JointRCNN: A Region-based Convolutional Neural Network for Optic Disc and Cup Segmentation

Yuming Jiang, Lixin Duan, Jun Cheng, Zaiwang Gu, Hu Xia, Huazhu Fu, Changsheng Li and Jiang Liu

Abstract*—*Objective: **The purpose of this paper is to propose a novel algorithm for joint optic disc and cup segmentation, which aids the glaucoma detection.** ***Methods:*** **By assuming the shapes of cup and disc regions are elliptical, we proposed an end-to-end region-based convolutional neural network for joint optic disc and cup segmentation (referred to as Joint RCNN). Atrous convolution is introduced to boost the performance of feature extraction module. In JointRCNN, disc proposal network (DPN) and cup proposal network (CPN) are proposed to generate bounding box proposals for optic disc and cup respectively. Given the prior knowledge that the optic cup is located in the optic disc, disc attention module is proposed to connect DPN and CPN, where a suitable bounding box of the optic disc is first selected and then continued to be forward propagated as the basis for optic cup detection in our proposed network. After obtaining the disc and cup regions which are the inscribed ellipses of the corresponding detected bounding boxes, the vertical cup-to-disc ratio (CDR) is computed and used as an indicator for glaucoma detection.** ***Results:*** **Comprehensive experiments clearly show that our JointRCNN model outperforms state-of-the-art methods for optic disc and cup segmentation task and glaucoma detection task.** ***Conclusion:*** **Joint optic disc and cup segmentation, which utilizes the connection between optic disc and cup, could improve the performance of optic disc and cup segmentation.** ***Significance:*** **The proposed method improves the accuracy of glaucoma detection. It is promising to be used for glaucoma screening.**

***Index Terms*—Glaucoma detection, optic disc segmentation, optic cup segmentation, convolutional neural network**

I. Introduction

Glaucoma is eye disease that would do harm to the optic nerve and cause vision loss [1]. The vision loss caused by glaucoma is permanent and cannot be cured with current treatment methods. However, glaucoma progresses silently without earlier noticeable symptoms, which makes the disease the silent thief of sight. Therefore, screening of glaucoma in an early stage is vital for timely treatment.

Intraocular pressure (IOP) assessment [2], visual field test [3] and optic nerve head (ONH) assessment [4] are three main techniques to detect glaucoma. IOP assessment measures the fluid pressure in the eyes, which is determined by Tonometry. However, the presence of glaucoma is not always accompanied by the change in IOP. The visual field test measures the range of sight when one focuses his eyes on a central point. The limitation lies in that not all hospitals have the equipment for the visual field test. Therefore, the first two methods are not suitable for glaucoma screening, while the third one, ONH assessment, has been found to be more promising clinically.

The ultimate goal of the ONH assessment is to do binary classifications between glaucomatous subjects and healthy subjects. While performing the assessment manually is costly, some automatic methods are proposed. These methods can be roughly divided into two categories: the first category of methods extracts image features for a direct binary classification [5], [6] while the second category computes some clinical indicators for glaucoma detection, such as the vertical cup-to-disc ratio (CDR), rim to disc area ratio and disc diameter [7]. The first category of methods is challenging because it is difficult to select features that could represent images properly. Moreover, the direct binary classification works as a black box without clinically explainable measurements. Thus, the second approach is more commonly used. Among these indicators, CDR is widely used in clinical assessment [8].

The optic disc is where the ganglion cell axons leave the eye. The ganglion cell axons are used to transmit the visual information of the photo-receptors to the brain. In a retinal fundus image, the optic disc can be divided into two parts: the optic cup and neuroretinal rim. The optic cup is the brightest area in the optic disc region. Both of the optic disc and the optic cup are approximated to have vertical elliptical shapes [9]. The CDR is often calculated as the ratio of vertical cup diameter to vertical disc diameter. In order to obtain an accurate measurement of CDR, we usually need to segment the optic disc and cup precisely.

Since manually segmenting optic disc and cup is time-consuming, many methods have been proposed for automatic optic disc and cup segmentation. Optic disc segmentation methods can be classified as template based methods, deformable model based methods and pixel classification based methods. In most template based methods [10]–[13], optic disc is approximated as a circle or an ellipse and thus Hough transform is utilized. For example, in [10], the Sobel method was used to detect the edge and then Hough transform was

This work is supported by Chinese Academy of Sciences (Grant No. Y61102DL03) and National Natural Science Foundation of China (Grant No. 61772118). Corresponding authors: Lixin Duan and Jun Cheng.

Y. Jiang is with Big Data Research Center at University of Electronic Science and Technology of China, Sichuan, China and also with Division of Intelligent Medical Imaging, Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Zhejiang, China (Email: yumingj80@gmail.com).

L. Duan, H. Xia and C. Li are with Big Data Research Center at University of Electronic Science and Technology of China, Sichuan, China and also with Youedata Research, Beijing, China (Email: lxduan@gmail.com, xiahu718@gmail.com, lichangsheng507@gmail.com).

J. Cheng, Z. Gu and J. Liu are with Division of Intelligent Medical Imaging, Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Zhejiang, China (Email: chengjun@nimte.ac.cn, guzaiwang@nimte.ac.cn, jimmyliu@nimte.ac.cn).

H. Fu is with Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (Email: hzfu@ieee.org).

performed to find the approximate margin of the optic disc. Aquino *et al.* [11] combined morphological and edge detection techniques with the circular Hough transform to perform optic disc segmentation. In deformable model based methods, Lowell *et al.* [14] located the optic disc firstly and then the optic disc was segmented using a deformable contour model, which employed a global elliptical model and a local deformable model. Joshi *et al.* [15] took advantage of robust multidimensional feature spaces, which are obtained from points of interest, to reduce the variations around optic disc regions. As for pixel classification based methods, Abromoff *et al.* [16] proposed a pixel classification method to segment the optic disc and cup. However, it is not easy to select the pixels and extract features to train the classifier from a large number of pixels. Cheng *et al.* [8] used the superpixel strategy to reduce the number of pixels and performed the optic disc and cup segmentation using superpixel classification. And transfer learning techniques were utilized in [17] to boost the optic disc segmentation performance in superpixel classifications.

Optic cup segmentation is more challenging than optic disc segmentation because of the obscure boundaries of optic cup. In retinal fundus images, two types of image-level information are often used to determine the optic cup boundary: pallor and vessel bends [18]. In [19], researchers used a threshold to extract the optic cup from fundus images according to intensity. Wong *et al.* [20] combined a variational level set and thresholding to segment the optic cup. All the aforementioned methods used pallor information to segment the optic cup. In [15], [21], blood vessels were found to be useful in detecting the optic cup as the vessels entering the cup are often bent. Many other methods like superpixel [8] were applied to optic cup segmentation as well. Xu *et al.* [22] addressed the superpixel classification as a low-rank superpixel representation problem.

Most of these methods treated the disc and cup segmentation as two individual segmentation problems. These methods segment the optic disc and cup separately without considering the relationship between the optic disc and cup. In [23], Zheng *et al.* performed the optic disc and cup segmentation through a graph-cut framework. They integrated some priors, such as the shapes of optic disc, the thickness of rim and the location relationship between the optic disc and cup, into their method.

The drawback of most of these methods lies in the utilization of hand-crafted features of optic disc and optic cup to obtain segmentation results, as good hand-crafted features are difficult to design and often cannot achieve good classification performance compared to the learned features in deep neural networks. Since deep learning showed its success on many computer vision tasks [24], it has been introduced into biomedical image analysis as well [25], [26]. Ronneberger *et al.* [27] proposed an architecture called U-Net, which has been widely used in biomedical image segmentation. Sevastopolsky *et al.* [28] employed U-Net in optic disc and cup segmentation. The segmentation results of the optic disc are used for further optic cup segmentation. Al-Bande *et al.* [29] adapted DenseNet into a U-Net shaped architecture to perform optic disc and cup segmentation. Shankaranarayana *et al.* [30] employed fully convolutional network to make pixel-wise prediction on retinal fundus images to generate optic disc and cup segmentation where adversarial training technique was used. Fu *et al.* [31] proposed M-Net framework, which is modified from U-Net framework, for joint optic disc and cup segmentation. Although the M-Net performed the optic disc and cup segmentation jointly, the segmentation problems are treated as a multi-label problem and the two segmentation problems are still considered as two independent problems without making use of some prior information, e.g., the optic cup is located within the optic disc. Moreover, the M-Net does not make full use of the prior knowledge that the optic disc and cup are approximately ellipses. Previously, we have also proposed to use Faster RCNN [32] for optic disc and cup segmentation, where two Faster RCNNs were employed for optic disc segmentation and optic cup segmentation separately. However, the method proposed in [32] treats the two segmentation problems as two independent ones.

In this paper, we focus the joint optic disc and cup segmentation problem on the following aspects:

1) **Object detection:** Most existing methods make pixelwise predictions to perform segmentation [28], [30], [31]. In the literature [9], the shapes of disc and cup are often approximated to be ellipses. Although such an approximation may lead to the loss of some local change in the optic disc or optic cup morphology, it often helps to remove noise. Moreover, the approximation does not affect the computation of the vertical CDR much. Inspired by the above discussions, we formulate the optic disc and optic cup segmentation problems as object detection problems, i.e., we find the minimal bounding boxes of optic disc and cup. By ignoring the rotational angle, the boundaries of the optic disc and cup can be determined by computing the inscribed ellipse of bounding boxes.
2) **Joint segmentation:** Optic disc segmentation and optic cup segmentation are two separate but dependent tasks as the optic cup locates at the center part of the disc. In order to utilize such important prior information, *attention* mechanism is employed, where the corresponding area of optic disc is cropped to guide the localization of optic cup. The two segmentation tasks are accomplished in a unified framework, which is trained end-to-end. In this way, the results of optic disc and optic cup influence each other positively.

In this paper, instead of directly segmenting the optic disc and cup, we propose to detect the minimum bounding boxes of the optic disc and cup. The boundaries of the optic disc and cup are then estimated as the inscribed ellipses within the boxes. We propose a region-based convolutional neural network for optic disc and cup segmentation, named as JointRCNN. Our method is an object detection based method. Many deep learning based methods, such as Single Shot MultiBox Detector (SSD) [33], You Only Look Once (YOLO) [34] and Faster RCNN [35], have been proposed for object detection problem. In this paper, we propose to chain two Faster RCNNs to detect the bounding boxes of the optic disc and cup. The main contributions of this work are as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2019.2913211, IEEE Transactions on Biomedical Engineering

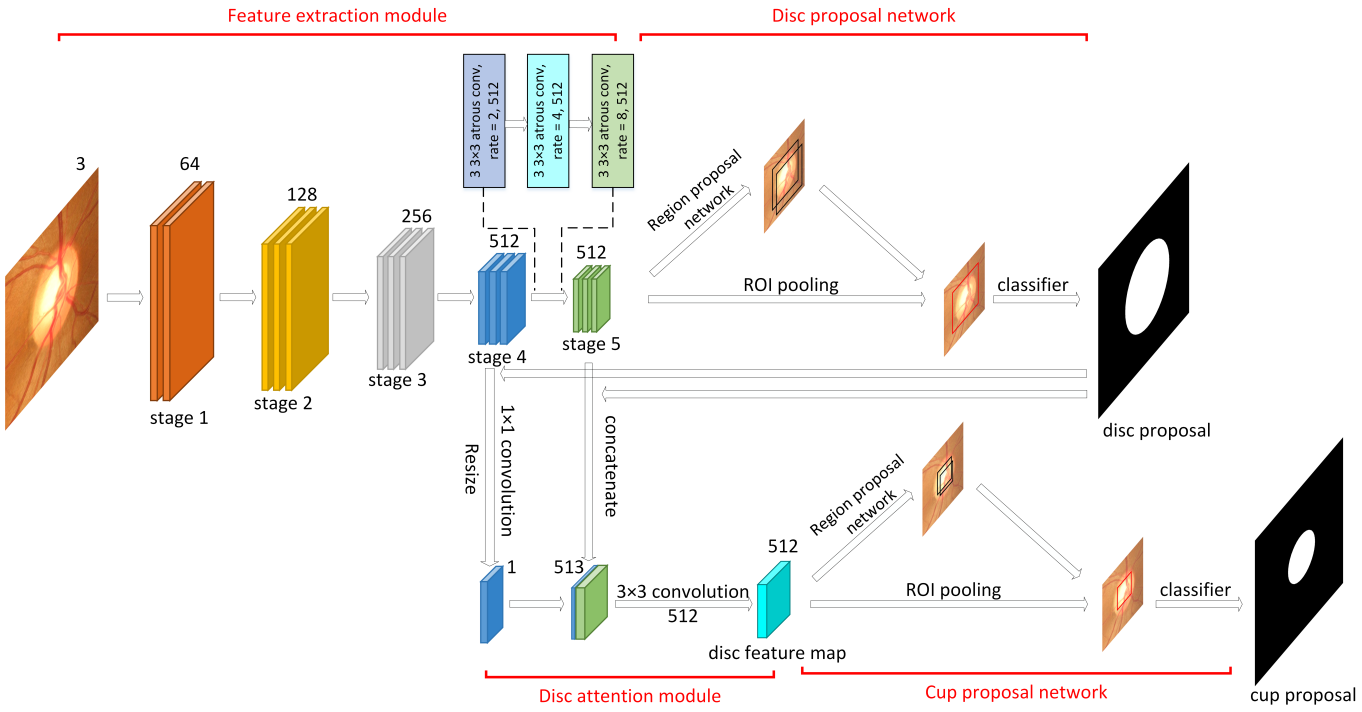


Fig. 1. Illustration of our JointRCNN method. Firstly, the images are fed into a feature extraction module, named Atrous-VGG16 framework. The backbone of Atrous-VGG16 is VGG16 [36]. The last three convolution layers of VGG16 are replaced with 9 atrous convolution layers, which enlarges the receptive field while keeping the size of feature map. We refer to feature maps with the same size as a stage. Between two adjacent stages, the max pooling is employed to reduce the size of feature map. The feature maps are further sent into following tasks. Disc proposal network and cup proposal network are proposed for optic disc and optic cup segmentation respectively. In disc/cup proposal networks, feature maps are fed into region proposal network. After that, the ROI pooling mechanism is used to crop feature maps according to the coordinates of candidate bounding boxes. Finally, cropped feature maps are fed into the classifier, where the final predictions are made. For disc proposal network, it predicts the candidate disc region proposal with the highest confidence. Then, disc attention module is proposed to crop the corresponding disc feature maps for cup proposal network. Disc attention module fuses different feature maps from different stages. In order to concatenate the cropped feature maps from stage 4 with that from stage 5, the cropped feature maps from stage 4 should be resized to the size of cropped feature maps from stage 5. After that, a 1×1 convolution is applied on cropped feature maps from stage 4. Then it was concatenated with cropped feature maps from stage 5. Finally, a 3×3 convolution is used to generate the disc feature map.

1) We relax the original segmentation task to an object detection problem, which aims to find the optic disc and cup bounding boxes. Segmentation of the optic disc and cup can be derived from computing the inscribed ellipses of bounding boxes.
2) We propose an end-to-end deep learning framework named JointRCNN, where feature maps are shared for different tasks (i.e., optic disc segmentation and optic cup segmentation) and *attention* mechanisms are employed to improve the performance. The proposed framework produces the output of the optic disc segmentation and the optic cup segmentation together, and the results will influence each other.
3) Our proposed method outperforms the state-of-the-art segmentation methods, by achieving an average overlapping error of 6.3% and 20.9% for optic disc and optic cup segmentation, respectively. We also achieve better glaucoma diagnosis results on ORIGA dataset and SCES dataset.

The remainder of this paper is organized as follows. Section II introduces the proposed JointRCNN model in detail. Section III presents our experimental results and discussions. And finally, in Section IV, we draw some conclusions.

II. Method

Our proposed JointRCNN consists of four major parts: feature extraction module, disc proposal network (DPN), disc attention module and cup proposal network (CPN). Feature extraction module is made up of deep convolutional layers, which is used to extract feature representations for original images. In feature extraction module, we introduce atrous convolution to extract more dense features [37]. The final output feature maps of this module are shared for the following tasks. The DPN is designed for optic disc segmentation, while the CPN is for optic cup segmentation. The DPN and CPN have similar designs, and they generate candidate bounding boxes for optic disc and optic cup respectively. Between DPN and CPN, the *attention* mechanism is utilized in the disc attention module to tell the CPN where the optic disc locates according to the output of DPN. The overall architecture is shown in Fig. 1.

A. Feature Extraction Module

The feature extraction module is the first step of our JointRCNN model. The feature extraction module consists of convolutional layers, pooling layers, and atrous convolution. This module generates feature maps for the segmentation tasks.

1) **Atrous convolution**: In semantic segmentation tasks and object detection tasks, deep convolutional layers have proved effective in extracting feature representations for images. However, the pooling layers generate some low resolution feature maps, which leads to the loss of semantic information in images. In order to overcome this limitation, atrous convolution is proposed for dense segmentation [37]. In this paper, we introduce atrous convolution to improve the accuracy of the bounding box detections. Mathematically, atrous convolution can be written as follows:

$$y[i] = \sum_{i=1}^{k} x[i + r \cdot k] w[k], \tag{1}$$

where y denotes the output of an atrous convolution layer, x denotes the input, w denotes the weights of filter and r denotes the rate of atrous convolution, which determines how densely the convolution layers extract features.

The atrous convolution works just like inserting holes into original inputs. Atrous convolution with a larger rate means more inserted holes. When the rate is 1, the atrous convolution is the same as the standard convolution. By atrous convolution, we can extract deeper feature maps of high resolution without increasing parameters. Besides, the atrous convolution can enlarge the field-of-views.

2) **Atrous-VGG16**: In this paper, we employ the atrous convolution in the feature extraction module, as shown in Fig. 1. We use VGG16 [36] as the backbone in this work, as it has shown good performance in semantic segmentation and object detection tasks to generate feature maps. In VGG16, input images are firstly fed into two 3×3 standard convolutions with a channel size of 64. After a max pooling layer with the stride of 2, three 3×3 standard convolutions with a channel size of 128 are used. Same design (a pooling layer with the stride of 2 + three 3×3 standard convolutions) are stacked three times, but the channels of these three modules are 256, 512 and 512, respectively.

In order to enlarge the field-of-views with keeping the feature maps in high resolution, we replace the last three standard convolutions with atrous convolutions. The design is as follows: three 3×3 atrous convolutions with the rate of 2, three 3×3 atrous convolutions with the rate of 4, three 3×3 atrous convolutions with the rate of 8. All of these atrous convolutions have a channel size of 512.

B. Disc/Cup Proposal Network

DPN and CPN are proposed to detect optic disc and optic cup in retinal fundus images respectively. Faster RCNN [35] is a well acknowledged deep learning based object detection model, which predicts bounding boxes of objects in images. Faster RCNN can be used to detect optic disc and cup in retinal fundus images directly. However, in order to make use of the prior knowledge that the optic cup is within the optic disc, we propose to detect the optic disc and optic cup jointly in an end-to-end framework, in which DPN and CPN are employed and a disc attention module is proposed to integrate the prior knowledge into the framework.

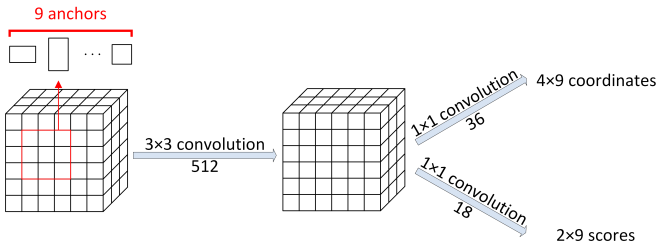


Fig. 2. Region Proposal Network (RPN): RPN is employed to generate proposal bounding boxes for optic disc and optic cup. For each location, 9 anchors will be generated as references. Generating bounding boxes is implemented as follows: the feature map is fed into a 3×3 convolutional layer with 512 channels, and then 1×1 convolutional layers are used to generate coordinates and scores for bounding boxes.

1) **Region proposal network (RPN)**: Feature maps from the feature extraction module are fed into RPN, which outputs a number of candidate region proposals. With each proposal, there is a score to indicate the probability that the box contains the optic disc/cup.

The architecture of RPN is shown in Fig. 2. In order to generate proposals accurately, anchors are introduced as bounding box regression references. At each location on feature maps, 9 anchors are generated (3 scales $\times$ 3 aspect ratios). Sliding windows are employed to generate anchors. Suppose that the size of the feature map is $W \times H$, the number of anchors for this feature map is $9WH$.

A set of candidate bounding boxes is generated by fully convolutional layers. The number of candidate bounding boxes is the same as the number of anchors. For each location, 9 candidate bounding boxes are generated. There is a score for every candidate bounding box which indicates if the box includes an object. It should be noted that the outputs of fully convolutional network are not coordinates of the bounding boxes directly but are encoded forms of coordinates as follows [38]:

$$\begin{aligned} t_x = \frac{(x - x_a)}{w_a}, t_y = \frac{(y - y_a)}{h_a}, \\ t_w = \log\left(\frac{w}{w_a}\right), t_h = \log\left(\frac{h}{h_a}\right), \end{aligned} \tag{2}$$

where x, y, w, h denote the bounding box's center coordinate, width, height, respectively. x_a, y_a, w_a, h_a denote the anchor's center coordinate, width, height, respectively.

Different from common tasks in computer vision, we aim at finding only one optic disc and one optic cup in each retinal image. Therefore, only a smaller number of fully convolutional network outputs are selected as disc region proposals according to the objectness scores. Non-maximum suppression (NMS) is applied for selection. Besides, the selection operation also reduces the computational cost.

2) **Region of interest (ROI) pooling**: The RPN generates some candidate bounding boxes. After that, ROI pooling is applied. ROI pooling [39] technique is to crop small regions of feature maps according to the coordinates of those candidate bounding boxes. We refer to the outputs of ROI pooling as RoIs. The ROI pooling utilizes the *attention* mechanism.

The RoIs are a set of cropped feature maps, which tell the classifiers where to look. The sizes of all cropped feature maps are the same. The procedures of ROI pooling are as follows:

- The coordinates of candidate bounding boxes are at image level. So we need to transform them in the size of feature maps by dividing a ratio.
- The region to be pooled is divided into $k \times k$ small blocks equally, where k denotes the desired size of output.
- Max pooling technique is applied on each small block. In one small block, only the maximum value is pooled out. Thus, the pooled feature maps are with the size of $k \times k$.

The RoIs for optic disc and optic cup are pooled from feature maps based on the outputs of RPNs in DPN and CPN respectively.

3) **Classifier**: The classifiers are deep convolutional layers, which generate the encoded forms of bounding box's coordinates (see in (2)) and probability of having the target objects. Suppose that the number of RoIs is *N*, the outputs of deep convolutional layer in classifier are 4*N* encoded coordinates and 2*N* scores. Softmax is applied on the scores to generate probability. The final prediction of bounding boxes is calculated according to the relationship (see in (2)) between coordinates of candidate bounding boxes generated from RPN and encoded coordinates.

C. Disc Attention Module

The architecture of disc attention module is shown in Fig. 1. The disc attention module is employed to chain two RCNN based modules (i.e., DPN and CPN). Inspired by the location relationship between the optic disc and the optic cup, we detect optic disc and optic cup in two stages. The first stage is to find the bounding box of the optic disc, then disc attention module is applied to crop the corresponding region of the optic disc in feature maps.

Fig. 1 shows some feature maps in different stages. We refer to stage as feature maps with the same sizes. In order to combine global and local information in retinal fundus images, we crop corresponding disc regions from different stages and then combine them. As is commonly discussed in [40], [41], feature maps with larger size (i.e, in lower stage) have more local information and feature maps with smaller size (i.e, in higher stage) have more global information. Therefore, combining large feature maps and small feature maps is useful in object detection and semantic segmentation.

In our disc attention module, we crop disc regions from stages 4 and 5. After cropping the disc region from stage 4, it is firstly resized so that it has the same shape with the cropped features from stage 5. Then, 1×1 convolution with the channel size of 1 is applied to the cropped feature map. This feature map is then concatenated with the cropped disc region from stage 5. Finally, in order to keep the channel same, a 3×3 convolution with the channel size of 512 is applied to the concatenated feature map. The final output feature map is the input of CPN.

D. Loss Function

As illustrated in Fig. 1, we need to train DPN and CPN together. The loss function of JointRCNN consists of two parts, which is defined as follows:

$$L = L_{\text{disc}} + L_{\text{cup}},$$

where L_{disc} and L_{cup} denote the losses for DPN and CPN, respectively.

In DPN, there is an RPN and a classifier to be trained. The L_{disc} is composed of two components:

$$L_{\text{disc}} = L_{\text{rpn}} + L_{\text{cls}}, \tag{3}$$

where L_{rpn} and L_{cls} denote the losses for RPN and the classifier, respectively. And the L_{cup} has a similar form like (3).

To train the RPN for DPN, we assign a binary label to each anchor, which indicates whether the anchor contains optic disc or not. The binary label is assigned according to the intersection-over-union (IOU) overlap between the anchor and ground truth (optic disc bounding boxes). An anchor with an IOU overlap higher than a threshold is labeled as positive. We set the threshold as 0.7 in this paper. For bounding box regression, the encoded forms of the coordinates of ground truth t_x^*, t_y^*, t_w^*, t_h^* are computed using (2). The smooth L_1 loss in [42] is used for computing the loss between t_x, t_y, t_w, t_h and t_x^*, t_y^*, t_w^*, t_h^* respectively, which is defined as follows:

$$L_{\text{smooth}}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1. \\ |x| - 0.5, & \text{otherwise.} \end{cases}$$

The loss function for RPN is as follows:

$$\begin{aligned} L_{\text{rpn}} = & \frac{1}{N_{\text{cls}}} \sum_i L_{\text{bce}}(p_i, p_i^*) \\ & + \frac{1}{N_{\text{reg}}} \sum_i p_i^* \cdot L_{\text{smooth}}(t_i - t_i^*), \end{aligned} \tag{4}$$

where i denotes the index of an anchor, p_i denotes the predicted probability, p_i^* denotes the binary label of each anchor. t_i is a 4-dimensional vector representing the encoded forms of coordinates of the i-th predicted bounding box t_x, t_y, t_w, t_h. And t_i^* is that for ground truth. L_{bce} is the binary cross entropy loss. N_{cls} and N_{reg} are normalization terms for L_{bce} and L_{smooth}, respectively.

The loss for classifier L_{cls} has a similar form as L_{rpn} (4). For the classifier's training, each RoI is assigned with a binary label according to the IOU overlap with disc ground truth, whose threshold is 0.5 empirically. The t_x^*, t_y^*, t_w^*, t_h^* for bounding box regression loss are computed using (2) based on the candidate bounding boxes generated from RPN and ground truth.

E. Joint Optic Disc and Cup Segmentation

Our JointRCNN outputs the bounding boxes of both the optic disc and the optic cup. The bounding boxes of the optic disc and cup we get from the framework is determined by two

points: the top left point (x_0, y_0) and the bottom right point (x_1,y_1). According to these two points, we can get the center point (x_c, y_c), the long axis a and the short axis b of inscribed ellipses as follows:

$$x_c = \frac{x_0 + x_1}{2},\ y_c = \frac{y_0 + y_1}{2},\ a = \frac{x_1 - x_0}{2},\ b = \frac{y_1 - y_0}{2}.$$

Therefore, the segmentation result of the optic disc and cup can be expressed as the standard elliptic function.

III. Experiment

In this section, the datasets are first introduced. Some implementation details and evaluation metrics are then stated as well. Finally, experimental results are given and discussed.

A. Datasets

In this paper, we conduct experiments on the ORIGA [43] and the Singapore Chinese Eye Study (SCES) datasets [44].

1) **ORIGA**: This dataset contains 650 retinal fundus images from 482 healthy eyes and 168 glaucoma patient eyes with image resolution of 3072×2048. Every image comes with an optic disc mask and an optic cup mask, CDR and the gold ground truth for glaucoma diagnosis. All of the optic disc and optic cup masks are manually annotated by ophthalmologists and graders who are trained for this task. And the gold ground truth for glaucoma diagnosis is diagnostic results by ophthalmologists who followed the world glaucoma diagnosis criteria. The 650 images have been divided into 2 sets: *Set A* for training and *Set B* for testing [45]. In this paper, we follow the same partition of the dataset to train and test our models.

2) **SCES**: It consists of 1676 images with each from a single subject. Among 1676 images, 46 images are from glaucoma patients. The diagnosis of the 1676 subjects has been given and we use this dataset to evaluate the glaucoma screening performance.

3) **Image preprocessing and data augmentation**: In order to detect the optic disc and cup in retinal fundus images based on their original resolution, we crop an 800×800 area based on the optic disc localization method proposed in [46].

Because of the limited number of training images in the ORIGA, the dataset should be properly augmented. Firstly, we move images in 8 directions with 50 pixels (i.e. up, down, right, left, top-left, top-right, bottom-left and bottom-right). Then, all images are horizontally flipped and rotated by 90 degrees, 180 degrees and 270 degrees. Therefore, each image in the original dataset is augmented to $9 \times 4 \times 2 = 72$ images.

B. Experimental Setup

Our experiments are separated in two sections: i) the optic disc and cup segmentation and ii) glaucoma screening.

1) **Optic disc and cup segmentation:** We use *Set A* in ORIGA as the training set and *set B* for testing. After data augmentation, the total number of images for training is $325 \times 72 = 23400$.

The deep convolutional layers in the feature extraction module are initialized by using parameters in the model pretrained on ImageNet. During training, we optimize the model by stochastic gradient descent. The learning rate starts from 0.001 and then drops to 0.0001 after 40,000 iterations. In the disc attention module, the disc bounding box with the highest score is selected to predict the optic cup further. For an input fundus image, the framework outputs many bounding boxes. The bounding box with the highest score is selected as the final detection result.

To evaluate the segmentation results of the optic disc and cup, we compute the commonly used overlapping error as follows:

$$E = 1 - \frac{Area\,(S \cap G)}{Area\,(S \cup G)},$$

where S represents the segmented mask and G denotes the ground truth.

2) **Glaucoma screening:** In glaucoma screening, two datasets were used. For ORIGA dataset, performance is evaluated on *Set B*. And the model is trained by using data augmented *Set A* with the size of 23400. For SCES dataset, we evaluate glaucoma screening performance on the whole datasets. And the training set for glaucoma screening performance evaluation is the same as ORIGA.

The cup-to-disc ratio (CDR) is an important factor to detect glaucoma, which can be calculated as follows:

$$CDR = \frac{VCD}{VDD},$$

where VCD and VDD are the vertical cup diameter and the vertical disc diameter, respectively.

The CDR error is computed as follows:

$$\delta = |CDR_{\mathrm{GT}} - CDR|\,,$$

where CDR_{GT} denotes the CDR computed from clinical annotation.

After obtaining CDR, we compute the Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC), which report the performance of glaucoma screening. A larger AUC indicates better performance.

C. Results

1) **Optic disc and cup segmentation:** We compare our optic disc and cup performance with some other segmentation methods: relevant-vessel bends (R-Bend) method [15], active shape model (ASM) [13], superpixel based method [8], quadratic divergence regularized SVM (QDSVM) [17], low-rand superpixel representation (LRR) [22], U-Net based segmentation method [28], fully convolutional DenseNet [29], Faster RCNN [32] and M-Net proposed in [31]. Three evaluation metrics, the overlapping error of optic disc segmentation, the overlapping error of optic cup segmentation and the CDR error are compared. The comparison results are shown in Table I.

As shown in Table I, the proposed method outperforms other optic disc and cup segmentation algorithms in all three evaluation metrics, which shows the benefits of formulating the segmentation tasks as bounding box detection problems in our work. In our method, we consider the optic disc and cup as objects in retinal fundus images. By doing so, we can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2019.2913211, IEEE Transactions on Biomedical Engineering

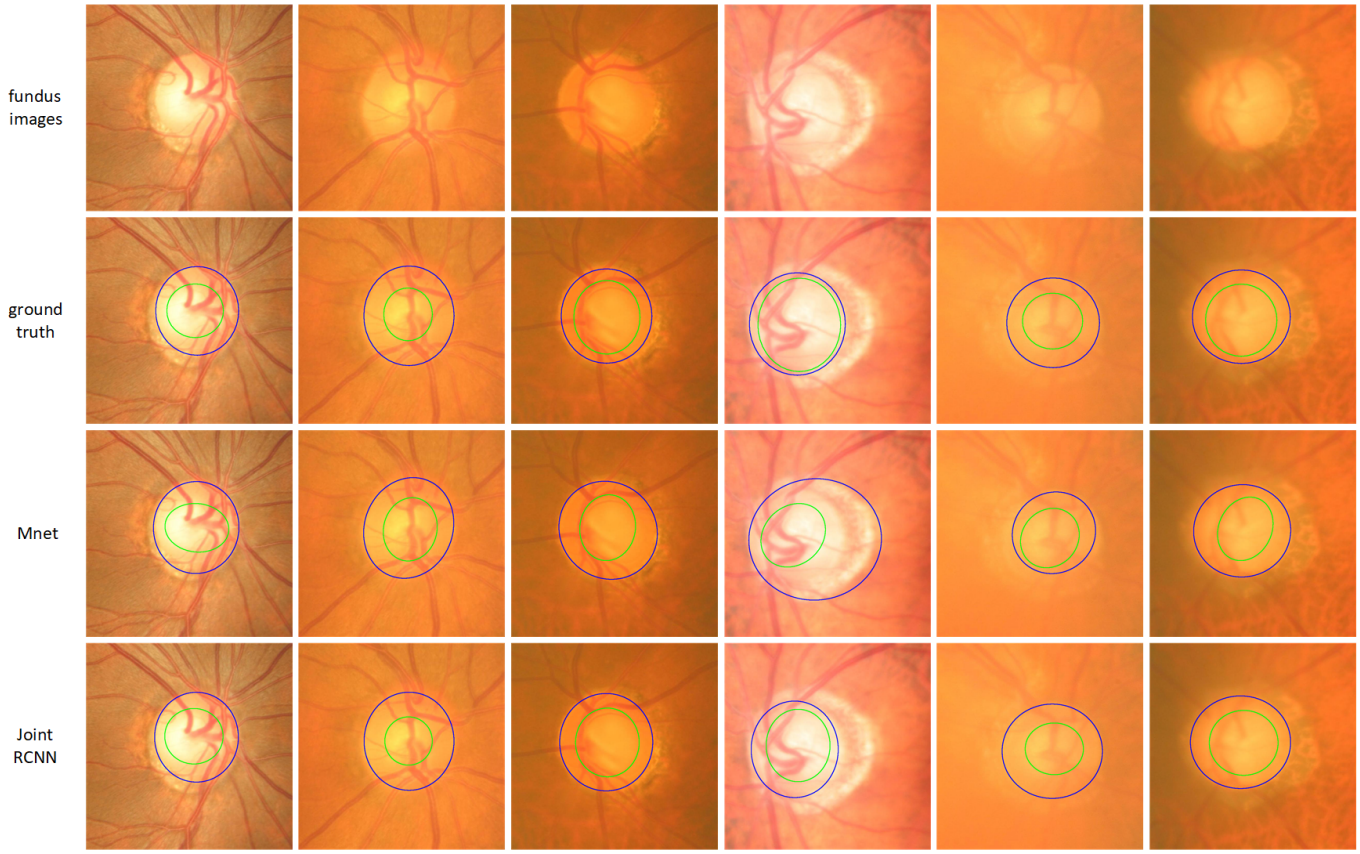


Fig. 3. Sample results. From top to bottom lines: original fundus images, ground-truth masks, state-of-the-art results obtained by M-Net [31], and our methods.

TABLE I
OVERLAPPING ERRORS AND CDR ERRORS OF DIFFERENT METHODS FOR GLAUCOMA DETECTION ON THE ORIGA DATASET.

Method	E_{disc}	E_{cup}	δ
R-Bend [15]	0.129	0.395	0.154
ASM [13]	0.148	0.313	0.107
Superpixel [8]	0.102	0.264	0.077
LRR [22]	-	0.244	-
QDSVM [17]	0.110	-	-
U-Net [27]	0.115	0.287	0.102
FC DenseNet [29]	0.067	0.231	-
Faster RCNN [32]	0.069	0.222	-
M-Net [31]	0.071	0.230	0.071
JointRCNN	**0.063**	**0.209**	**0.068**

relax the original object segmentation problem into an easier object detection problem. Besides, in other methods, some complex post-processing procedures need to be done, such as selection of the largest connected region and the ellipse fitting, which could boost performance. In our proposed method, such complicated post-processing procedures are redundant as we only need to compute the inscribed ellipse after obtaining the bounding boxes. Compared to our previous method [32], the proposed method segments optic disc and cup jointly by employing the disc attention module, which enables the results of optic disc and cup influence each other positively.

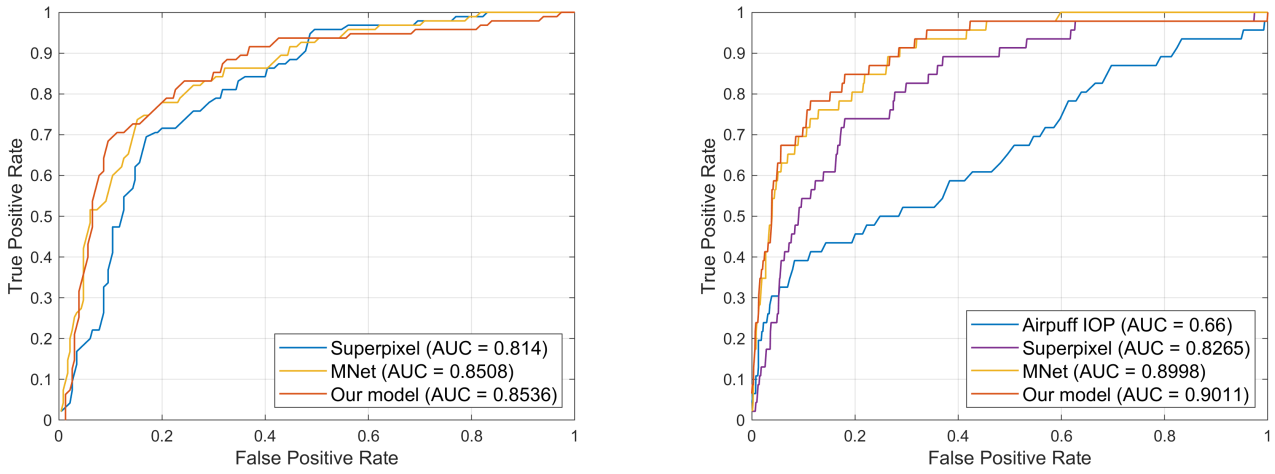
We also present six sample results in Fig. 3 to visually compare our method with M-Net. The images clearly show more accurate segmentation results of our method.

In order to justify the effectiveness of every component in our model, we conduct some ablation experiments and the results are given in Table II. To evaluate the significance level of the improvement by every component, we have also conducted student t-test.

Ablation study for chaining two RCNNs: The ablation study for chaining two RCNNs is the setting that we apply atrous convolution to the feature extraction module of Faster RCNN directly. The experimental result is shown in Table II. When we treat the optic disc and cup as two totally different objects in retinal fundus images, the performances of optic disc segmentation and optic cup segmentation drop by 0.7% and 1.0% respectively. Comparing the "JointRCNN" with "without chaining two RCNNs", we get $p < 0.001$ for disc segmentation and get $p = 0.001$ for cup segmentation, which indicates that JointRCNN significantly reduces the error in optic disc and optic cup segmentation. The ablation study that regards the optic disc and cup as two independent objects neglects the spatial connection between the optic disc and cup. According to the student t-test, chaining two RCNNs could help the segmentation of optic disc and cup. After chaining two RCNNs, the results of optic disc and cup influence each other in a positive way.

Ablation study for the disc attention module: For this ablation study, when chaining two RCNNs, we directly use

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2019.2913211, IEEE Transactions on Biomedical Engineering



(a) ORIGA dataset (b) SCES dataset

Fig. 4. ROC curves of the two datasets for glaucoma screening. Our JointRCNN method achieves the best results.

TABLE II
ABLATION STUDY FOR OUR PROPOSED JOINTRCNN METHOD.

-	E_{disc}	p-value	E_{cup}	p-value
without chaining two RCNNs	0.070	< 0.001	0.219	0.001
without disc attention module	0.066	0.028	0.215	0.052
without atrous convolution	0.064	0.709	0.226	< 0.001
JointRCNN	**0.063**	-	**0.209**	-

TABLE III
SENSITIVITY VALUES OF ALL METHODS AT DIFFERENT SPECIFICITY VALUES (I.E. s) FOR GLAUCOMA DIAGNOSIS ON THE ORIGA DATASET.

Method	$s = 0.90$	$s = 0.85$	$s = 0.80$	$s = 0.70$	$s = 0.60$	$s = 0.50$
Superpixel (AUC = 0.814) [8]	0.568	0.684	0.716	0.811	0.863	**0.947**
M-Net (AUC = 0.851) [31]	0.578	**0.726**	0.779	0.832	0.863	0.937
JointRCNN (AUC = 0.854)	**0.684**	**0.726**	**0.789**	**0.853**	**0.916**	0.937

TABLE IV
SENSITIVITY VALUES OF ALL METHODS AT DIFFERENT SPECIFICITY VALUES (I.E. s) FOR GLAUCOMA DIAGNOSIS ON THE SCES DATASET.

Method	$s = 0.90$	$s = 0.85$	$s = 0.80$	$s = 0.70$	$s = 0.60$	$s = 0.50$
Superpixel (AUC = 0.827) [8]	0.544	0.609	0.739	0.826	0.891	0.913
M-Net (AUC = 0.900) [31]	0.696	0.761	0.804	**0.913**	0.935	**0.978**
JointRCNN (AUC = 0.901)	**0.714**	**0.804**	**0.848**	**0.913**	**0.957**	**0.978**

the cropped feature maps from the output of feature extraction module. As is shown in Table II, our disc attention module, which combines the corresponding disc regions from different stages of feature maps, drops the error rates of optic disc and cup segmentation by 0.3% and 0.6% respectively. Cropping the corresponding disc region purely from one feature map causes some loss of local information, which leads to poor performance. The p-value indicates that the disc attention module significantly reduces the error in optic disc segmentation while the reduction in optic cup segmentation is not significant. This is reasonable as the disc attention module helps the cup segmentation indirectly. Therefore, the reduction in cup segmentation is less significant.

Ablation study for atrous convolution: We compare our proposed Atrous-VGG16 net with VGG16 net to conduct this ablation study. The comparison in Table II shows the effect of atrous convolution, which improves the performance of the optic cup by 1.7%. Comparing the "JointRCNN" with "without atrous convolution", we get $p = 0.709$ for disc segmentation and $p < 0.001$ for optic cup segmentation. This shows that the atrous convolution does not contribute much to the disc segmentation. This might be explained as the optic disc segmentation is already very good with the disc attention module and chaining two RCNNs. However, it significantly reduces the error in optic cup segmentation as there is still room for improvement.

2) **Glaucoma screening:** We evaluate the glaucoma diagnosis performance both on ORIGA dataset and SCES dataset. Our model is trained on the augmented training set of ORIGA dataset. We evaluate the model both on these two datasets. The ROC curves of these two datasets are shown in Fig. 4. As we can see, the proposed method achieves higher AUC value than prior methods.

Besides, we also present sensitivity values at different specificity values for these two datasets in Table III and IV, respectively. As the aim of glaucoma screening is to detect glaucoma and glaucoma suspicious from a large scale of population study where most of subjects do not have glaucoma, therefore, a high false positive rate will result in too many false alarms and make the system not practical. And clinicians are more interested to improve the sensitivity for false positive rates below 0.5. The JointRCNN gives a lower sensitivity than M-Net and superpixel classification methods in high false positive rate region because the computed CDRs for a few glaucomatous cases in ORIGA happened to be smaller. Based on the above discussion, we show 6 sample specificity values, which are all larger than 0.5 in both tables. On both datasets, our method achieves the best performance as compared to other state-of-the-art methods.

Although CDR is not a single decisive parameter for glaucoma diagnosis, it is the most common one used by many clinicians. Very often, a higher CDR indicates a higher risk of glaucoma. Clinicians often use CDR to screen for subjects with higher CDRs for following up and the changes in CDR in two visits will be used as evidence. Therefore, more accurate CDR estimation is expected to the accuracy of glaucoma screening. In addition, the CDR can also be combined with other factors to get a binary assessment. In [47], Liu *et al.* shows that the CDR can be combined with patients' personal data, medical retinal image and genome information for automatic glaucoma diagnosis and screening in a large dataset from a population study.

IV. Conclusion

In this work, we propose an end-to-end deep learning framework named JointRCNN for optic disc and cup segmentation. Based on the observation that the shape of the optic disc and cup regions are approximately ellipses, we formulate the original disc and cup segmentation problems to an object detection problem. Motivated by the inherent location relationship that the optic cup is contained in the optic disc, we cascade disc proposal network (DPN) and cup proposal network (CPN), in which the bounding boxes of optic disc and cup are detected sequentially. Between DPN and CPN, the *attention* mechanism is employed in the disc attention module to tell the network where to look for the optic cup. In order to boost the performance of the detection problem, we also introduce the atrous convolution to extract feature maps of fundus images. The proposed method achieves the state-of-the-art performance for the optic disc and cup segmentation on the ORIGA dataset, as well as the glaucoma screening on both the ORIGA and SCES datasets.

References

[1] Y.-C. Tham *et al.*, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[2] J. Baum *et al.*, "Assessment of intraocular pressure by palpation," *American journal of ophthalmology*, vol. 119, no. 5, pp. 650–651, 1995.

[3] S. Drance *et al.*, "Risk factors for progression of visual field abnormalities in normal-tension glaucoma," *American journal of ophthalmology*, vol. 131, no. 6, pp. 699–708, 2001.

[4] D. Garway-Heath and R. Hitchings, "Quantitative evaluation of the optic nerve head in early glaucoma," *British Journal of Ophthalmology*, vol. 82, no. 4, pp. 352–361, 1998.

[5] J. Meier *et al.*, "Effects of preprocessing eye fundus images on appearance based glaucoma classification," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2007, pp. 165–172.

[6] R. Bock *et al.*, "Glaucoma risk index: automated glaucoma detection from color fundus images," *Medical image analysis*, vol. 14, no. 3, pp. 471–481, 2010.

[7] A. Almazroa *et al.*, "Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey," *Journal of ophthalmology*, vol. 2015, 2015.

[8] J. Cheng *et al.*, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[9] ——, "Automatic optic disc segmentation with peripapillary atrophy elimination," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 6224–6227.

[10] X. Zhu and R. M. Rangayyan, "Detection of the optic disc in images of the retina using the hough transform," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 3546–3549.

[11] A. Aquino *et al.*, "Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques," *IEEE transactions on medical imaging*, vol. 29, no. 11, pp. 1860–1869, 2010.

[12] H. Tjandrasa *et al.*, "Optic nerve head segmentation using hough transform and active contours," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 3, pp. 531–536, 2012.

[13] F. Yin *et al.*, "Model-based optic nerve head segmentation on retinal fundus images," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 2626–2629.

[14] J. Lowell *et al.*, "Optic nerve head segmentation," *IEEE Transactions on medical Imaging*, vol. 23, no. 2, pp. 256–264, 2004.

[15] G. D. Joshi *et al.*, "Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment," *IEEE transactions on medical imaging*, vol. 30, no. 6, pp. 1192–1205, 2011.

[16] M. D. Abramoff *et al.*, "Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features," *Investigative ophthalmology & visual science*, vol. 48, no. 4, pp. 1665–1673, 2007.

[17] J. Cheng *et al.*, "Quadratic divergence regularized svm for optic disc segmentation," *Biomedical optics express*, vol. 8, no. 5, pp. 2687–2696, 2017.

[18] M. E. Brezinski *et al.*, "Imaging of coronary artery microstructure (in vitro) with optical coherence tomography," *The American journal of cardiology*, vol. 77, no. 1, pp. 92–93, 1996.

[19] G. D. Joshi *et al.*, "Optic disk and cup boundary detection using regional information," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*. IEEE, 2010, pp. 948–951.

[20] D. Wong *et al.*, "Level-set based automatic cup-to-disc ratio determination using retinal fundus images in argali," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 2266–2269.

[21] ——, "Automated detection of kinks from blood vessels for optic cup segmentation in retinal images," in *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260. International Society for Optics and Photonics, 2009, p. 72601J.

[22] Y. Xu *et al.*, "Optic cup segmentation for glaucoma detection using low-rank superpixel representation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 788–795.

[23] Y. Zheng *et al.*, "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.
[24] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[25] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2493–2501, 2018.
[26] Z. Gu *et al.*, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, 2019.
[27] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[28] A. Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network," *Pattern Recognition and Image Analysis*, vol. 27, no. 3, pp. 618–624, 2017.
[29] B. Al-Bander *et al.*, "Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis," *Symmetry*, vol. 10, no. 4, p. 87, 2018.
[30] S. M. Shankaranarayana *et al.*, "Joint optic disc and cup segmentation using fully convolutional and adversarial networks," in *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 168–176.
[31] H. Fu *et al.*, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging*, 2018.
[32] Y. Jiang *et al.*, "Optic disc and cup segmentation with blood vessel removal from fundus images for glaucoma detection," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 862–865.
[33] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
[34] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[35] S. Ren *et al.*, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[37] L. Chen *et al.*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence*, 2017.
[38] R. Girshick *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
[39] K. He *et al.*, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *european conference on computer vision*. Springer, 2014, pp. 346–361.
[40] H. Zhao *et al.*, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
[41] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *CVPR*, vol. 1, no. 2, 2017, p. 4.
[42] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[43] Z. Zhang *et al.*, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 3065–3068.
[44] M. Baskaran *et al.*, "The prevalence and types of glaucoma in an urban chinese population: the singapore chinese eye study," *JAMA ophthalmology*, vol. 133, no. 8, pp. 874–880, 2015.
[45] J. Cheng *et al.*, "Similarity regularized sparse group lasso for cup to disc ratio computation," *Biomedical optics express*, vol. 8, no. 8, pp. 3763–3777, 2017.
[46] Z. Zhang *et al.*, "Optic disc region of interest localization in fundus image for glaucoma detection in argali," in *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on*. IEEE, 2010, pp. 1686–1689.
[47] J. Liu *et al.*, "Automatic glaucoma diagnosis through medical imaging informatics," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1021–1027, 2013.