. Transformer

The transformer was first proposed in the machine translation task [42], and it has produced many state-of-the-art methods in the field of natural language processing (NLP) [43,44]. Inspired by transformers, many works tried to extend transformers to the field of computer vision. In [45], the image size that the model can process increases by restricting the self-attention mechanism in the local information. Scalable approximations to global self-attention were employed in Sparse Transformers[46]. Recently, some efficient and effective transformer-based work has been proposed. For example, Vision Transformer (ViT) [47] designed the classification network based on the standard transformer module to achieve the state-of-the-art method on ImageNet classification. TransUNet [48] combined transformer and U-Net for medical image segmentation. Swin-UNet[49] designed a UNet-like pure transformer architecture based on shifted windows mechanism for medical image segmentation. Different from the existing transformer layers, we design a new relation transformer layer that can model the relationship of features from two different branches. The proposed relation transformer is able to exploit not only the intra-branch relationship in each branch, but also the inter-branch relationship between a local edge feature branch and a prior knowledge branch in medical image processing.

3. Methods

We propose a PKRT-Net framework to segment the OC and OD, as shown in Fig. 3. The main difference between the proposed approach and existing ones is that our framework utilizes the relation transformer to model long-range feature dependencies guided by the clinical prior information. The proposed PKRT-Net framework consists of three major parts: the dual-branch module, the relation transformer fusion module, and the weighted fusion-based decoder. We first locate the OD center by our previous method [9,10]. The region of the fundus image that contains the OC is cropped out automatically. Region of interest (ROI) in a fundus image $X \in \mathbb{R}^{H\times W\times 3}$ is fed into the dual-branch module (DBM), which is composed of two independent branches: general feature extraction branch and vessel feature guidance branch. DBM decouples fundus images into two different feature spaces. The two

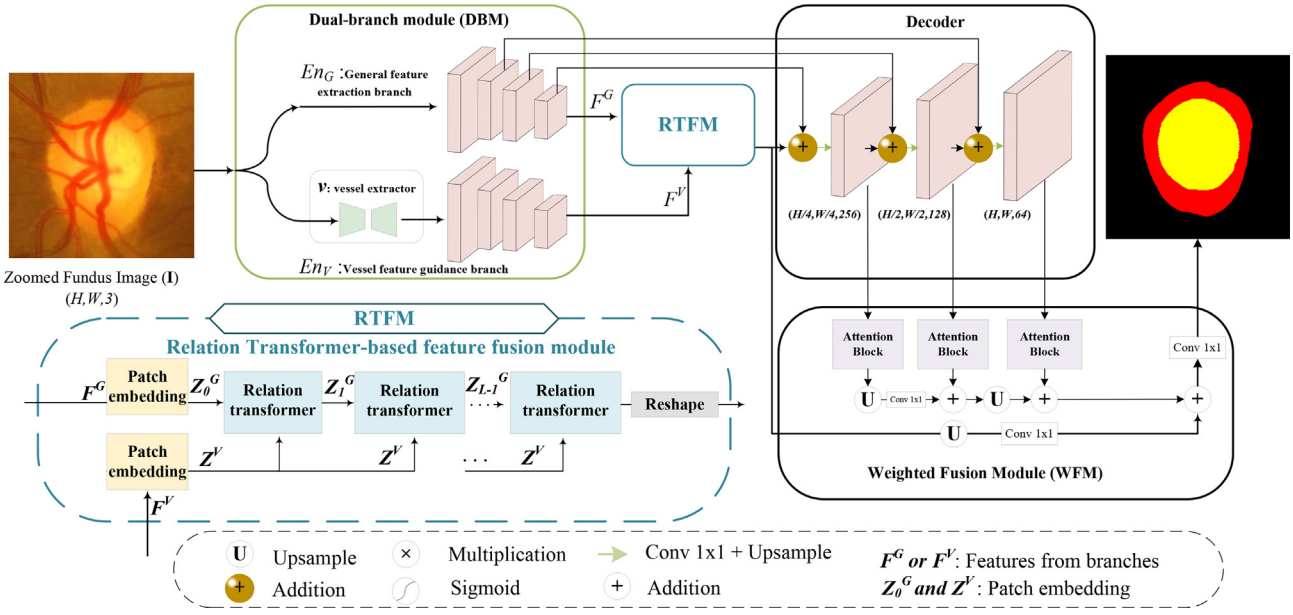

Fig. 3. Illustration of our PKRT-Net architecture, which consists of the dual-branch module (DBM), the relation transformer fusion module (RTFM), and the decoder with weighted fusion module (WFM). ROI of the fundus image is fed into the dual-branch module, which extracts two decoupled features from two different branch networks. Further, the relation transformer fusion module (RTFM) fuses two decoupled features from the dual branch network. Finally, the weighted fusion module (WFM) fuses RTFM's features and the decoder's features to obtain the final segmentation result.

decoupled features are further fed into the relation transformer feature fusion module simultaneously. RTFM can improve the accuracy of OC by fusing the clinical prior knowledge and modeling local and long-range relations. The output of the RTFM module is further passed through the decoder, which gradually restores the scale of the high-level features to the width and height of the fundus image. Finally, the weighted fusion module incorporated decoder's multi-slice output with the representations of the relation transformer to improve segmentation performance. The proposed approach can produce efficient results without using any intensity normalization stage, such as presented in [50,51], which increases computational costs. In this section, we introduce the proposed segmentation framework in detail.

3.1. Dual-branch module

Our target task is closely related to vessel information of the fundus image. So we consider creating an additional branch to extract features of the target from the perspective of the vessel space, which is different from existing methods. In this way, the fundus image is decoupled into a general feature space and vessel feature space, respectively. To achieve this, we propose the dual-branch module, as shown in the top left green box in Fig. 3. The dual-branch module consists of the general feature extraction branch En_G and the vessel feature guidance branch En_V. F^G and F^V represent the features extracted by En_G and En_V, respectively. I and $\mathcal{V}$ denote the fundus image and vessel extractor function, respectively. Eq. (a) indicates that the general feature extraction branch En_G extracts the decoupled feature F^G from the fundus image. Eq. (b) indicates that the vessel feature guidance branch En_V extracts decoupled features F^V from the fundus image under the constraints of the vessel feature space.

$$F^G = En_G(I), \quad \text{(a)}$$

$$F^V = En_V(I|\mathcal{V}(I)). \quad \text{(b)}$$

General feature extraction branch. It mainly relies on the edge information and local information of the feature. The general feature extraction branch uses the ResNet network as its backbone. The global average pooling layer and the final fully connected layer of ResNet are deleted, and the four stages of the ResNet backbone are retained. Each stage is composed of multiple residual blocks. Each residual block consists of two 3×3 convolutional layers and a residual connection. Each convolutional layer is followed by a batch normalization (BN) layer and rectified linear unit (ReLU) activation function. Activation functions should be chosen carefully in deep networks with residual blocks [52]. Although, various activation functions have been applied in recent works [53–57], ReLU has been used in the proposed architecture due to its efficiency.

Vessel feature guidance branch. It extracts features from the perspective of blood vessels to guide the segmentation of OC. The vessel feature guidance branch includes a blood vessel extractor and a vessel feature encoder network. First, the fundus image is used to extract the vessel information through the vessel extractor. The vessel extraction method is the extractor in [39], and its main structure is the U-shaped network. The blood vessel extractor first extracts the vessel structure, then the vessel is sent to the vessel feature encoder network. ResNet-34 is used as the backbone of the vessel feature encoder network, and its structure is similar to the general feature extraction branch network.

3.2. Relation transformer fusion module

In this section, we introduce how to model both general feature space and vessel feature space relationships in a unified model using our relation transformer. Due to the limitations of the intrinsic locality of convolution operations, we propose a relation transformer fusion module to fuse the features extracted from the two-branch network. Compared with the transformer module in [42,47], the relation transformer we proposed can not only promote the feature interaction within a single branch, but also increase the feature information interaction in both branch networks. The relation transformer fusion module mainly includes patch embedding, position embedding, and a relation transformer layer. Firstly the decoupled features from 2D images are split into patch sequences by patch embedding, and then the patch sequences are assigned position information. Finally, the relation

transformer layers fuse the patch sequences from the two decoupled spaces.

Patch embedding and position embedding. F^G and F^V are the features extracted from the dual-branch network as shown in Fig. 3. F^G represents general local features, and F^V represents clinical prior information features. We reshape the features $F^G \in \mathbb{R}^{h\times w\times C}$ and $F^V \in \mathbb{R}^{h\times w\times C}$ into two sequences of flattened 2D patches $\boldsymbol{X}_{p^g} \in \mathbb{R}^{N\times(C\cdot P\cdot P)}$ and $\boldsymbol{X}_{p^v} \in \mathbb{R}^{N\times(C\cdot P\cdot P)}$ with patches of size $P \times P$, as shown in Eq. (2). (h, w) and C represent the resolution and channel number of the features, respectively. $N = \frac{h\cdot w}{P\cdot P}$ is the number of patches. Then $\boldsymbol{X}_{p^g}$ and $\boldsymbol{X}_{p^v}$ are mapped to a D-dimensional space with a learnable linear projection (*i.e.*, $\boldsymbol{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$), where D represents the dimension of the latent space. Further a learnable position embedding (*i.e.*, $\boldsymbol{E}_{pos} \in \mathbb{R}^{N\times D}$) is added to the patch sequence to preserve the spatial position information of the sequence, as shown in following Eqs. (3) and (4). $\boldsymbol{Z}_0^G$ and $\boldsymbol{Z}^V$ represent the sequence feature embeddings from the general feature branch and the vessel feature branch, respectively.

$$\begin{aligned}\boldsymbol{X}_{p^g} &= \left[x_{p^g}^1, \ldots, x_{p^g}^i \ldots, x_{p^g}^N\right], i = 1, \ldots, N,\\ \boldsymbol{X}_{p^v} &= \left[x_{p^v}^1, \ldots, x_{p^v}^i \ldots, x_{p^v}^N\right], i = 1, \ldots, N,\end{aligned} \tag{2}$$

$$\begin{aligned}&\boldsymbol{Z}_0^G = \left[x_{p^g}^1\boldsymbol{E}; x_{p^g}^2\boldsymbol{E}; \cdots; x_{p^g}^N\boldsymbol{E}\right] + \boldsymbol{E}_{pos},\\ &\boldsymbol{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \boldsymbol{E}_{pos} \in \mathbb{R}^{N\times D},\end{aligned} \tag{3}$$

$$\begin{aligned}&\boldsymbol{Z}^V = \left[x_{p^v}^1\boldsymbol{E}; x_{p^v}^2\boldsymbol{E}; \cdots; x_{p^v}^N\boldsymbol{E}\right] + \boldsymbol{E}_{pos},\\ &\boldsymbol{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \boldsymbol{E}_{pos} \in \mathbb{R}^{N\times D}.\end{aligned} \tag{4}$$

Relation transformer layer. Compared with the transformer in [42,47], the proposed relation transformer can fuse the features of the two branch networks. Relation transformer layer consists of a multi-head cross-attention (MCA), Layer norm (LN), and multi-layer perceptron (MLP) blocks, as shown in Fig. 4. First, the correlation matrices between two different patch sequences are calculated in the cross-attention sub-layer, and then the output of the cross-attention is normalized, and the final result is obtained through the multi-layer perceptron.

In the multi-head cross-attention sub-layer, h groups parallel cross-attention heads are concatenated together. The h groups cross-attention are realized by mapping queries ($\boldsymbol{Q}$), keys ($\boldsymbol{K}$), and values ($\boldsymbol{V}$) h times through different learnable linear projection. Specifically, given two different sequences $\boldsymbol{Z}^G$ and $\boldsymbol{Z}^V$ (representing features from two decoupled branches), we compute the i-th cross-attention inputs query, key and value. $\boldsymbol{Z}^V$ is linearly projected to $\boldsymbol{Q}_i^V$ with $\boldsymbol{W}_i^Q$, as shown in Fig. 4; $\boldsymbol{Z}^G$ is linearly projected to $\boldsymbol{K}_i^G$ and $\boldsymbol{V}_i^G$ with $\boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V$, respectively. The query, key and value are formed with the following equations:

$$\begin{aligned}\boldsymbol{Q}_i^V &= \boldsymbol{Z}^V\boldsymbol{W}_i^Q,\\ \boldsymbol{K}_i^G &= \boldsymbol{Z}^G\boldsymbol{W}_i^K,\\ \boldsymbol{V}_i^G &= \boldsymbol{Z}^G\boldsymbol{W}_i^V,\end{aligned} \tag{5}$$

where $\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{model}\times d_k}, \boldsymbol{W}_i^K \in \mathbb{R}^{d_{model}\times d_k}, \boldsymbol{W}_i^V \in \mathbb{R}^{d_{model}\times d_v}$, and $d_k = d_v = d_{model}/h = 64$. Then we use the Scaled Dot-Product attention to calculate the correlation matrix, and this correlation matrix is used for a weighted combination of value. The cross-attention (CA) formula is described as follows:

$$CA\left(\boldsymbol{Q}_i^V, \boldsymbol{K}_i^G, \boldsymbol{V}_i^G\right) = SoftMaX\left(\frac{\boldsymbol{Q}_i^V\boldsymbol{K}_i^{G^T}}{\sqrt{d_k}}\right)\boldsymbol{V}_i^G. \tag{6}$$

Further all heads are calculated and they are concatenated using the following formula:

$$\begin{aligned}&MCA\left(\boldsymbol{Z}^V, \boldsymbol{Z}^G\right) = Concat(head_1, \ldots, head_h)\boldsymbol{W}^O,\\ &head_i = CA\left(\boldsymbol{Z}^V\boldsymbol{W}_i^Q, \boldsymbol{Z}^G\boldsymbol{W}_i^K, \boldsymbol{Z}^G\boldsymbol{W}_i^V\right),\end{aligned} \tag{7}$$

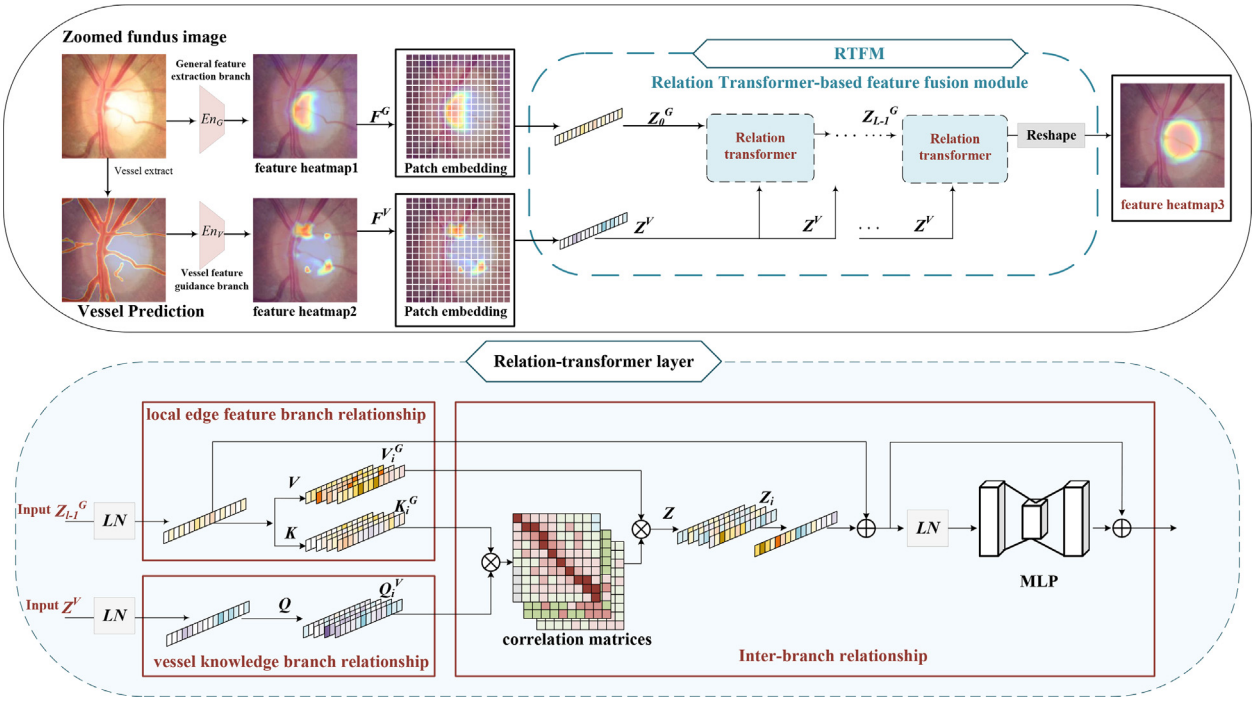

Fig. 4. Illustration of relation transformer layer. $\boldsymbol{Z}_{l-1}^G$ and $\boldsymbol{Z}^V$ represent output of (l-1)-th relation transformer and patch sequence features from vessel space, respectively. $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$ represent queries, keys, and values. LN and MLP denote the layer norm and the multi-layer perceptron, respectively. $\oplus$ and $\otimes$ denote the element-wise addition and the element-wise multiplication, respectively.

where $\boldsymbol{W_i^O} \in \mathbb{R}^{h \cdot d_v \times d_{model}}$.

The output of the MCA is further fed into an MLP with two fully connected layers to adjust the representation. Given the input of the MLP x, the MLP can be described as follows:

$$MLP(x) = ReLU(x\boldsymbol{W_1} + b_1)\boldsymbol{W_2} + b_2. \tag{8}$$

Our proposed relation transformer layer can be fully described in Eq. (9) and (10). $\boldsymbol{Z}^V$ denotes the patch embedding from vessel space, and $\boldsymbol{Z}^G_{l-1}$ denotes the output of (l-1)-th relation transformer layer.

$$\widehat{\boldsymbol{Z}^G_l} = MCA\left(LN\left(\boldsymbol{Z}^V\right), LN\left(\boldsymbol{Z}^G_{l-1}\right)\right) + \boldsymbol{Z}^G_{l-1}, \quad l = 1 \cdots L, \tag{9}$$

$$\boldsymbol{Z}^G_l = MLP\left(LN\left(\widehat{\boldsymbol{Z}^G_l}\right)\right) + \widehat{\boldsymbol{Z}^G_l}, l = 1 \cdots L. \tag{10}$$

Below we will briefly analyze the reasons why RTFM is effective on our target task. To visualize feature maps more concisely, we only train OC segmentation for feature heatmap visualization. The feature heatmap shown in Fig. 4 comes from OC segmentation prediction. In Fig. 4, feature heatmap1 represents edge features; feature heatmap2 represents vessel structure features; the distribution of feature heatmap2 is more discrete than that of feature heatmap1. Feature map heatmap3 represents the features generated by the visual transformer, and feature map heatmap3 shows that the visual transformer is able to establish long-range relationships between features in heatmap1 and features in heatmap2. It can be observed that the visual transformer can achieve performance improvement on our target task. Some recent works [58–61] have also demonstrated that visual transformers have strong cross-modal modeling capabilities in cross-modal tasks.

3.3. Decoder with weighted fusion module

The decoder branch network has three decoder blocks, three multi-slice output layers, and a weighted fusion module. Each decoder block is composed of two convolutional layers and a bilinearly interpolated up-sampling layer. Each convolutional layer is followed by a BatchNorm (BN) layer and a ReLU activation function. The weighted fusion module merges the features of RTFM and the multi-slice-layer output for the final prediction.

Weighted Fusion Module. Our proposed weighted fusion module integrates the multi-scale features in the decoder and the representation of RTFM to improve the performance of segmentation. As shown in Fig. 3, the multiple slices output of the decoder will go through an attention block to pay more attention to useful information.

In the attention block, the attention among the feature channels is redistributed, and the feature channel with high saliency will get a larger attention weight. Specifically, 1×1 convolution is first used to adjust the channel dimension of the input feature F. Input feature with new dimensions is denoted as F', as shown in Fig. 5. Then global max-pooling and global average-pooling are used to squeeze the spatial information of the input feature F', respectively. Two $1 \times 1 \times C$ channel descriptors F_{Avg} and F_{Max} are generated. F_{Avg} and F_{Max} represent the average-pooling and max-pooling features, respectively. The number of channels of the two pooling descriptors is the same as that of the feature F'. F_{Avg} and F_{Max} share two consecutive fully connected layer. The hidden layer size of the fully connected layer is set to C/r to reduce the parameter amount of the fully connected layer, where r represents the reduction rate. Finally, the feature vectors of the output of F_{Avg} and F_{Max} are added element by element, and the result of the addition is used to generate the attention weight through the sigmoid function. After that, the input feature and attention weights are computed by element-wise multiplication.

Objective Function.The final loss function is formulated as follows,

$$\begin{aligned} L &= -\sum_k^K \frac{2w_k\left(P_k \bigcap G_k\right)}{|P_k| + |G_k|} \\ &= -\sum_k^K \frac{2w_k \sum_i^N p_{(k,i)} g_{(k,i)}}{\sum_i^N p^2_{(k,i)} + \sum_i^N g^2_{(k,i)}}, \end{aligned} \tag{11}$$

where K represents the number of classes, and k represents the k-th class; P_k and G_k represent the prediction of the k-th class and the ground truth of the k-th class, respectively; $w_k = \frac{1}{K}$ denotes class weight; For a pixel-level perspective, N represents the number of pixels, and i represents the i-th pixel; $g_{(k,i)}$ in $\{0, 1\}$ represents whether the i-th pixel belongs to the k-th category; $p_{(k,i)}$ in $[0, 1]$ represents whether the i-th pixel predicts whether it belongs to class k.

4. Experiments

4.1. Datasets

In our experiment, we evaluated our method and state-of-the-art methods on three public datasets, namely Drishti-GS [3], RIM-ONE(r3) [2], and REFUGE [38]. These three datasets come from different fundus cameras, and their image quality varies greatly. In our experiment, we evaluated our method and state-of-the-art methods on three public datasets, namely Drishti-GS, RIM-ONE(r3), and REFUGE. These three datasets come from different fundus cameras, and their image quality varies greatly. The datasets have been divided into a training set, a validation set, and a test set and we follow the settings in all the experiments,

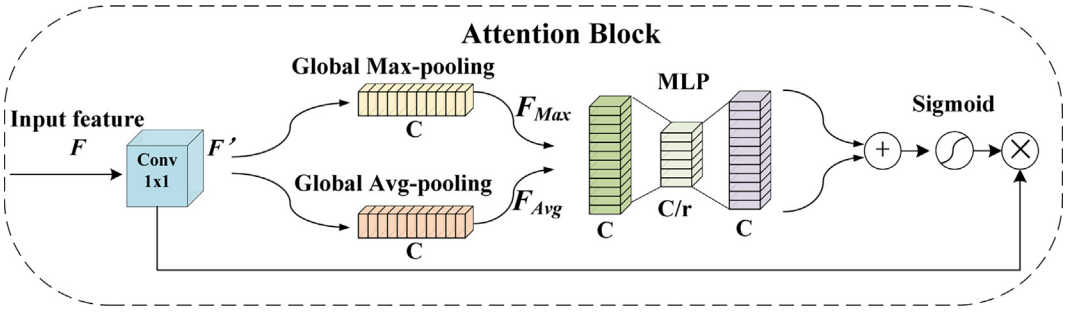

Fig. 5. Illustration of attention block. F, C, r and MLP denote the input feature, the number of channels, the reduction rate and the fully connected layer. $\oplus$ and $\otimes$ denote the element-wise addition and the element-wise multiplication, respectively.

which is consistent with previous methods. The datasets are further described as follows:

Drishti-GS dataset: It contains 101 fundus images, including 50 training sets and 51 test sets. The distribution of glaucoma and normal eyes in the training and test sets is slightly different. The training set includes 32 glaucoma and 18 normal eyes; the test set includes 38 glaucoma and 13 normal eyes. The fundus images are centered on OD, with a field of view of 30 degrees and a resolution of 2896 × 1944. For each fundus image, the annotations were marked by four glaucoma experts with 3, 5, 9, and 20 years of experience.

RIM-ONE(r3) dataset: It contains 159 stereo images, including 75 healthy cases and 84 glaucoma suspect cases. The fundus images were captured by the fundus camera Nidek AFC-210 with a resolution of 2144 × 1424. The partition of training and test sets of RIM-ONE(r3) is not given. We divide the dataset into the training set and test set (99/60) according to [30,28].

REFUGE dataset: It contains 1200 fundus images, which are equally divided into a training set, validation set, and test set. The ratio of glaucoma to normal eyes remains the same in the three sets. The fundus images of the training set were captured by Zeiss Visucam 500 fundus camera, and the validation/test set was captured by the Canon CR-2 device [38]. The resolution of the fundus images taken by the two fundus cameras is shown in Table 1. Seven experts and a senior specialist voted on the segmentation of OC and OD.

4.2. Evaluation metric

The comparison methods and the proposed method are evaluated using the classic metric Dice, which is consistent with the references [34–38]. Dice coefficient is defined as follows: $Dice = \frac{2TP}{TP+FN+TP+FP}$, where *TP* and *TN* represent the number of true positives and true negatives, respectively, and *FP* and *FN* represent the number of false positives and false negatives, respectively.

4.3. Implementation details

For a more fair comparison, we compare our method with other approaches in the same test environment, which is Pytorch 1.11 on top of AMD Ryzen Threadripper 3960X 24-Core Processor and NVIDIA GeForce RTX™3090 graphic card. Mmsegmentation [62] and segmentation models [63] are two open source libraries that can fairly compare model parameters and runtimes. A lot of the latest work code is now integrated into these two libraries. The maximum number of model epoch for training is 300. We use the Adam optimizer to update the model parameters, and the initial learning rate of the parameters is 3×10^{-4}. Classic data augmentation techniques are used, such as random rotation, flip, and movement. The cropped fundus image is compressed to 256×256 resolution. Transformer-based fusion module contains 3 relation transformer blocks. Patch size *P*, Hidden size *D* and heads *h* of MCA are set to 1, 768, and 12 respectively. For different linear projections of dimensions we use $d_k = d_v = d_{model}/h = 64$. MLP size is set to 3072.

4.4. Comparison with state-of-the-art methods

For a fair comparison, we verified the model's performance on three public data sets. We compared our method with several state-of-the-art OD and OC methods. Table 2 and 3 show the performance comparison of different deep learning methods on public datasets Drishti-GS, RIM-ONE(r3), and REFUGE. In addition, we also compared the latest transformer-based segmentation methods: TransUNet [48] and swin-UNet [49]. These state-of-the-art methods can be grouped as follows:

- **U-Net methods** are characterized by a U-shaped structure. Modified U-Net [30] and level-set U-Net [34] made some structural changes based on U-Net, such as adjusting the number of convolutional layers. In addition, a constraint loss was introduced into level-set U-Net.
- **Multi-scale U-Net methods** include M-Net [32], Stack-U-Net [31], CE-Net [39], MSMKU-Net [40], and NENet [64]. The main feature of this group methods is that multi-scale spatial features are introduced into the U-shaped network. Many ways were utilized to realize multi-scale spatial features, such as multi-scale input and output features, multi-scale high-level latent space features, etc.
- **GAN-based methods** contain pOSAL [35], WGAN-seg [36], and ISFA [37]. This group methods employed an adversarial learning strategy to construct a domain adaptation framework.
- **Transformer-based methods** contain TransUNet [48] and swin-UNet [49]. Transformer blocks were utilized in both methods for medical image segmentation.

4.4.1. ***Results on Drishti-GS and RIM-ONE(r3) dataset***

We first compared our PKRT-Net with the state-of-the-art approaches on Drishti-GS and RIM-ONE(r3) datasets. Table 2 shows a performance comparison among these methods. The GAN-based approaches did not show a segmentation improvement than U-Net and multi-scale U-Net on the Drishti-GS and RIM-ONE (r3), which are single-domain datasets. Single-domain dataset means that the data in the training and test sets of the Drishti-GS and RIM-ONE(r3) datasets are captured by the same kind of camera, respectively. This shows that the GAN-based method has no obvious performance improvement on the single-domain dataset. Multi-scale U-Net networks (*i.e.*, M-Net, stack-U-Net, CE-Net, MSMKU-Net, and NENet) achieve higher segmentation Dice than the U-Net network employing multiscale spatial information modules/blocks. Due to the properties of local convolution operation, CNNs lack the ability to model long-range dependency. The transformer-based approaches (*i.e.*, TransUNet and swin-UNet) achieve better performance than the multi-scale networks. Compared with transformer-based methods, our PKRT-Net implements the vessel information constraints to enforce the model to exploit

Table 1
Description of fundus image datasets.

Dataset	Release year	Train/val/test	Number of samples	Image size	Cameras
REFUGE	2018	train	400	2124 × 2056	Zeiss Visucam 500
		val	400	1634 × 1634	Canon CR-2
		test	400	1634 × 1634	Canon CR-2
RIM-ONE(r3)	2011	train	99	2144 × 1424	Nidek AFC-210
		test	60	2144 × 1424	Nidek AFC-210
Drishti-GS	2015	train	50	2896 × 1944	unknown
		test	51	2896 × 1944	unknown

Table 2
Performance comparison with different deep learning methods on datasets Drishti-GS and RIM-ONE(r3).

Method	Drishti-GS		RIM-ONE(r3)		Time(ms)	FPS	Params
	$Dice_{Cup}$	$Dice_{Disc}$	$Dice_{Cup}$	$Dice_{Disc}$			
Modified U-Net [30]	0.8500	–	0.8200	0.9500	**1.24**	**807.95**	**1 M**
Level-set U-Net [34]	0.8706	0.9623	–	–	5.45	183.21	31 M
POSAL [35]	0.8580	0.9650	0.7870	0.8650	4.95	202.03	4 M
WGAN-seg [36]	0.8400	0.9540	–	–	5.68	176.18	23 M
ISFA [37]	0.8920	0.9660	0.8220	0.9080	19.14	52.24	7 M
M-Net [32]	0.8860	0.9658	–	–	4.45	224.63	9 M
Stack-U-Net [31]	0.8900	0.9700	0.8400	0.9500	12.23	81.78	15 M
CE-Net [39]	0.8818	0.9642	0.8435	0.9527	6.82	146.47	39 M
MSMKU-Net [40]	–	–	0.8564	0.9561	5.49	181.90	12 M
NENet [64]	0.8401	0.9632	0.8680	0.9552	18.21	54.89	30 M
TransUNet [48]	0.8875	0.9575	0.8423	0.9517	14.95	66.88	105 M
Swin-UNet [49]	0.8907	0.9687	0.8574	0.9536	7.57	132.08	21 M
Our PKRT-Net	**0.9120**	**0.9766**	**0.8723**	**0.9582**	11.14	89.75	56 M

- stands for value not available

Table 3
Performance comparison with different deep learning methods on dataset REFUGE.

Team/Method	Year	$Dice_{Cup}$	$Dice_{Disc}$	$Time_{ms}$
Winter_Fell		0.6861	0.8772	4.10
Cvblab		0.7728	0.9077	10.30
SDSAIRC		0.8315	0.9436	4.45
NIGHTOwl		0.8257	0.9487	–
SMILEDeepDR		0.8367	0.9386	–
Mammoth		0.8667	0.9361	11.05
AIML	REFUGE	0.8519	0.9505	43.19
VRT	Challenge	0.8600	0.9532	7.31
NKSG	2018 [38]	0.8643	0.9488	10.68
BUCT		0.8728	0.9525	**1.24**
Masker		0.8837	0.9464	21.33
M-Net [32]	2018	0.8648	0.9359	4.45
POSAL [35]	2019	0.8826	0.9602	4.95
NENet [64]	2021	0.8990	0.9616	18.21
Our PKRT-Net		**0.8997**	**0.9751**	11.14

useful clinical vessel knowledge for OC segmentation, resulting in a remarkable improvement of Dice of 2.1% and 1.5% on the Drishti-GS and RIM-ONE(r3), respectively. Our proposed PKRT-Net outperforms other state-of-the-art methods. Specifically, PKRT-Net achieves Dice of 0.9120 and 0.9766 for OC and OD, respectively, on the Drishti-GS dataset; it also achieves Dice of 0.8723 and 0.9582 for OC and OD, respectively, on the RIM-ONE(r3) dataset. The experimental results demonstrate that transformer and prior knowledge-based modules are useful to guide segmentation training.

We compare computing time and parameters in Table 2. Runtime, FPS (Frames Per Second), and number of parameters are used as three evaluation metrics. The runtime represents the time it takes for the model to complete forward propagation of an image. FPS is the frequency (rate) at which images are processed. The above evaluation indicators have also been used in recent research [65]. It can be seen from Table 2 that the model with the shortest runtime was Modified U-Net. Models with multi-scale modules (*e.g.*, Stack-U-Net, M-Net, and NENet) have a larger time consumption than Modified U-Net method. The time consumption of transformer-based methods (*i.e.*, TransUNet, Swin-UNet, and our PKRT-Net) is comparable to that of multi-scale methods, because there are improvements to the original transformer method in these three transformer-based methods. Both TransUNet and our PKRT-Net consist of CNN modules and transformer modules. The advantage of this combination is that only the high-level features extracted by CNN will be used in the transformer for global feature modeling, which can save a lot of computational consumption. In addition, CNN modules mainly contain standard 3×3 convolutions and transformer modules are also based on standard transformers, which can be well accelerated by the Pytorch Deep Learning framework. Swin-UNet is an improved work based on the swin-transformer. The advantage of swin-transformer is based on local window calculation, so the problem of a large amount of calculation in the transformer is greatly solved. Therefore, the computational efficiency of the transformer-based method is comparable to that of the multi-scale methods in Table 2. For the comparison of number of parameters, the transformer method is generally larger than that of the multi-scale CNN methods (*e.g.*, M-Net and CE-Net). Because the amount of model parameters does not determine the inference time of the model, there are examples where the CNN method takes longer than the transformer method. As shown in Table 2, NENet is more time-consuming than the transformer based methods. The main reason is that its backbone network EfficientNetB4 has a dense branch network, and its low parallel efficiency leads to more time-consuming inference speed.

4.4.2. ***Results on REFUGE dataset***

We also compared our method with the top-performing methods from the REFUGE challenge [38] and the latest method based on the REFUGE dataset, as shown in Table 3. Top-performing methods and the REFUGE dataset were released in Retinal Fundus Glaucoma Challenge held by MACCAI (Medical Image Computing and Computer Assisted Intervention Society). Different from Drishti-

GS and RIM-ONE(r3) dataset, REFUGE is a cross-domain dataset. The training and validation/test data of REFUGE dataset are from two different types of fundus cameras. Our method outperforms the state-of-the-art methods and achieves Dice of 0.8997 and 0.9751 for OC and OD, respectively.

We compare different models with runtime in Table 3. The runtime represents the time it takes for the model to complete the forward propagation of an image. As can be seen from Table 3, the model with the shortest run time was BUCT. The models that took the longest were AIML and Masker, as these two models contain multiple sub-segmentation networks for model voting.

4.5. Ablation study

We conducted an ablation study based on a U-Net baseline to demonstrate the effectiveness of the proposed modules: dual-branch module, relation transformer fusion module, and weighted fusion module. Table 4 shows the performance of our proposed different modules on three public datasets.

4.5.1. Ablation study for dual-branch module (DBM)

The DBM module contains a general feature extraction branch and a vessel feature guidance branch. The feature output of the vessel feature guidance branch depends on the vessel space, which is correlated with OC features. To verify the effectiveness of The DBM module, we use the U-Net network with ResNet encoder as our baseline. U-Net with DBM means that the baseline uses the DBM module as the encoder; similar to the skip connection of U-Net, the multiple layer outputs in the general feature extraction branch are summed with the corresponding layer of the decoder; the highest level feature output of the vessel feature guidance branch is summed with the corresponding general feature extraction branch for feature fusion; the decoder part is consistent with U-Net. As Table 4 shows, U-Net with DBM outperforms U-Net by average 1% on OC segmentation. To further demonstrate the performance of our proposed DBM module, We compare the performance of our PKRT-Net without DBM. The performance of PKRT-Net without DBM drops by about 2% on OC segmentation. The results demonstrate that the DBM module can capture effective features from prior knowledge (*i.e.*, vessel space) and extract a more robust feature representation to improve the performance of OC segmentation.

4.5.2. Ablation study for relation transformer fusion module (RTFM)

The RTFM module performs correlation calculations on long-range dependent features from two branches using multiple transformer blocks. To verify the effectiveness of RTFM, we compared the performance of U-Net with and without RTFM. U-Net with RTFM represents that the highest-level features of U-Net are replicated as two separate features for the dual inputs of RTFM, respectively; the output of RTFM is fed into the decoder of U-Net. It can be seen from Table 4 that U-Net with RTFM achieves about 3% and 1% improvement on OC and OD respectively on three public datasets. The results demonstrate that the RTFM module can effectively mine the long dependencies of features and facilitate the segmentation of OC and OD.

4.5.3. Analysis for relation transformer fusion module (RTFM)

As shown in Fig. 4, we will analyze the reasons why the inter-branch relationship and the intra-branch relationship in the RTFM can improve the final performance at the medical aspect and the engineering aspect.

Medical aspect: The more accurate OC boundaries in the fundus image are determined by a combination of the clear boundary information and the clinical prior knowledge. Especially for fundus images with unclear OC boundary, clinicians extract OC boundary in three main steps: first use clear boundary information to determine part of OC boundary, and then use effective vessel kinks to fit the remaining boundary information. The final OC boundary is obtained by combining the information from both perspectives.

Engineering aspect: To mimic the three steps in the clinical process described above, our proposed RTFM incorporates three relationships: the edge feature branch (intra-branch) relationship, the prior knowledge branch (intra-branch) relationship, and the inter-branch relationship. (1) The local edge relationship is used to imitate clinicians to extract relationships between edges. Specifically, the h-group projection transforms were used to generate the new h-group local edge-based features (*i.e.*, V_i^G and K_i^G). The h-group projection transforms method was proposed in [42] and it is widely used in the field of vision. From an engineering point of view, the h-group projection transforms increase the diversity of features, which has a positive effect on the final performance improvement. (2) The prior knowledge relationship was used to model the relationship of the vessel kinks clinically. Specifically, we have used the h-group projection transforms to generate the new h-group prior knowledge-based features (*i.e.*, Q_i^V), which is important for enhancing representations between prior knowledge. (3) The inter-branch relationship is used to imitate the way clinicians combine boundary information and vessel information to determine OC boundary. We design cross-attention to model the long dependencies between the two branches to generate a relation matrix; then the relation matrix is used to adjust the attention weights; finally, a multilayer perceptron (MLP) is used to enhance the joint of the features from two branches. Similar to the clinical experience, the combination of edge information and vessel information outperforms either of them independently.

4.5.4. Ablation study for weighted fusion module (WFM)

The WFM module mainly fuses features from the RTFM and the multi-slice output of the decoder. To verify the effectiveness of the WFM module, we append WFM to U-Net + RTFM (U-Net with RTFM) to observe the performance changes of the model. Compared with U-Net + RTFM, U-Net + RTFM with WFM achieves an average improvement of 0.7% in the segmentation of OC and OD on three public datasets, which illustrates the effectiveness of the WFM.

Table 4
Ablation tests on REFUGE, Drishti-GS and RIM-ONE(r3).

Baseline	DBM	RTFM	WFM	REFUGE		Drishti-GS		RIM-ONE(r3)	
				$Dice_{Cup}$	$Dice_{Disc}$	$Dice_{Cup}$	$Dice_{Disc}$	$Dice_{Cup}$	$Dice_{Disc}$
U-Net(R34)	–	–	–	0.8369	0.9492	0.8584	0.9565	0.8125	0.9392
	√	–	–	0.8478	0.9491	0.8676	0.9559	0.8245	0.9401
	–	√	–	0.8637	0.9584	0.8893	0.9654	0.8473	0.9483
	–	√	√	0.8712	0.9693	0.8946	0.9695	0.8564	0.9546
	√	√	–	0.8809	0.9587	0.8986	0.9658	0.8585	0.9493
	√	√	√	**0.8997**	**0.9751**	**0.9120**	**0.9766**	**0.8723**	**0.9582**

4.5.5. Ablation study for different non-local modules

Theoretical analysis: Attention as a non-local module has been a hot research topic. Many improvements to the attention mechanism based on QKV have been proposed, among which the most influential work is the transformer. The transformer structure was first proposed in the field of natural language processing (NLP). Further, the transformer has shown to be effective in computer vision as well. Compared with the attention structure, the classic transformer structure includes 12 parallel attention heads to enhance the diversity of features [42]. In addition, the transformer also includes MLP (multiple fully connected layers). MLP is an important part of improving transformer performance [66]. In our research, the attention in the transformer is modified to cross-attention in order to have the ability to model across branches. Therefore, the RTFM module proposed in this paper inherits the original advantages of the transformer and also has the ability to model across branches.

Experiment analysis: To compare the performance of different local blocks, we remove the RTFM module in our proposed framework as the baseline, and then use the following three non-local modules (*i.e.*, CSAM [67], BAM [68], and PAM [68]) for performance comparison.

- CSAM: The channel & spatial attention module (CSAM) [67] is proposed for vessel segmentation in fundus images, and the CSAM module can be inserted in the middle of any encoder and decoder network. CSAM contains spatial and channel attention blocks. In this experiment, the input to CSAM is the concatenation of two encoder network features.
- BAM: The basic spatial attention module (BAM) [68] is proposed to model the relationship of two branch networks, whose input is the output of the two branch networks. It can be integrated into the baseline framework with little modification.
- PAM: The pyramid spatial attention module (PAM) [68] introduces a pyramid structure based on the BAM module. Its input is the output features of the two branch networks, and it can also be combined with the baseline framework with little modification.

As can be seen from Table 5, CSAM, BAM and PAM can improve the performance compared with baseline. Compared with non-local blocks, our proposed RTFM module can achieve better performance.

4.6. Visualization

Fig. 6 shows visual examples of OC and OD segmentation from the Drishti-GS, REFUGE, and RIM-ONE(r3) datasets. The closed curves in green and blue represent the boundaries of OC and OD, respectively. The fundus images in the first three rows (*i.e.*, rows A, B, and C) of Fig. 6 are samples from Drishti-GS dataset. The fundus images in the fourth and fifth rows (*i.e.*, rows D and E) of Fig. 6 are samples from the REFUGE dataset. The last row (*i.e.*, row F) of Fig. 6 is the fundus image from RIM-ONE(r3) dataset. It can be seen from Fig. 6 that our PKRT-Net can achieve better OC and OD segmentation performance than U-Net on three public available datasets. Our proposed feature extraction module (*i.e.*, DBM) and feature fusion modules (*i.e.*, RTFM and WFM) have significant improvements for OC segmentation.

Table 5
Comparison of different non-local modules on REFUGE.

Method	non-local modules	REFUGE	
		$Dice_{Cup}$	$Dice_{Disc}$
Baseline	–	0.8725	0.9659
	+CSAM	0.8773	0.9698
	+BAM	0.8831	0.9655
	+PAM	0.8876	0.9693
	+RTFM	**0.8997**	**0.9751**

The fundus images in the third and last two rows (*i.e.*, C, E, and F) of Fig. 6 are examples of normal eyes (**N**), and the fundus images in the rest rows (*i.e.*, rows A, B and D) are examples from glaucoma patients (**G**). It can be observed from the ground truth in the second column that OC occupies a larger proportion of OD in glaucoma examples compared with the normal examples.

For better illustration, the fundus image samples in Fig. 6 are divided into two groups: **cases with clear boundaries (*i.e.*, cases C and E) and cases without obvious boundaries (*i.e.*, cases A, B, D and F)**. For cases without obvious boundaries, there is a large gap between the OC predicted by the U-Net and the Ground Truth; this shows that it is difficult for the U-Net to effectively predict the OC boundary when the boundary is unclear; U-Net with DBM achieves better performance than U-Net. As shown in Fig. 6 (A and B), U-Net with DBM finds kinks to determine the OC boundaries. However, U-Net with DBM does not have enough feature fusion capability, so the OC contour predicted by U-Net with DBM is not smooth enough. The OC boundaries predicted by U-Net with fusion modules (RTFM and WFM) are smoother. Our proposed PKRT-Net obtains more accurate OC and OD boundaries by effectively combining prior knowledge and feature fusion modules. For cases with clear boundaries (*i.e.*, cases C and E), both U-Net and our proposed PKRT-Net can accurately predict the boundaries of OC and OD. This shows that the CNN network can predict the accurate boundaries of OC and OD when the boundary is obvious.

To clearly demonstrate the effect of vessel segmentation on OC segmentation, we used the proposed framework for training OC segmentation only. Fig. 7 shows features generated from our proposed framework for a fundus image from REFUGE dataset, and the feature maps at different stages are used to demonstrate the role played. The features are displayed as heatmaps on the fundus image to visualize the model's attention to different positions of the image. Fig. 7 shows that the location of the vessel guidance branch focus is similar to that of the clinician focus, which provides the model with a clinical prior knowledge. The vessel feature guidance branch is beneficial for the model to obtain results closer to the ground truth. How vessel kink works is described in detail below.

Sub-figure (a) in Fig. 7 represents the zoomed fundus image and the predicted vessel structure. To better visualize the vessel segmentation, we show the predicted vessels as a heatmap on the fundus image. The black circles in the vessel feature prediction represent areas of concern to clinicians. Sub-figure (b) shows that the focus of the vessel guidance branch is similar to clinicians' focus in sub-figure (a). Sub-figure (c) shows that the features of the general feature extraction branch are more dependent on locations with clear boundary information. This dependency leads to a large gap between the region of feature focus and the Ground Truth. In this case, the vessel branch will be able to play a more positive role. The features of the vessel branch were fused with the features of the general feature extraction by RTFM, which made the fused features more similar to Ground Truth. Because RTFM contains multiple sublayers, we visualized features in the first sublayer and the final sublayer of RTFM to demonstrate changes in model attention. From sub-figure (d), it can be observed that the position of the vessel branch focus gradually guides the model to obtain results close to the Ground Truth.

4.7. Discussion

Automated OC and OD segmentation methods are very important to achieve large-scale glaucoma screening. Although existing

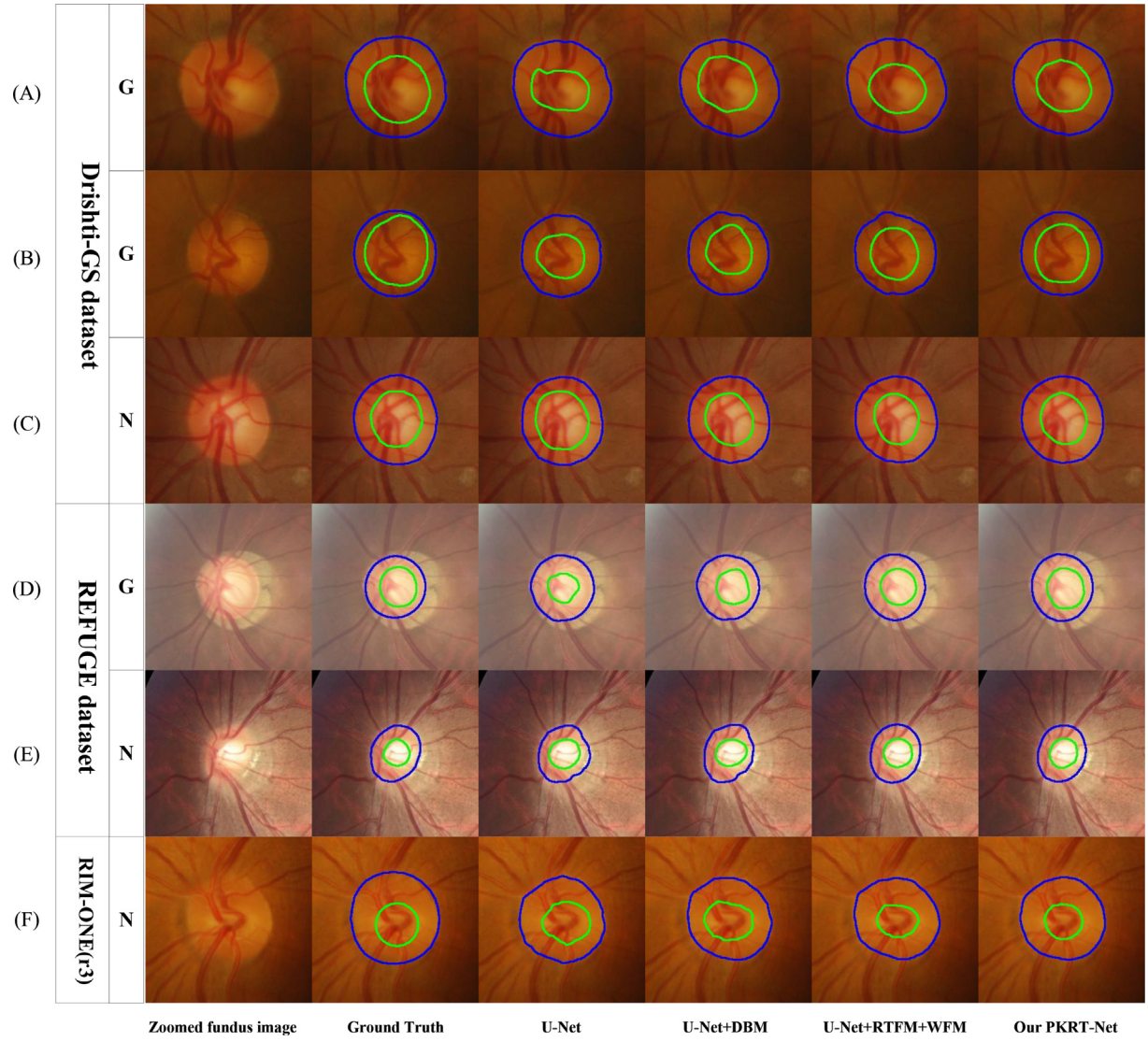

Fig. 6. Illustration of experimental results on the REFUGE, RIM-ONE(r3), and Drishti-GS datasets. The closed curves in green and blue represent OC and OD boundary, respectively. The first column: zoomed fundus image; the second column: Ground Truth of OC and OD boundary; the third column: the results of U-Net model (Baseline); the fourth column: U-Net with Dual-branch Module(DBM); the fifth column: U-Net with relation transformer fusion module (RTFM) and Weighted Fusion Module (WFM); the last column: our proposed PKRT-Net. N and G represent examples from normal eyes and glaucoma patients, respectively.

deep learning methods have achieved impressive performance, the accuracy of OC segmentation is limited due to the fact that deep learning methods focus more on boundaries and ignore clinical prior knowledge. In the present study, we introduce clinical prior knowledge into deep network to improve the performance of OC segmentation, in which a novel hybrid CNN and transformer are used to model global features. Our results demonstrate that clinical prior knowledge can help deep networks improve the performance of OC segmentation. Furthermore, we found that using CNN to extract shallow features and using transformers to fuse deep features is a good solution for balancing computational cost and modeling global feature relationships. To our knowledge, our proposed method is the first to combine clinical prior knowledge and deep learning methods for OC and OD segmentation. Our observation of the effect of clinical prior knowledge on deep networks is consistent with the recent report that clinical prior knowledge contributes to the segmentation of diabetic retinopathy multi-lesion [69]. It also implies that the clinical prior knowledge can be introduced to other medical images, *e.g.*, breast ultrasound image [70,71] and spine image [72].

Strengths and Weaknesses. We propose a relation transformer module to model the relationship between clinical prior knowledge and features automatically extracted by deep networks, enabling clinical prior knowledge to guide OC segmentation. The proposed transformer is only used to model high-level features, so the computational cost and the expensive memory consumption in the visual transformer framework are limited. While modeling the vessel-based prior information, some misleading vessel structures are introduced. As shown in Fig. 8, some vessel structure in the fundus is very similar to the real vessel kink, and such a vessel structure is called invalid kink. These invalid vessel structures are also challenging for clinicians, and only experienced clinicians can exclude those invalid vessel structures based on the vessel morphology analysis. Our model has limited ability to exclude indistinguishable invalid kinks. As shown in Fig. 8, our model deviates from the Ground Truth due to invalid kinks in a small number of

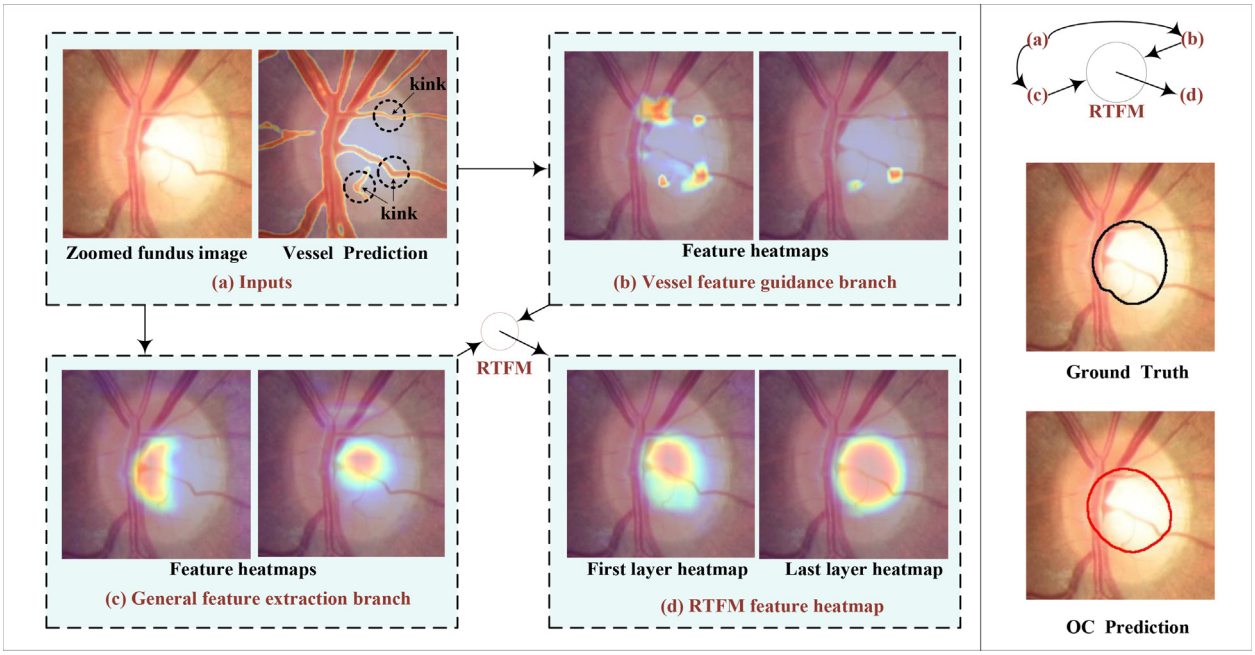

Fig. 7. Schematic diagram of vessel kink guided OC segmentation.

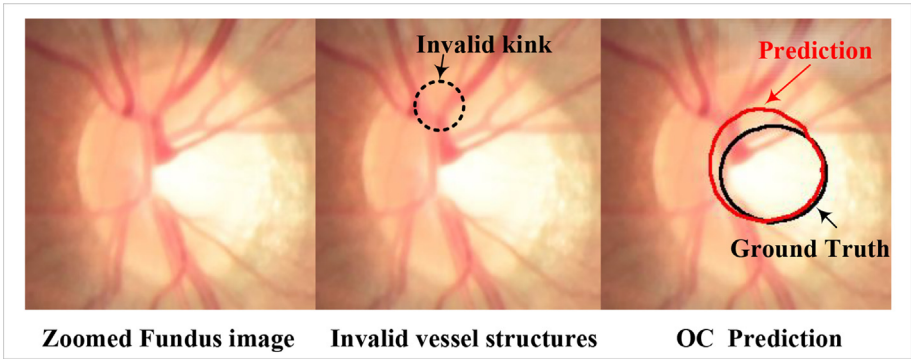

Fig. 8. Illustration of the invalid kink.

samples. In future work, we plan to improve our model to avoid being affected by invalid vessel structures.

5. Conclusion

In this paper, we presented a novel Prior Knowledge-based Relation Transformer Network (PKRT-Net) for OC and OD segmentation. The proposed PKRT-Net employed a dual-branch module to extract features from two aspects of clinical prior knowledge and local edge information. Relation transformer fusion module can exploit not only the intra-branch relationship in each branch, but also the inter-branch relationship between a local edge feature branch and a prior knowledge branch. Moreover, a weighted fusion-based decoder has been employed to assign weights to effective features of high-level layers, incorporated with the representations of the transformer, to supervise the final result. Experiments show the superiority of our proposed PKRT-Net compared with other state-of-the-art methods. The experimental results indicate that the clinical prior knowledge is essential for OC extraction. We will extend the proposed method to introduce prior knowledge in other medical image tasks in the future.

CRediT authorship contribution statement

Shuai Lu: Investigation, Methodology, Writing - original draft. **He Zhao:** Validation, Writing - review & editing. **Hanruo Liu:** Resources, Data curation. **Huiqi Li:** Supervision, Writing - review & editing, Funding acquisition. **Ningli Wang:** Resources, Conceptualization.

Data availability

Data will be made available on request.

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 research work is supported by the National Natural Science Foundation of China (NSFC) (Grant No. 82072007) and China Postdoctoral Science Foundation (No.2020M680387)

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.

[2] F. Fumero, S. Alayón, J.L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, Rim-one: An open retinal image database for optic nerve evaluation, 24th international symposium on computer-based medical systems (CBMS), IEEE 2011 (2011) 1–6.

[3] J. Sivaswamy, S. Krishnadas, G.D. Joshi, M. Jain, A.U.S. Tabish, Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation, IEEE 11th international symposium on biomedical imaging (ISBI), IEEE 2014 (2014) 53–56.

[4] P.N. Schacknow, J.R. Samples, The glaucoma book: a practical, evidence-based approach to patient care, Springer Science & Business Media, 2010.

[5] K. Lee, M. Niemeijer, M.K. Garvin, Y.H. Kwon, M. Sonka, M.D. Abramoff, Segmentation of the optic disc in 3-d oct scans of the optic nerve head, IEEE Trans. Med. Imaging 29 (2009) 159–168.

[6] M. Wu, T. Leng, L. de Sisternes, D.L. Rubin, Q. Chen, Automated segmentation of optic disc in sd-oct images and cup-to-disc ratios quantification by patch searching-based neural canal opening detection, Opt. Express 23 (2015) 31216–31229.

[7] H. Fu, D. Xu, S. Lin, D.W.K. Wong, J. Liu, Automatic optic disc detection in oct slices via low-rank reconstruction, IEEE Trans. Biomed. Eng. 62 (2014) 1151–1158.

[8] H. Fu, Y. Xu, S. Lin, X. Zhang, D.W.K. Wong, J. Liu, A.F. Frangi, M. Baskaran, T. Aung, Segmentation and quantification for angle-closure glaucoma assessment in anterior segment oct, IEEE Trans. Med. Imaging 36 (2017) 1930–1938.

[9] H. Li, O. Chutatape, Automatic location of optic disk in retinal images, Proceedings 2001 International Conference on Image Processing (Cat No. 01CH37205), 2, IEEE, 2001, pp. 837–840.

[10] H. Li, O. Chutatape, Automated feature extraction in color retinal images by a model based approach, IEEE Trans. Biomed. Eng. 51 (2004) 246–254.

[11] W.W.K. Damon, J. Liu, T.N. Meng, Y. Fengshou, W.T. Yin, Automatic detection of the optic cup using vessel kinking in digital retinal fundus images, in: 2012 9th IEEE international symposium on biomedical imaging (ISBI), IEEE, 2012, pp. 1647–1650

[12] B. Schwartz, Cupping and pallor of the optic disc, Arch. Ophthalmol. 89 (1973) 272–277.

[13] D. Wong, J. Liu, J. Lim, H. Li, T. Wong, Automated detection of kinks from blood vessels for optic cup segmentation in retinal images, in: Medical Imaging 2009: Computer-Aided Diagnosis, volume 7260, SPIE, 2009, pp. 459–466.

[14] G.D. Joshi, J. Sivaswamy, S. Krishnadas, Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment, IEEE Trans. Med. Imaging 30 (2011) 1192–1205.

[15] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, T.Y. Wong, Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, IEEE Trans. Med. Imaging 32 (2013) 1019–1032.

[16] Y. Zheng, D. Stambolian, J. O'Brien, J.C. Gee, Optic disc and cup segmentation from color fundus photograph using graph cut with priors, in: International conference on medical image computing and computer-assisted intervention, Springer, 2013, pp. 75–82.

[17] A. Almazroa, R. Burman, K. Raahemifar, V. Lakshminarayanan, Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey, J. Ophthalmol. 2015 (2015).

[18] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, L. Kennedy, Optic nerve head segmentation, IEEE Trans. Med. Imaging 23 (2004) 256–264.

[19] A. Aquino, M.E. Gegúndez-Arias, D. Marín, Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques, IEEE Trans. Med. Imaging 29 (2010) 1860–1869.

[20] S. Lu, Accurate and efficient optic disc detection and segmentation by a circular transformation, IEEE Trans. Med. Imaging 30 (2011) 2126–2133.

[21] M.D. Abramoff, W.L. Alward, E.C. Greenlee, L. Shuba, C.Y. Kim, J.H. Fingert, Y.H. Kwon, Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features, Investig. Ophthalmol. Visual Sci. 48 (2007) 1665–1673.

[22] H. Li, O. Chutatape, Boundary detection of optic disk by a modified asm method, Pattern Recogn. 36 (2003) 2093–2104.

[23] H. Li, O. Chutatape, Automatic detection and boundary estimation of the optic disk in retinal images using a model-based approach, J. Electron. Imaging 12 (2003) 97–105.

[24] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, M. Pietikäinen, Capturing correlations of local features for image representation, Neurocomputing 184 (2016) 99–106.

[25] A. Li, Z. Niu, J. Cheng, F. Yin, D.W.K. Wong, S. Yan, J. Liu, Learning supervised descent directions for optic disc segmentation, Neurocomputing 275 (2018) 350–357.

[26] X. Zhu, R.M. Rangayyan, Detection of the optic disc in images of the retina using the hough transform, in: 2008 30th annual international conference of the IEEE engineering in medicine and biology society, IEEE, 2008, pp. 3546–3549.

[27] A. Almazroa, S. Alodhayb, K. Raahemifar, V. Lakshminarayanan, Optic cup segmentation: type-ii fuzzy thresholding approach and blood vessel extraction, Clinical ophthalmology (Auckland, NZ) 11 (2017) 841.

[28] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, L.V. Gool, Deep retinal image understanding, in: International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 140–148.

[29] 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 Graph. 55 (2017) 28–41.

[30] A. Sevastopolsky, Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network, Pattern Recogn. Image Anal. 27 (2017) 618–624.

[31] A. SEVASTOPOLSKY, S. DRAPAK, K. KISELEV, B.M. SNYDER, J.D. KEENAN, A. GEORGIEVSKAYA, Stack-u-net: Refinement network for image segmentation on the example of optic disc and cup. arxivpreprintarxiv: 1804.11294, 2018.

[32] H. Fu, J. Cheng, Y. Xu, D.W.K. Wong, J. Liu, X. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, IEEE Trans. Med. Imaging 37 (2018) 1597–1605.

[33] Q. Liu, X. Hong, S. Li, Z. Chen, G. Zhao, B. Zou, A spatial-aware joint optic disc and cup segmentation method, Neurocomputing 359 (2019) 285–297.

[34] P. Yin, Y. Xu, J. Zhu, J. Liu, H. Huang, Q. Wu, et al., Deep level set learning for optic disc and cup segmentation, Neurocomputing 464 (2021) 330–341.

[35] S. Wang, L. 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.

[36] S. Kadambi, Z. Wang, E. Xing, Wgan domain adaptation for the joint optic disc-and-cup segmentation in fundus images, Int. J. Comput. Assist. Radiol. Surg. 15 (2020) 1205–1213.

[37] H. Lei, W. Liu, H. Xie, B. Zhao, G. Yue, B. Lei, Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation, IEEE J. Biomed. Health Inform. (2021).

[38] J.I. Orlando, H. Fu, J.B. Breda, K. van Keer, D.R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, et al., Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, Med. Image Anal. 59 (2020).

[39] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imaging 38 (2019) 2281–2292.

[40] Y.-L. Xu, S. Lu, H.-X. Li, R.-R. Li, Mixed maximum loss design for optic disc and optic cup segmentation with deep learning from imbalanced samples, Sensors 19 (2019) 4401.

[41] Y. Zhang, X. Cai, Y. Zhang, H. Kang, X. Ji, X. Yuan, Tau: Transferable attention u-net for optic disc and cup segmentation, Knowl.-Based Syst. 213 (2021).

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inform. Process. Syst. 30 (2017).

[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[44] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Adv. Neural Inform. Process. Syst. 32 (2019).

[45] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, International Conference on Machine Learning, PMLR (2018) 4055–4064.

[46] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, arXiv preprint arXiv:1904.10509 (2019).

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

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

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

[50] E. Goceri, Fully automated and adaptive intensity normalization using statistical features for brain mr images, Celal Bayar Univ. J. Sci. 14 (2018) 125–134.

[51] E. Goceri, Intensity normalization in brain mr images using spatially varying distribution matching, in: 11th Int. Conf. on computer graphics, visualization, computer vision and image processing (CGVCVIP 2017), 2017, pp. 300–4.

[52] E. Goceri, Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases, Ninth international conference on image processing theory, tools and applications (IPTA), IEEE 2019 (2019) 1–6.

[53] Y. Yu, K. Adu, N. Tashi, P. Anokye, X. Wang, M.A. Ayidzoe, Rmaf: Relu-memristor-like activation function for deep learning, IEEE Access 8 (2020) 72727–72741.

[54] E. Goceri, Diagnosis of skin diseases in the era of deep learning and mobile technology, Comput. Biol. Med. 134 (2021).

[55] M. Tanaka, Weighted sigmoid gate unit for an activation function of deep neural network, Pattern Recogn. Lett. 135 (2020) 354–359.

[56] E. Goceri, Deep learning based classification of facial dermatological disorders, Comput. Biol. Med. 128 (2021).

[57] E. Goceri, Skin disease diagnosis from photographs using deep learning, in: ECCOMAS thematic conference on computational vision and medical image processing, Springer, 2019, pp. 239–246.

[58] C.-M. Feng, Y. Yan, G. Chen, Y. Xu, Y. Hu, L. Shao, H. Fu, Multi-modal transformer for accelerated mr imaging, IEEE Trans. Med. Imaging (2022).
[59] S. Luo, H. Dai, L. Shao, Y. Ding, C4av: learning cross-modal representations from transformers, European Conference on Computer Vision, Springer (2020) 33–38.
[60] A. Shin, M. Ishii, T. Narihira, Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision, Int. J. Comput. Vision 130 (2022) 435–454.
[61] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021) 1–23.
[62] M. Contributors, MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark,https://github.com/open-mmlab/mmsegmentation, 2020.
[63] P. Yakubovskiy, Segmentation models pytorch,https://github.com/qubvel/segmentation_models.pytorch, 2020.
[64] S. Pachade, P. Porwal, M. Kokare, L. Giancardo, F. Mériaudeau, Nenet: Nested efficientnet and adversarial learning for joint optic disc and cup segmentation, Med. Image Anal. 74 (2021).
[65] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[66] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inform. Process. Syst. 34 (2021) 24261–24272.
[67] L. Mou, Y. Zhao, H. Fu, Y. Liu, J. Cheng, Y. Zheng, P. Su, J. Yang, L. Chen, A.F. Frangi, et al., Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging, Med. Image Anal. 67 (2021).
[68] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sensing 12 (2020) 1662.
[69] S. Huang, J. Li, Y. Xiao, N. Shen, T. Xu, Rtnet: Relation transformer network for diabetic retinopathy multi-lesion segmentation, IEEE Trans. Med. Imaging (2022).
[70] Q. Huang, Y. Huang, Y. Luo, F. Yuan, X. Li, Segmentation of breast ultrasound image with semantic classification of superpixels, Med. Image Anal. 61 (2020).
[71] Q. Huang, Z. Miao, S. Zhou, C. Chang, X. Li, Dense prediction and local fusion of superpixels: A framework for breast anatomy segmentation in ultrasound image with scarce data, IEEE Trans. Instrum. Meas. 70 (2021) 1–8.
[72] Q. Huang, H. Luo, C. Yang, J. Li, Q. Deng, P. Liu, M. Fu, L. Li, X. Li, Anatomical prior based vertebra modelling for reappearance of human spines, Neurocomputing (2022).

Shuai Lu received B.E. and M.S. degrees from Beijing University of Chemical Technology, Beijing, China, in 2017 and 2020, respectively. He is currently a Ph.D. candidate at Beijing Institute of Technology. His research interests are image processing and machine learning.

He Zhao received Ph.D. degree from Beijing Institute of Technology, China in 2020 and received B.E. degree from Beijing Institute of Technology, China in 2014. His research interest is medical image processing, deep learning, computer vision.

Hanruo Liu received Ph.D. degree from University of East Anglia, UK in 2013. She is currently an associate professor in Ophthalmology with Beijing Tongren Eye Center. Her research interests are intelligent ophthalmology big data research.

Huiqi Li received Ph.D. degree from Nanyang Technological University, Singapore in 2003. She is currently a professor at Beijing Institute of Technology. Her research interests are medical image processing and computer-aided diagnosis.

Ningli Wang serves as the Director of Beijing Tongren Eye Center, Dean of School of Ophthalmology, Capital Medical University, Head of National Committee for the Prevention of Blindness, Advisory Board Member of Chinese Academy of Medical Sciences, President of Asia-Pacific Academy of Ophthalmology. His research interests are pathogenesis, diagnosis and treatment of glaucoma.