BMC Ophthalmology

RESEARCH

Open Access

Self-supervised pre-training for joint optic disc and cup segmentation via attention-aware network

Zhiwang Zhou[1*], Yuanchang Zheng[1,2], Xiaoyu Zhou[3], Jie Yu[4] and Shangjie Rong[5]

Abstract

Image segmentation is a fundamental task in deep learning, which is able to analyse the essence of the images for further development. However, for the supervised learning segmentation method, collecting pixel-level labels is very time-consuming and labour-intensive. In the medical image processing area for optic disc and cup segmentation, we consider there are two challenging problems that remain unsolved. One is how to design an efficient network to capture the global field of the medical image and execute fast in real applications. The other is how to train the deep segmentation network using a few training data due to some medical privacy issues. In this paper, to conquer such issues, we first design a novel attention-aware segmentation model equipped with the multi-scale attention module in the pyramid structure-like encoder-decoder network, which can efficiently learn the global semantics and the long-range dependencies of the input images. Furthermore, we also inject the prior knowledge that the optic cup lies inside the optic disc by a novel loss function. Then, we propose a self-supervised contrastive learning method for optic disc and cup segmentation. The unsupervised feature representation is learned by matching an encoded query to a dictionary of encoded keys using a contrastive technique. Finetuning the pre-trained model using the proposed loss function can help achieve good performance for the task. To validate the effectiveness of the proposed method, extensive systemic evaluations on different public challenging optic disc and cup benchmarks, including DRISHTI-GS and REFUGE datasets demonstrate the superiority of the proposed method, which can achieve new state-of-the-art performance approaching 0.9801 and 0.9087 $F1$ score respectively while gaining 0.9657 DC_{disc} and 0.8976 DC_{cup}. The code will be made publicly available.

Keywords Deep learning, Optic disc and cup segmentation, Medical image processing

Introduction

Glaucoma is the leading cause of irreversible vision damage, and it is reported that the number of glaucoma patients will increase to 110 million worldwide by 2040 [1, 2]. Glaucoma progresses silently without earlier noticeable symptoms. To prevent permanent vision loss, early treatment is extremely important. Recently, there have been three common diagnostic techniques for glaucoma including optic nerve head assessment [3], function-based visual field examination [4, 5], and intraocular pressure (IOP) assessment [6, 7]. Among these, some manual assessment methods of intraocular pressure

*Correspondence:
Zhiwang Zhou
zhiwangzhou@email.ncu.edu.cn
[1] Institute for Advanced Study, Nanchang University, Nanchang 330031, China
[2] Institute of Science and Technology, Waseda University, Tokyo 63-8001, Japan
[3] School of Transportation Engineering, Tongji University, Shanghai 200000, China
[4] School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300000, China
[5] School of Mathematical Sciences, Xiamen University, Xiamen 361000, China

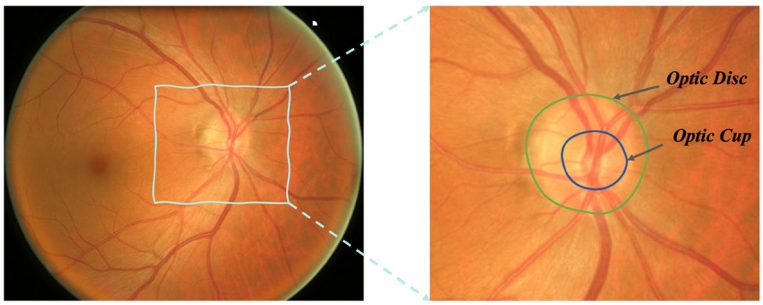

Fig. 1 Visualization of the retinal fundus images and the corresponding OD and OC images

measurement have not been widely used due to the differences in the human and equipment resources of each hospital. At the same time, these manual assessment methods consume a lot of manpower and are not conducive to large-scale pathological screening in hospitals, which may hinder their application in the real-life world.

To this end, the automatic retinal fundus photography strategy [8, 9] using deep neural networks becomes popular, which can help doctors to screen glaucoma. As shown in Fig. 1, the retinal fundus image shoots the main structure of the fundus including the optic disc (OD) and optic cup (OC). The vertical cup-to-disc ratio (CDR) can be calculated by the comparison of the diameter of the cup-to-disc. The normal CDR is 0.3 to 0.4. A larger CDR may indicate glaucoma. The accurate CDR can be calculated from the segmented optic disc and cup area [10].

Currently, deep learning methods [11] show great performance towards the accurate optic disc and cup segmentation. The most prominent architecture is U-Net [12], which performs skip connections to fuse multi-level information. Later, M-Net [13] further improves the performance by injecting the domain-specific knowledge that the optic cup lies in the optic disc and adds side-output layers to acquire more supervision. JointRCNN [14] explores joint OD and OC segmentation with a disc attention module and makes full use of the prior knowledge that the optic disc and cup are approximately ellipses. PM-Net [15] performs OD and OC detection and also utilises the prior knowledge that the optic cup lies in the optic disc. Afterwards, Yin et al. [16] inject a guided filter into U-Net to restore the structure information loss caused by down-sampling operations. The domain-specific knowledge helps to increase the performance. Some algorithms [17, 18] try to adopt GAN [19] to assist in enhancing segmentation performance. Recently, some of the up-to-date networks [20, 21] utilize vision transformer [22, 23] to conduct medical image segmentation and achieve state-of-the-art performance.

However, two challenging questions remain on the optic disc and cup segmentation task: (1) How to design an effective network to capture the global information of the input images and enjoy fast execution; (2) How to solve the problem that optic disc and cup training samples are not sufficient enough. For the first issue, most previous works will explore non-local skill [24] to capture the semantics of the medical images. However, excessive convolution operations complicate the calculation, which is prone to overfitting. Besides, some of the token-based transformer methods are too large, which take a lot of computing resources and execute slowly, which is not practical in a real medical environment. For the second point, collecting medical images (e.g. optic disc and cup images) is much more difficult than in the common computer vision field data (i.e. the public COCO [25] and PASCAL [26] segmentation datasets can be widely collected on the Internet) due to some pathological privacy issues. Therefore, training the deep learning networks especially some transformer or GAN-based networks may achieve unsatisfactory results when training data is rare.

In this paper, to tackle the above-mentioned issues, we propose a novel attention-aware segmentation model equipped with the multi-scale attention module in the pyramid structure-like encoder-decoder network, which can efficiently learn the global semantics and the long-range dependencies of the input images. The proposed multi-scale attention module is different from the traditional attention mechanism in transformer [22], we design a more powerful multi-scale nearest neighbour semantic pixel matching operation to enable the network to capture more useful visual hints. Besides, different from some previous methods that require multi-stages [27] for segmentation, our framework is a one-stage network, which does not need first to crop the key region and then segment the image. Furthermore, considering that the scarcity of medical imaging images leads to instability in training deep network models, we designed a new self-supervised contrastive learning training paradigm, which can learn the discriminative representation of the image in an unsupervised manner.

Meanwhile, we also proposed a novel loss function to make use of this knowledge by constraining the subtraction of the optic cup from the optic disc in the optic rim.

To demonstrate the effectiveness of the proposed method, extensive systemic evaluations on different public challenging optic disc and cup benchmarks including DRISHTI-GS and REFUGE datasets reveal the superiority of the proposed method, which can achieve new state-of-the-art performance. Our main contributions are summarized as follows:

- We experimentally analyze unsolved challenges in optic disc and cup segmentation tasks, and we take the early step to explore self-supervised contrastive learning to tackle the drawbacks in the medical image field.
- We propose a brand new attention-aware segmentation model equipped with the multi-scale attention module, which explores nearest neighbour semantic pixel matching operation to enable the network to capture more useful visual cues for optic disc and cup segmentation.
- We design a novel loss function to make use of the knowledge by constraining the subtraction of the optic cup from the optic disc in the optic rim, which can help to enhance the learning representation.
- Extensive experimental results conducted on different challenging benchmarks validate the effectiveness of the proposed network and training paradigm, which can achieve new state-of-the-art performance.

Related work

In this section, we will provide a brief overview of different types of existing traditional and medical image segmentation methods. Specifically, we will summarize the ordinary scene image segmentation methods based on CNNs or transformers, and then review the expansions of these methods in the medical image domain. Finally, we will discuss the self-supervised training methods.

Non-Learning-Based Image Segmentation: Image segmentation is a crucial preprocessing for image recognition and computer vision. Conventional image segmentation usually means traditional semantic segmentation. Image segmentation in this period (about 2010), due to limited computer computing power, could only process some grayscale images in the early days, and later could process rgb images. The segmentation in this period mainly depends on extracting low-level features of images and then segmenting them, some methods have emerged: Ostu [28], FCM [29], watershed [30], N-Cut [31], etc. Subsequently, with the improvement of computing power, people began to consider obtaining semantic segmentation of images. The semantics here are currently low-level semantics, which mainly refers to the categories of segmented objects. At this stage (probably from 2010 to 2015), people considered using machine Learning methods for image semantic segmentation. With the emergence of FCN [32], deep learning officially enters the field of image semantic segmentation.

Image segmentation based on CNNs: Image segmentation is a vital branch in the field of deep learning, which can help analyze the pixel-level content of images. The first step of traditional image segmentation usually need to collect a large amount of data (i.e. collect the images from the Internet), and then requires enormous annotations to train a strong network for satisfactory performance. Long et al. [32] proposed fully convolutional networks (FCNs), which enjoys advantageous in end-to-end dense representation modeling, laying the foundation for modern semantic segmentation algorithms. However, FCNs suffer from the limited visual context with local receptive fields of the convolutional operations. Later, DeepLab [33–35] explores new solution by enlarging receptive fields with dilated operation and spatial pyramid pooling. Moreover, scholars try to design different pyramid-like structure network [36, 37] for multi-scale learning. Some other researchers utilize U-Net [12] like structure and devise many promising encoder-decoder network [38, 39] solutions. Furthermore, many existing works adopt auxiliary information like boundary clues [40, 41] and optical flow [23] hints to boost performance. Recently, many cutting-edge semantic segmentation methods inject neural attention [24, 42–45] for improving the extracted semantic features. As for medical image segmentation, U-Net series can help achieve competitive performance. Edupuganti et al. [46] adopts an end-to-end encoder-decoder network to segment optic disc and cup with the edge loss function. Shankaranarayana et al. [47] utilizes FCN network with adversarial training for OD and OC joint segmentation. Later Fu et al. [13] proposes M-Net with multi-label strategy for segmentation. More recently, some other variants networks like U-Net++ [48], U-Net3+ [49] and DenseU-Net [50] also shows acceptable performance in medical image segmentation. In MDC-Net [51], multi-scale dilated convolution is adopted to increase the receptive field of the model and multiple residual connections are used to utilize feature information from different scales. Zhu et al. [52] designed a network consisting a down-sampling path extracting the features and an up-sampling path restoring the down-sampled features. The features are automatically extracted from the images through the convolutional operators during the down-sampling procedure. Besides, some other latest works [53, 54] both adopted deep- learning-based

method to automaticly for pathological analysis. Nevertheless, although the mentioned methods have used variants of encoder-decoder architecture, they limit the local context encoding by convolutional layers. To this end, some researchers' focus gradually shifts to vision transformer.

Image segmentation based on transformer: Recently, more and more segmentation models [55, 56] are built upon the attention vision transformer (ViT) [22] to capture the global long-range dependencies of the image pixels. Zheng et al. [57] explores ViT as backbone and utilize a standard CNN as decoder for segmentation. Swin Transformer [58] designs a variant of ViT architecture with shifted windows and equipped with a pyramid FCN decoder. Robin et al. [56] proposes a transformer encoder-decoder architecture for semantic image segmentation inspired by DETR [59]. As in medical image segmentation, TransUNet [60] designs a U-Net like transformer network to locate the image token spatial information. TransAttUnet [61] improves the U-shaped architecture segmentation network with multi-level guided attention and multi-scale skip connection. DS-TransUNet [62] adopts the Swin Transformer block [58] to both the encoder and the decoder and achieve competitive performance. Liu et al. [63] proposed a network which consists of a transformer-based branch and a convolution-based branch, and the information is exchanged between the inner layers. However, all the above-mentioned medical segmentation methods fails to take full advantage of the spatial detail information from the transformer-based network since the medical training images are insufficient, which greatly increases the difficulty of transformer network training.

Self-supervised training methods: In recent years, unsupervised or self-supervised learning has attracted much attention. Some previous methods design the pretext task like image colorization [64], image jigsaw complement [65, 66] and rotation prediction [67], etc. With the birth of the contrastive learning paradigm, MOCO [68] learns the feature representation using a dictionary look-up pretext task from a perspective on contrastive learning. SimpleCLR [69] learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. SimpleCLR and MOCO both adopt a siamese network with contrastive learning and achieve encouraging performance. Besides, some video self-supervised learning methods focus on temporal hints for learning. Xu et al. [70] model the self-supervised learning by shuffling the video and predicting the final orders. Some other works [71, 72] pay attention to complete video playback speeds. Therefore, in this paper, we explore the self-supervised learning for optic disc and cup segmentation based on contrastive learning and explore several data augmentation methods to acquire a pre-train model for segmentation.

Methodology

In this section, we will introduce the pipeline of the proposed network. First, the overview of the network architecture is presented, followed by a detailed description of each component. Then, we introduce the self-supervised pre-training paradigm of the proposed network. Finally, we will explicitly elaborate on loss functions to train our network.

Network architecture

As shown in Fig. 2, we depict the overall network architecture of the proposed network. Given the input image I, our target is to segment out the accurate mask output O that represents the optic disc and cup. To achieve this goal, we propose an attention-aware segmentation network, which is based on an encoder-decoder structure. Specifically, we adopt the CNN encoder (i.e. ResNet [73]) to extract the multi-scale feature map (F_1, F_2, F_3, F_4). For each layer, we add the proposed multi-scale attention module followed by the convolutional feature maps to model the global semantic hints. The aggregation attention module is followed by the last feature layer for enhancement. The overall network is in the pyramid structure-like architecture with different skip-connection. The decoder is responsible for upsampling and predicting the final masks.

Multi-Scale Attention Module: Concretely, we introduce the proposed multi-scale attention module followed by each convolutional layer for global feature modelling, as shown in Fig. 3. For simplicity, we consider (F_1, F_2, F_3, F_4) to be F. Specifically, we first reshape the feature map $F \in \mathbb{R}^{B*N\times C\times h\times w}$ (w and h denote the spatial size of the feature map, C is the feature dimension, B and N denote the batch size and group size, respectively) into a sequential flattened patch tokens $X \in \mathbb{R}^{B\times C\times P}$, where $P = N * h * c$. Then, we adopt the multi-head sel-attention mechanism (MHSA) for these tokens, which can be computed as follows:

$$X = MHSA(LN(X)) + X, \tag{1}$$

where LN indicates LayerNorm [74] function.

Afterwards, different from the standard vision transformer [22] operation that directly applies linear layers after the MHSA, we propose the multi-scale attention multilayer perceptron here. Specifically, as shown in Fig. 3 right, for each query patch token, it will first select its corresponding top-k nearby potential tokens. Mathematically, for each token, we first use ℓ_2- distance to measure the relationship between the two arbitrary

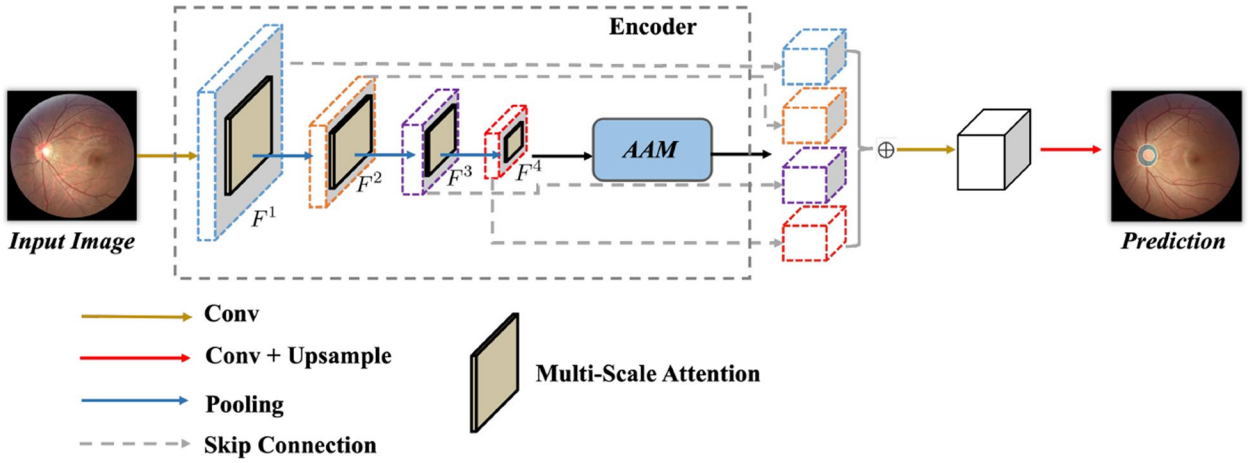

Fig. 2 The overall architecture of the network. The given input image *I* is first fed into the encoder, yielding the multi-scale feature maps *F*. We employ the proposed multi-scale attention module followed by each convolutional layer for feature enhancement. Then, we inject the designed aggregation attention module followed by the last layer for feature fusion. The decoder is bridged behind the encoder in the pyramid-like structure for final mask prediction

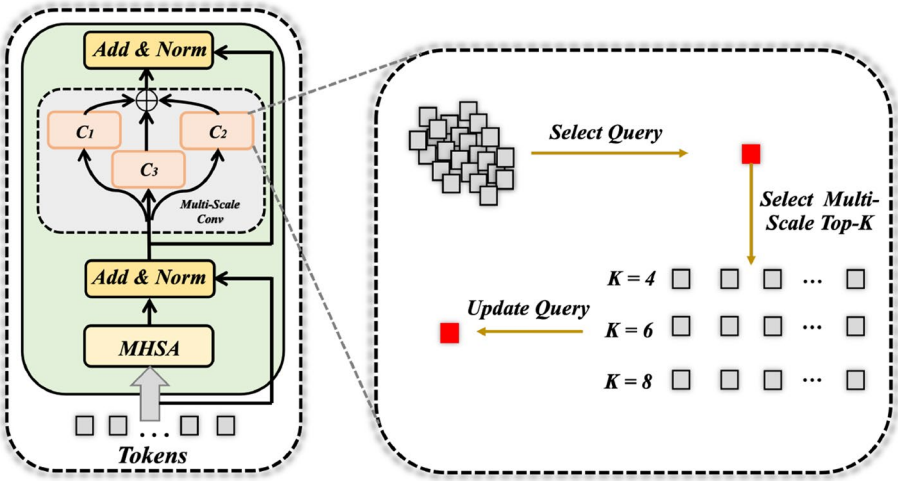

Fig. 3 Illustration of the proposed multi-scale attention module. For each query image token pixel, it will match with its top-*K* potentially corresponding tokens. Afterwards, it will be updated by aggregating different sub-region representations using the multi-layer perceptron operation

patches. Since we use normalized channel features, by removing the constant, the matrix $S \in \mathbb{R}^{B*N\times Q\times Q}$ ($Q = h * w$ is the number of patches) can be formulated as:

$$S = F^T F. \tag{2}$$

We then perform *KNN* operation on the matrix using the PyTorch built-in function (*torch.topk*(*S*)) to select its potentially corresponding target patches, which will produce a tensor $\hat{X} \in \mathbb{R}^{B\times C\times N\times K}$, which indicates the patches along with their top-*K* semantically related patches. Here, we chose k = 4, 6, 8. In a word, for each query token, it can match tokens of different scales that are semantically similar. When *k* becomes larger, it contains more relevant tokens. In order to update the token's features, the multi-scale Multi-layer Perceptron (MSMLP) embedding operation is performed as:

$$X = \sum_{i}^{k=4,6,8} \underset{x\in\hat{X}}{\text{Max}}(\ \varphi\ (x_1, x_2, ..., x_{k_i})),\ \ X \in \mathbb{R}^{B\times C\times N}, \tag{3}$$

where $\varphi(\cdot)$ is the local feature modelling function, and we here use two 1×1 convolution layers followed by the

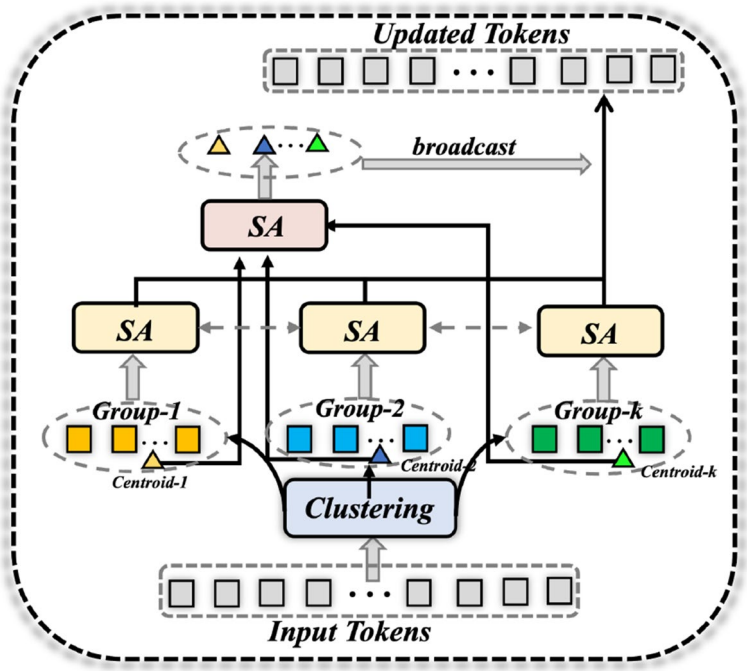

Fig. 4 Illustration of the proposed aggregation attention module. The input tokens are first clustered into different groups. For each group, the self-attention operation is performed individually over the cluster centroid and cluster tokens. Ultimately, the updated cluster centroid and the group features are aggregated together to form a new feature vector

ReLU activation function, and we fuse the multi-scale features by tensor addition.

Ultimately, the whole process of each multi-scale attention module can be formulated as follows:

$$\begin{aligned} X' &= \mathrm{MHSA}(\mathrm{LN}(X)) + X, \\ X &= \mathrm{MSMLP}(\mathrm{LN}(X')) + X', \end{aligned} \tag{4}$$

Aggregation Attention Module: Furthermore, we design an aggregation attention module following the last layer F_4. As shown in Fig. 4, for the input tokens, we first utilize a mini-batch k-means clustering algorithm to group query into $C = 3$ clusters adaptively followed the implementation in [75]. Here, we want to cluster the image tokens into the optic disc, cup and background. Concretely, for the cluster centroid and cluster tokens, we both adopt the self-attention mechanism to update the global features as follows:

$$Q = \mathbf{W}_Q X + \mathbf{B}_Q, \;\; K = \mathbf{W}_K X + \mathbf{B}_K, \;\; V = \mathbf{W}_V X + \mathbf{B}_V, \tag{5}$$

where $\mathbf{W}_Q, \mathbf{W}_K$ and $\mathbf{W}_V$ are three learnable linear weight matrices, while $\mathbf{B}_Q, \mathbf{B}_K$ and $\mathbf{B}_V$ are weight vectors. After having the Q, K, V, the global self-attention mechanism can be formulated as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{SoftMax}(QK^T/\sqrt{D})V, \tag{6}$$

where D is the feature dimension. Afterwards, the updated cluster centroid are broadcasted into the shape of the input tokens and combined with the updated group cluster tokens to produce the output.

Finally, the decoder with skip connections fuse multi-level information and produces the segmentation prediction of the optic disc and cup. However, the deep guidance network does not inject domain-specific knowledge. Motivated by the prior knowledge that the subtraction of the optic cup from the optic disc is the optic rim, we design a multi-label loss to inject this knowledge. As suggested by PM-Net [15], the multi-label head learns an independent binary classifier for each class. Furthermore, to segment a glaucoma fundus image that the OC occupies the most area of OD, the multi-label head balances the pixel number for OD and OC since the classifier is independent for OD and OC. However, the multi-label head does not make full use of the ground truth. We further add the constraint that the subjection of OC from OD is the optic rim (OR). The proposed loss function treats the segmentation problem as three binary classification problems with single label: [OD, OD^{-}], [OC, OC^{-}], [OD-OC, OR^{-}] ($^{-}$ represents negative examples). Then, the loss function can be defined as follows:

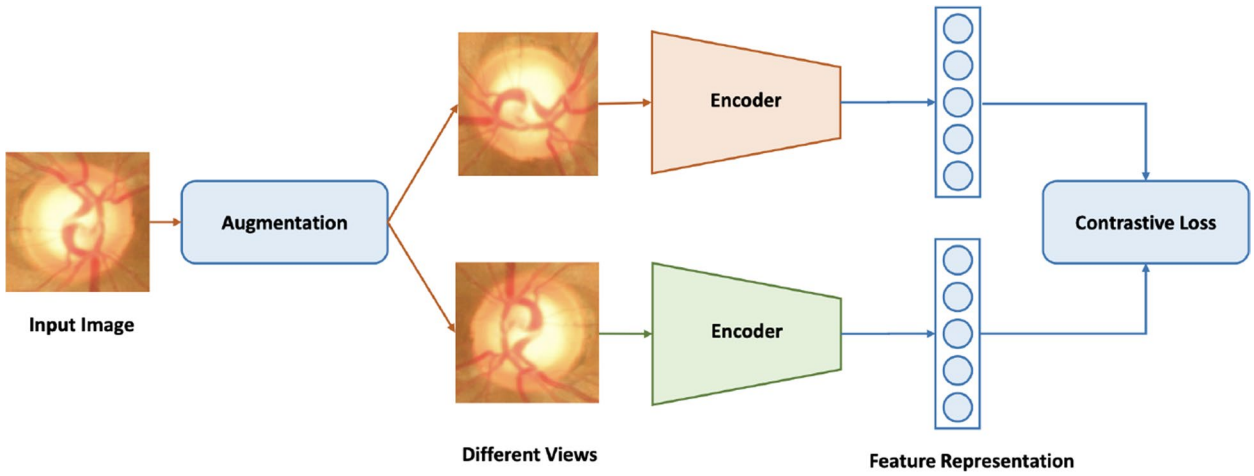

Fig. 5 The framework of the proposed self-supervised method. An input image is augmented into two different views. Then the network learns to maximize agreement using a contrastive loss

$$\begin{aligned} Loss_D &= g_{D,i} log p_{D,i} + (1 - g_{D,i}) log(1 - p_{D,i}), \\ Loss_C &= g_{C,i} log p_{C,i} + (1 - g_{C,i}) log(1 - p_{C,i}), \\ Loss_R &= g_{R,i} log(p_{D,i} - p_{C,i}) + (1 - g_{R,i}) log(1 - (p_{D,i} - p_{C,i})), \end{aligned} \tag{7}$$

where, $g_{R,i}$ represents the ground truth of the optic rim and can be calculated by $g_{R,i} = g_{D,i} - g_{C,i}$. Finally, the total segmentation loss can be defined as follows:

$$Loss = \frac{1}{3} * (Loss_D + Loss_C + Loss_R). \tag{8}$$

Here we treat the segmentation problem as three binary classification problems. $p_{D,i}$, $p_{C,i}$ represents the predicted probability of OD and OC respectively for pixel i. $g_{D,i}$, $g_{C,i}$ represents the ground-truth label of pixel i for OD and OC respectively. We add a constraint that the predicted OD area subject the predicted OC area is close to the ground truth OR area.

Self-supervised pretrainng

Self-supervised learning aims to learn feature representations from a large amount of unlabeled data, which is usually achieved by setting different pretext tasks and utilize easy-to-obtain automatically generated supervision. In the image domain, [64] perform image colorization pretext to establish a mapping from objects to colors that learn the potential features of the images. Some previous works [65, 66] try to solve jigsaw problems to learn the information of different patches in the images. Komodakis et al. [67] proposed a simple rotation transformation to make the network to predict different rotation degrees of the images to identify objects' features. Later, such transformations as scaling, warping and inpainting have been applied to the latest work [76]. Leveraging the merits of contrastive learning that focus on semantic information rather than too much on pixel details, most of the current works [68, 77, 78] explored to construct positive pairs and negative pairs for feature learning.

The self-supervised framework is shown in Fig. 5. We perform two separate data augmentation operations to obtain two different views of an input image. Then, we train our network to maximize the agreement using a contrastive loss. We randomly sample a mini-batch of N examples and define the contrastive prediction task on pairs of augmented examples derived from the mini-batch. The two views are similar to each other and dissimilar to other pairs. The similarity is measured by the dot product. The InfoNCE loss function [79] is considered in the paper to train the network, and it is defined as follows:

$$L_q = -log \frac{exp(q \cdot k_+)/\tau}{\sum_{i=0}^{K} exp(q \cdot k_i)/\tau}. \tag{9}$$

where, q and k_+ are the positive pair. q and k_i are K negative pairs. The sum is over one positive and K negative samples. τ denotes a temperature parameter [80].

Moreover, RotNet [67] trains a network to recognize the rotation transformation for unsupervised learning and motivated by this, we also apply the rotation transform augmentations. Similar to RotNet, we only rotate the image to 0°, 90°, 180°, 270°. MOCOv2 [81] states that the gaussian blur is also helpful for learning, so we also perform gaussian blur with σ between [0, 0.5]. The sharpening operation is to sharpen images and alpha-blend the result with the original input images. When $\alpha = 0$, only the original image is visible. When $\alpha = 1$, only its sharpened version is visible. We also conduct γ contrast with γ between [0.5, 2] to augment the data. For segmentation, the output resolution is usually large (For example,

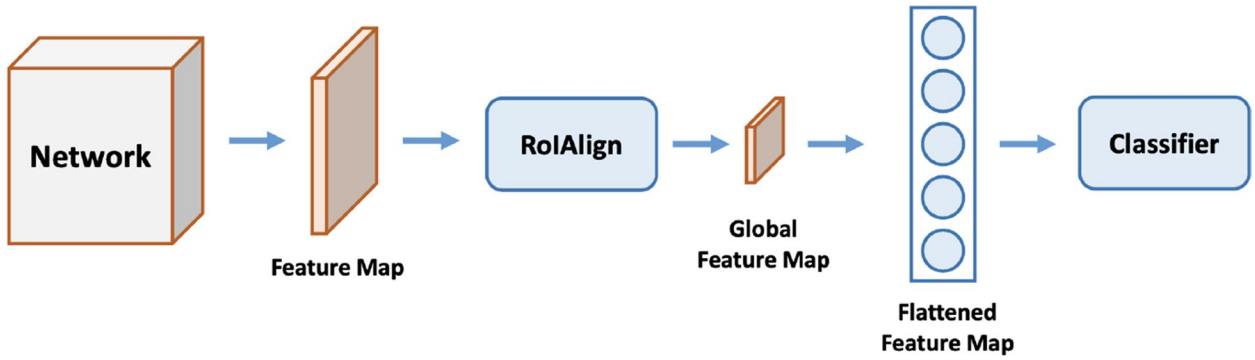

Fig. 6 The self-supervised training head for segmentation. The input image is first encoded by the network encoder. Then RoiAlign operation is applied to obtain a smaller global feature map for efficient learning. The final fully connected layer flattens the feature for contrastive learning

512 × 512), after the flatten operation, it will make the following fully connected layers too large to train. To solve this issue, we adopt the RoiAlign layer proposed by Mask-RCNN [82] to obtain a smaller global feature map (24 × 24). The global feature map is flattened and sent to the classifier for contrastive learning, where the entire process is shown in Fig. 6. Moreover, Table 1 shows the augmentation we used for self-supervised learning.

Experiment

In this section, we will first elaborate on the details of experiment settings including datasets, metrics and detailed implementation. We then analyze the ablation studies and the evaluation results are finally given to compare with state-of-the-art methods.

Datasets

The experiments are conducted on different challenging datasets including the REFUGE dataset[1] [83] and DRISHTI-GS datasets[2] [84]. Specifically, the DRISHTI-GS dataset contains 101 images while the REFUGE dataset contains 1200 images. For the REFUGE dataset, we pre-train the network on the whole DRISHTI-GS dataset (101 images), REFUGE training and testing dataset (800 images) and evaluate the model on the REFUGE validation dataset (400 images). For the DRISHTI-GS dataset, it contains 101 retinal fundus images that 50 images are for training and 51 images are for testing. We pre-train the network on the DRISHTI-GS training set (50 images), the whole REFUGE datasets(1200 images) and fine-tune the pretrained on DRISHTI-GS training dataset(50 images), and finally evaluate the model on the DRISHTI-GS test dataset.

[1] https://refuge.grand-challenge.org/Home2020/

[2] https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation

Note that the original images for the DRISHTI-GS dataset were provided by Aravind eye hospital, Madurai, who selected an approximately equal number of men and women, aged 40-80 years, with glaucoma and non-glaucoma patients for fundus image acquisition. All images were acquired with dilated pupils and captured according to the following data collec- tion protocol: OD-centred High-resolution fundus images of 2896 × 1944 pixels were acquired with a field of view of 30°. Finally, by removing the surrounding non-fundus black area, the image area with the retinal structure is extracted from the original image, thereby obtaining a fundus image with a resolution of about 2047 × 1760. As shown in Fig. 2, each image was manually labelled by four glaucoma specialists with 3, 5, 9 and 20 years of experience, respectively. REFUGE dataset was organized as a half day Challenge in conjunction with the 5th MICCAI Workshop on Ophthalmic Medical Image Analysis (OMIA) with the goal of the challenge is to evaluate and compare automated algorithms for glaucoma detection and optic disc/cup segmentation on a common dataset of retinal fundus images. With this challenge, a large dataset of 1200 annotated retinal fundus images are made available. In addition, an evaluation framework has been designed to allow all the submitted results to be evaluated and compared with one another in a uniform manner. In general, these two datsets are currently the largest, most authoritative, and most challenging datasets. Therefore, we choose

Table 1 The data augmentation used in pretraining

Augmentation	Parameters
Rotation	0°, 90°, 180°, 270°
Sharpen	$\alpha = [0, 1]$
GammaContrast	$\gamma = [0.5, 2.0]$
GaussianBlur	$\sigma = [0, 0.5]$

Table 2 Analysis of the different proposed components on the DRISHTI-GS dataset

Method	OD		OC	
	F1	*BLE*	*F1*	*BLE*
Baseline	0.9334	9.07	0.8265	19.97
+ Standard Attention	0.9388	8.98	0.8314	19.19
+ Multi-Scale Attention	0.9488	8.17	0.8486	18.24
+ Aggregation Attention Module	0.9433	8.21	0.8432	18.47
+ Self-supervised Pretrain	0.9517	7.83	0.8636	15.72
+ Multi-Scale Attention + Aggregation Attention Module	0.9652	7.01	0.8879	13.98
+ Multi-Scale Attention + Self-supervised Pretrain	0.9788	6.28	0.9001	11.03
+ Aggregation Attention Module + Self-supervised Pretrain	0.9763	6.34	0.8978	11.54
Ours (full)	**0.9801**	**6.21**	**0.9087**	**10.07**

these two data sets to verify the effectiveness of our proposed network.

Metircs

Following the previous works [8, 27] strictly, we evaluate the performance of the proposed method using the F1 score, Boundary distance Localization Error (BLE) and the Dice coefficients (DC). Among them, the definition of F1 can be computed as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{10}$$

where Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$. TP, TN, FP and FN represent true-positive, true-negative, false-positive and false-negative cases, respectively.

As for Dice coefficients (DC), it can be defined as follows:

$$DC = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{11}$$

As for BLE, it can better reflect the segmentation effect of the boundary, which can be computed as follows:

$$BLE(C_0, C_g) = \frac{1}{N} \sum_{\theta=0}^{N-1} \sqrt{(d_g^\theta)^2 - (d_0^\theta)^2}, \tag{12}$$

where d_g^θ and d_0^θ indicate the Euclidean distance from the centre point of OD in the θ direction to C_g and C_0, and 24 equidistant points (N = 24) are set in the evaluation. Note that the smaller the BLE, the better the segmentation effect.

Implementations details

For supervised training, we train the entire network for 100 epochs. The learning rate decays 10 times every 50 epochs. The training is performed on one NVIDIA TITAN XP GPU. The initial learning rate is set to 0.0001. The batch size is set to 1. It takes almost 6 hours to train a network. For self-supervised pretraining, we train the network for 30 epochs with an initial learning rate of 0.0001 and decay 10 times every 15 epochs. The batch size is set to 8 on a single GPU since a bigger batch size may be hard to train and unstable as suggested by [69].

Ablation study

To explore the components of our proposed method, we first conduct extensive analysis on DRISHTI-GS datasets [84] to demonstrate how they help to improve feature learning for optic disc and cup segmentation. Specifically, we will analyze the effect of the proposed multi-scale attention module, aggregation attention module, self-supervised learning strategy, and standard attention module, etc.

As shown in Table 2, we conduct plenty of ablative analysis of the proposed modules. Specifically, we can observe that our baseline model can only achieve acceptable but not competitive performance. When we inject the proposed multi-scale attention module, the segmentation performance of both the optic disc and cup can be improved by around 0.1 - 0.2 and 1.0 - 1.5 in terms of the F1 score and BLE metric. It is worth noting that we also conduct experiments using the standard attention block proposed in [22], as can be seen in the second row of Table 2. The results show that our proposed modified multi-scale attention module is better than the traditional one, which reveals the effectiveness of the proposed module. Furthermore, we explore the usefulness of the proposed aggregation attention module. Likewise, this component can also boost the network performance. In terms of self-supervised training, since we can use unlabeled data from other datasets for a large amount of unsupervised pre-training, we can first learn the encoder weights of a segmentation network with appropriate parameters. Then, we can fine-tune

Table 3 5-fold Cross-validation on the DRISHTI-GS dataset

Method	OD		OC	
	F1	*BLE*	*F1*	*BLE*
1-fold	0.9801	6.21	0.9087	10.07
2-fold	0.9799	6.23	0.9066	10.12
3-fold	0.9800	6.19	0.9085	10.10
4-fold	0.9797	6.25	0.9083	10.11
5-fold	0.9798	6.24	0.9082	10.10

Table 4 Comparison of quantitative results of different methods on the DRISHTI-GS dataset. Some of the results are derived from [8]

Method	Year	Params (M)	FLOPs (G)	Times (ms)	OD		OC	
					F1	*BLE*	*F1*	*BLE*
FCN [32]	2014	48.2	136.2	500	0.9321	8.90	0.8170	21.83
U-Net [12]	2015	65.9	158.3	623	0.9600	7.23	0.8500	19.53
M-Net [13]	2018	71.8	164.5	650	0.9590	7.97	0.866	17.05
Stack-U-Net [86]	2018	86.6	178.0	700	0.9700	6.47	0.8900	14.39
POSAL [17]	2019	68.5	-	-	0.9650	-	0.8580	-
CE-Net [85]	2019	71.3	125.0	550	0.9688	5.04	0.8699	16.06
JointRCNN [14]	2020	-	-	-	0.9640	-	0.8640	-
BGA-Net [87]	2021	75.6	148.5	600	0.9750	7.01	0.8980	14.37
DCGAN [88]	2022	102.9	-	-	0.9746	7.35	0.8631	18.69
RSAP-Net [8]	2022	87.6	235.0	800	0.9752	6.33	0.9012	11.97
Ours	2023	54.6	147.6	560	**0.9801**	**6.21**	**0.9087**	**10.07**

the entire segmentation network from the perspective of a global optimal solution. As can be seen in the fifth row of Table 2, self-supervised pre-training can benefit the segmentation performance by a large margin, which also demonstrates the necessity of self-supervised pre-training from unlabeled data. Ultimately, we have tried to combine the different proposed modules in pairs, and we can find that there are different levels of advanced improvements in this task. When we use our full model (all the proposed components are used), we can achieve the best performance.

Cross-validation

As shown in Table 3, we conduct a 5-fold verification experiment. Specifically, since the dataset itself is divided into a training set and a test set, here we divide the data set into 5 equal parts and conduct cross-validation experiments. It can be seen that the effect of our algorithm on each fold is relatively average, which also reflects the robustness and effectiveness of our algorithm.

Compare with the state-of-the-art methods

Compared with DRISHTI-GS challenge: To demonstrate the superiority of the proposed network, we compare the experimental results with the existing state-of-the-art segmentation methods, as shown in Table 4. We can observe that some previous representative works like FCN [32] and U-Net [12] networks fail to achieve satisfactory performance, whose F1 score and BLE metric are all below average. Although there were some improved methods later, such as POSAL [17], CE-Net [85] and JointRCNN [14]. Most of these methods only focus on how to improve the design of the model artificially and do not take into account the scarcity of medical data and the expansion of the convolutional receptive field. These methods are easily interfered by fundus blood vessels and can not segment the boundary contour well. Our designed method comprehensively considers the existing segmentation problems from both the data and model perspectives, and we can finally achieve the best performance over the previous methods. Besides, the parameter of our network is also competitive, which guarantees the effectiveness of execution speed.

Compared with REFUGE challenge: We also compare our segmentation results with state-of-the-art methods on the REFUGE challenge task. As shown in Table 5, the first 12 rows are the results from different participating

Table 5 Comparison of quantitative results of different methods on the REFUGE dataset. Some of the results are derived from [27]

Method	DC_{disc}	DC_{cup}
CHUKMED	0.9602	0.8826
Masker	0.9496	0.8837
BUCT	0.9525	0.8728
NKSG	0.9488	0.8643
VRT	0.9532	0.8600
AIML	0.9505	0.8519
Mammoth	0.9361	0.8667
SMILEDeepDR	0.9386	0.8367
NightOwl	0.9487	0.8257
SDSAIRC	0.9436	0.8315
Cvblab	0.9077	0.7728
WinterFell	0.8772	0.6861
M-Net [13]	0.9436	0.8315
POSAL [17]	0.9602	0.8826
Mask R-CNN [82]	0.9504	0.8546
Liu et al. [27]	0.9601	0.8903
Ours	**0.9657**	**0.8976**

teams, and the rest are the results of some publicly available deep learning methods. Notably, our method can achieve the best segmentation result on both the optic disc and cup. The performance on both the REFUGE dataset and DRISHTI-GS datasets all reflect the generalizability and feasibility of the proposed network and training paradigm.

Qualitative visualization: Fig. 7 shows some qualitative visualizations of our proposed method on both the REFUGE dataset and DRISHTI-GS datasets. As can be seen, our method can yield high-quality accurate masks, which demonstrates that our method can be applied to practical medical applications.

Computational Complexity Analysis: As shown in Table 4, we also provide the computational complexity of different state-of-the-art networks, including network parameters, floating-point operations per second (GFLOPs) and running time. We can observe that although some of the previous CNN-based networks enjoyed low computational complexity, they failed to achieve satisfactory performance. The proposed framework can make a good balance between network performance and computational complexity.

Discussion: Through systematic experiments and evaluations such as the qualitative comparative experiments shown in Tables 4 and 5, we can see that our method has more advantages than existing advanced methods. We believe there are the following reasons: (1) First, we make an early attempt to adopt an unsupervised pre-training strategy, which can use a large amount of unlabeled data for image representation learning so that the network can be optimized in a better direction; (2) The proposed attention mechanism can effectively help the network expand the receptive field of learning, allowing the network to learn the global information of medical images and effectively improve the segmentation effect. Overall, the network we designed can efficiently solve the current joint optic disc and cup segmentation tasks.

Limitation: As a common practice in the deep learning area, every framework will have certain limitations. Among them, we generously admit that our method will be somewhat cumbersome in terms of training time because it is trained in two steps (i.e., self-supervised pre- train and then combined with supervised training). However, we believe that self-supervised training is a new training strategy that does not increase the number of parameters of the network operation. In addition, because we use the transformer-based attention mechanism, this will cause our network to be more computationally intensive than traditional CNN-based networks. However, the current GPU acceleration technique can already solve these problems well.

Future Work: In future work, we will continue to explore various variants of attention mechanism structures, hoping to effectively solve specific problems in the field of medical images. At the same time, we will also focus more on designing lightweight networks to be more suitable for practical applications in the medical field. Finally, we will also focus on other optimization methods in the field of self-supervised learning to solve problems such as training model collapse and parameter sensitivity and use the capabilities of large models to solve some issues such as data imbalance and data migration.

Conclusion

In this paper, we deeply discuss and analyze the unresolved challenges in medical segmentation especially for optic disc and cup segmentation. We then propose a novel attention-aware encoder-decoder network equipped with the designed multi-scale attention block and the aggregation attention module, which is capable of helping the network to capture the global dependencies of the input image tokens. Furthermore, we introduce a novel loss function to make use of the knowledge by constraining the subtraction of the optic cup from the optic disc in the optic rim and adopt contrastive learning for self-supervised pre-training. This strategy can alleviate the shortcomings of a small

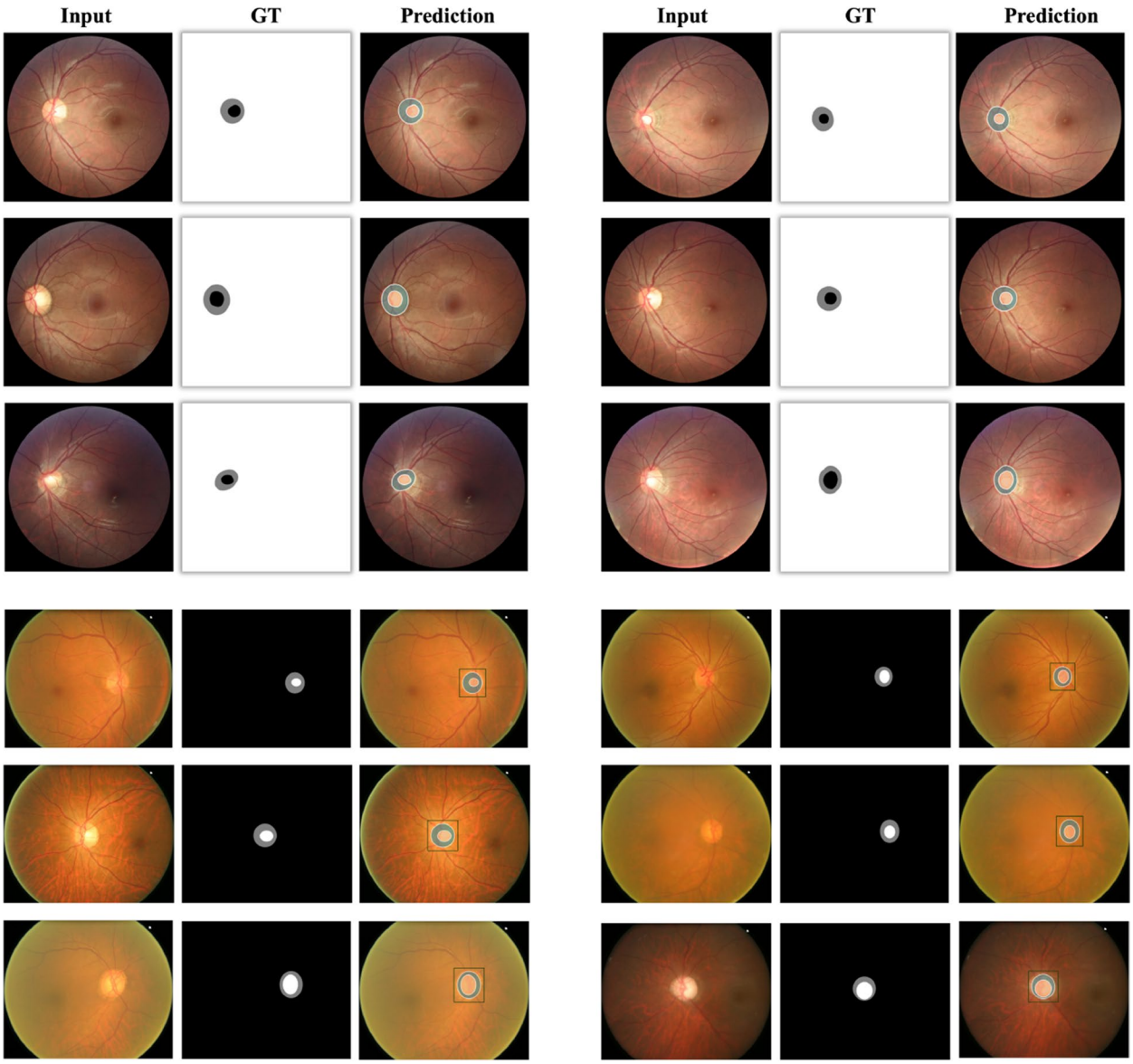

Fig. 7 Visualizations of the optic disc and cup segmentation on REFUGE dataset and DRISHTI-GS dataset

amount of image training data in the medical field. Finally, extensive experimental results conducted on different challenging benchmarks all demonstrate the superiority of the proposed network and training paradigm, which can outperform other state-of-the-art methods.

Acknowledgements
There is no Acknowledgments.

Authors' contributions
Conceptualization, Z.Z. and Y.Z.; methodology, Z.Z. and Y.Z.; software, Z.Z. and Y.Z.; validation, Z.Z. and Y.Z.; formal analysis, Z.Z. and Y.Z.; investigation, X.Z. and J.Y. and S.R; resources, Z.Z. and Y.Z.; data curation, Z.Z. and Y.Z.; writing-original draft preparation, Z.Z. and Y.Z.; writing-review and editing, Z.Z. and Y.Z. and X.Z. and J.Y. and S.R; visualization, Z.Z. and Y.Z.; supervision, Z.Z. and Y.Z.; project administration, Z.Z. and Y.Z.; All authors have read and agreed to the published version of the manuscript.

Funding
This research was funded by the Project of Development and Reform Commission of Jiangxi Province(2022C046-3).

Availability of data and materials
The data presented in this study are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate
DRISHTI-GS is a public dataset, where all images were collected at Aravind eye hospital, Madurai from visitors to the hospital, with their consent. Galucoma patient selection was done by clinical investigators based on clinical findings

during examination. REFUGE is a public Retinal Fundus Glaucoma Challenge held in conjunction with MICCAI 2018 (https://refuge.grand-challenge.org/), which publicly released a data set of 1200 fundus images with ground truth segmentations and clinical glaucoma labels. All the experimental protocol was established according to the ethical guidelines and permission.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 December 2023 Accepted: 28 February 2024
Published online: 04 March 2024

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology. 2014;121(11):2081–90.
2. Liu S, Zhao H, Huang L, Ma C, Wang Q, Liu L. Vascular features around the optic disc in familial exudative vitreoretinopathy: findings and their relationship to disease severity. BMC Ophthalmol. 2023;23(1):1–11.
3. Chauhan BC, Burgoyne CF. From clinical examination of the optic disc to clinical assessment of the optic nerve head: a paradigm change. Am J Ophthalmol. 2013;156(2):218–27.
4. Drance S, Anderson DR, Schulzer M, Collaborative Normal-Tension Glaucoma Study Group, et al. Risk factors for progression of visual field abnormalities in normal-tension glaucoma. Am J Ophthalmol. 2001;131(6):699–708.
5. Hung KH, Kao YC, Tang YH, Chen YT, Wang CH, Wang YC, et al. Application of a deep learning system in glaucoma screening and further classification with colour fundus photographs: a case control study. BMC Ophthalmol. 2022;22(1):1–12.
6. Somfai GM, Tátrai E, Laurik L, Varga BE, Ölvedy V, Smiddy WE, et al. Fractal-based analysis of optical coherence tomography data to quantify retinal tissue damage. BMC Bioinformatics. 2014;15(1):1–10.
7. Lim AB, Park JH, Jung JH, Yoo C, Kim YY. Characteristics of diffuse retinal nerve fiber layer defects in red-free photographs as observed in optical coherence tomography en face images. BMC Ophthalmol. 2020;20:1–7.
8. Jiang Y, Ma Z, Wu C, Zhang Z, Yan W. RSAP-Net: joint optic disc and cup segmentation with a residual spatial attention path module and MSRCR-PT pre-processing algorithm. BMC Bioinformatics. 2022;23(1):523.
9. Lin CL, Wu KC. Development of revised ResNet-50 for diabetic retinopathy detection. BMC Bioinformatics. 2023;24(1):1–18.
10. Thakur N, Juneja M. Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. Biomed Signal Process Control. 2018;42:162–89.
11. Cho BH, Lee DY, Park KA, Oh SY, Moon JH, Lee GI, et al. Computer-aided recognition of myopic tilted optic disc using deep learning algorithms in fundus photography. BMC Ophthalmol. 2020;20(1):1–9.
12. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Munich: Springer; 2015. p. 234–41.
13. Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans Med Imaging. 2018;37(7):1597–605.
14. Jiang Y, Duan L, Cheng J, Gu Z, Xia H, Fu H, et al. JointRCNN: a region-based convolutional neural network for optic disc and cup segmentation. IEEE Trans Biomed Eng. 2019;67(2):335–43.
15. Yin P, Wu Q, Xu Y, Min H, Yang M, Zhang Y, et al. PM-Net: Pyramid multi-label network for joint optic disc and cup segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer; 2019. pp. 129–37.
16. Yin P, Yuan R, Cheng Y, Wu Q. Deep guidance network for biomedical image segmentation. IEEE Access. 2020;8:116106–16.
17. Wang S, Yu L, Yang X, Fu CW, Heng PA. Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans Med Imaging. 2019;38(11):2485–95.
18. Wang S, Yu L, Li K, Yang X, Fu CW, Heng PA. Boundary and entropy-driven adversarial learning for fundus image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Shenzhen: Springer; 2019. p. 102–10.
19. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Process Mag. 2018;35(1):53–65.
20. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Strasbourg: Springer; 2021. p. 36–46.
21. Pan S, Liu X, Xie N, Chong Y. EG-TransUNet: a transformer-based U-Net with enhanced and guided models for biomedical image segmentation. BMC Bioinformatics. 2023;24(1):85.
22. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. arXiv preprint arXiv:201011929.
23. Su Y, Deng J, Sun R, Lin G, Su H, Wu Q. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. IEEE Trans Multimed. 2023;26:313–25.
24. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE; 2017. p. 7794–803.
25. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Zurich: Springer; 2014. p. 740–55.
26. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vis. 2010;88:303–38.
27. Liu B, Pan D, Song H. Joint optic disc and cup segmentation based on densely connected depthwise separable convolution deep network. BMC Med Imaging. 2021;21:1–12.
28. Jiao S, Li X, Lu X. An improved Ostu method for image segmentation. In: 2006 8th international Conference on Signal Processing. vol. 2. Guilin: IEEE; 2007.
29. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci. 1984;10(2–3):191–203.
30. Beasley D, Huggins L, Monke A. ANSWERS: A model for watershed planning. Trans ASAE. 1980;23(4):938–0944.
31. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905.
32. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway Township: IEEE; 2015. pp. 3431–40.
33. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. International Conference on Learning Representations. Banff Canada: ICLR; 2014
34. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell. 2017;40(4):834–48.
35. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. 2017. https://doi.org/10.48550/arXiv.1706.05587.
36. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE; 2017. p. 2117–25.
37. Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE; 2017. p. 1925–34.
38. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95.

39. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). Salt Lake City: IEEE; 2018. p. 801–18.
40. Ding H, Jiang X, Liu AQ, Thalmann NM, Wang G. Boundary-aware feature propagation for scene segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE; 2019. p. 6819–29.
41. Yuan Y, Xie J, Chen X, Wang J. Segfix: Model-agnostic boundary refinement for segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Glasgow: Springer; 2020. p. 489–506.
42. Sun G, Wang W, Dai J, Van Gool L. Mining cross-image semantics for weakly supervised semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Glasgow: Springer; 2020. p. 347–65.
43. Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, et al. Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European conference on computer vision (ECCV). Munich: Springer. 2018. p. 267–83.
44. Zhou T, Li L, Li X, Feng CM, Li J, Shao L. Group-wise learning for weakly supervised semantic segmentation. IEEE Trans Image Process. 2022;31:799–811.
45. Zhu Z, Xu M, Bai S, Huang T, Bai X. Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE; 2019. pp. 593–602.
46. Edupuganti VG, Chawla A, Kale A. Automatic optic disk and cup segmentation of fundus images using deep learning. In: 2018 25th IEEE international conference on image processing (ICIP). Athens: IEEE; 2018. p. 2227–31.
47. Shankaranarayana SM, Ram K, Mitra K, Sivaprakasam M. Joint optic disc and cup segmentation using fully convolutional and adversarial networks. In: Fetal, Infant and Ophthalmic Medical Image Analysis: International Workshop, FIFI 2017, and 4th International Workshop, OMIA 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 4. Québec: Springer; 2017. p. 168–76.
48. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Granada: Springer; 2018. p. 3–11.
49. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). Barcelona: IEEE; 2020. p. 1055–9.
50. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imaging. 2018;37(12):2663–74.
51. Liu X, Guo Z, Cao J, Tang J. MDC-net: A new convolutional neural network for nucleus segmentation in histopathology images with distance maps and contour information. Comput Biol Med. 2021;135:104543.
52. Zhu F, Gao Z, Zhao C, Zhu Z, Tang J, Liu Y, et al. Semantic segmentation using deep learning to extract total extraocular muscles and optic nerve from orbital computed tomography images. Optik. 2021;244:167551.
53. Qader SM, Hassan BA, Rashid TA. An improved deep convolutional neural network by using hybrid optimization algorithms to detect and classify brain tumor using augmented MRI images. Multimed Tools Appl. 2022;81(30):44059–86.
54. Meena G, Mohbey KK, Kumar S, Lokesh K. A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. Decis Anal J. 2023;7:100243.
55. Cheng B, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. Adv Neural Inf Process Syst. 2021;34:17864–75.
56. Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE; 2021. p. 7262–72.
57. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Montreal: IEEE; 2021. p. 6881–90.
58. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE; 2021. p. 10012–22.
59. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision. Glasgow: Springer; 2020. p. 213–29.
60. Chen J, Lu Y, Yu Q, Luo X, Zhou Y. Transunet: transformers make strong encoders for medical image segmentation. 2021. https://doi.org/10.48550/arXiv.2102.04306.
61. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Trans Emerg Top Comput Intell. 2024;8:55–68.
62. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE Trans Instrum Meas. 2022;71:1–15.
63. Liu X, Zhang D, Yao J, Tang J. Transformer and convolutional based dual branch network for retinal vessel segmentation in OCTA images. Biomed Signal Process Control. 2023;83:104604.
64. Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. In: European conference on computer vision. Amsterdam: Springer; 2016. p. 577–93.
65. Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. Amsterdam: Springer; 2016. p. 69–84.
66. Wei C, Xie L, Ren X, Xia Y, Su C, Liu J, et al. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp. 1910–9.
67. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. Vancouver: Int Conf Learn Representations (ICLR); 2018.
68. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle. 2020. p. 9729–38.
69. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. Cancun: PMLR; 2019. p. 1597–607.
70. Xu D, Xiao J, Zhao Z, Shao J, Xie D, Zhuang Y. Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seoul: IEEE; 2019. p. 10334–43.
71. Benaim S, Ephrat A, Lang O, Mosseri I, Freeman WT, Rubinstein M, et al. Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE; 2020. p. 9922–31.
72. Wang J, Jiao J, Liu YH. Self-supervised video representation learning by pace prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Glasgow: Springer; 2020. p. 504–21.
73. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE; 2016. p. 770–8.
74. Ba JL, Kiros JR, Hinton GE. Layer normalization. 2016. https://doi.org/10.48550/arXiv.1607.06450.
75. Li S, Cao Q, Liu L, Yang K, Liu S, Hou J, et al. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE; 2021. p. 13668–77.
76. Jenni S, Jin H, Favaro P. Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE; 2020. p. 6408–17.

77. Xie S, Gu J, Guo D, Qi CR, Guibas LJ, Litany O. Pointcontrast: unsupervised pretraining for 3d point cloud understanding. In: Computer Vision–ECCV 2020. Glasgow: Springer; 2020.
78. Hassani K, Khasahmadi AH. Contrastive multi-view representation learning on graphs. 2020. https://doi.org/10.48550/arXiv.2006.05582.
79. Oord AVD, Li Y, Vinyals O. Representation learning with contrastive predictive coding. 2018. https://doi.org/10.48550/arXiv.1807.03748.
80. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City: IEEE; 2018. p. 3733–42.
81. Chen X, Fan H, Girshick R, He K. Improved baselines with momentum contrastive learning. 2020. https://doi.org/10.48550/arXiv.2003.04297.
82. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. Honolulu: IEEE; 2017. p. 2961–9.
83. Orlando JI, Fu H, Breda JB, Van Keer K, Bathula DR, Diaz-Pinto A, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal. 2020;59:101570.
84. Sivaswamy J, Krishnadas S, Joshi GD, Jain M, Tabish AUS, Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). IEEE; 2014. pp. 53–6.
85. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, et al. Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans Med Imaging. 2019;38(10):2281–92.
86. Sevastopolsky A, Drapak S, Kiselev K, Snyder BM, Keenan JD, Georgievskaya A. Stack-u-net: refinement network for image segmentation on the example of optic disc and cup. 2018. https://doi.org/10.48550/arXiv.1804.11294.
87. Luo L, Xue D, Pan F, Feng X. Joint optic disc and optic cup segmentation based on boundary prior and adversarial learning. Int J Comput Assist Radiol Surg. 2021;16(6):905–14.
88. Yu L. Joint segmentation of optic cup and optic disc using deep convolutional generative adversarial network. In: Journal of physics: conference series. vol. 2234. IOP Publishing; 2022. p. 012008.

Publisher's Note