tvst

Special Issue

A Retrospective Comparison of Deep Learning to Manual Annotations for Optic Disc and Optic Cup Segmentation in Fundus Photographs

Huazhu Fu[2,*], Fei Li[1,*], Yanwu Xu[3], Jingan Liao[4], Jian Xiong[1], Jianbing Shen[2], Jiang Liu[5,6], and Xiulan Zhang[1], for iChallenge-GON study group[#]

[1] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong, China
[2] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[3] Intelligent Healthcare Unit, Baidu, Beijing, China
[4] School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China
[5] Department of Computer Science and Engineering, Southern University of Science and Technology, Guangzhou, Guangdong, China
[6] Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Ningbo, Zhejiang, China

Correspondence: Yanwu Xu, Intelligent Healthcare Unit, Baidu, Beijing 100085, China.
e-mail: ywxu@ieee.org
Xiulan Zhang, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510275, China.
e-mail: zhangxl2@mail.sysu.edu.cn

Received: January 30, 2020
Accepted: April 22, 2020
Published: June 24, 2020

Keywords: optic disc; optic cup; segmentation; artificial intelligence

Citation: Fu H, Li F, Xu Y, Liao J, Xiong J, Shen J, Liu J, Zhang X. A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photographs. Trans Vis Sci Tech. 2020;9(2):33, https://doi.org/10.1167/tvst.9.2.33

Purpose: Optic disc (OD) and optic cup (OC) segmentation are fundamental for fundus image analysis. Manual annotation is time consuming, expensive, and highly subjective, whereas an automated system is invaluable to the medical community. The aim of this study is to develop a deep learning system to segment OD and OC in fundus photographs, and evaluate how the algorithm compares against manual annotations.

Methods: A total of 1200 fundus photographs with 120 glaucoma cases were collected. The OD and OC annotations were labeled by seven licensed ophthalmologists, and glaucoma diagnoses were based on comprehensive evaluations of the subject medical records. A deep learning system for OD and OC segmentation was developed. The performances of segmentation and glaucoma discriminating based on the cup-to-disc ratio (CDR) of automated model were compared against the manual annotations.

Results: The algorithm achieved an OD dice of 0.938 (95% confidence interval [CI] = 0.934–0.941), OC dice of 0.801 (95% CI = 0.793–0.809), and CDR mean absolute error (MAE) of 0.077 (95% CI = 0.073 mean absolute error (MAE)0.082). For glaucoma discriminating based on CDR calculations, the algorithm obtained an area under receiver operator characteristic curve (AUC) of 0.948 (95% CI = 0.920 mean absolute error (MAE)0.973), with a sensitivity of 0.850 (95% CI = 0.794–0.923) and specificity of 0.853 (95% CI = 0.798–0.918).

Conclusions: We demonstrated the potential of the deep learning system to assist ophthalmologists in analyzing OD and OC segmentation and discriminating glaucoma from nonglaucoma subjects based on CDR calculations.

Translational Relevance: We investigate the segmentation of OD and OC by deep learning system compared against the manual annotations.

Introduction

Glaucoma is the leading cause of irreversible blindness around the world.[1] In clinical practice, glaucoma is diagnosed by evaluating the thickness of the retinal nerve fiber layer (RNFL), and the morphology of the optic nerve head (ONH).[2,3] Some other features are considered when making a diagnosis of glaucoma,[1,4] including visual field (VF), intraocular pressure (IOP), family history, corneal thickness, history of disc hemorrhages, etc. In fundus examinations, glaucoma is usually characterized by a larger cup-to-disc ratio (CDR), focal notching of the neuroretinal rim, etc.[5,6] An enlarged CDR may also indicate the existence of other ocular ailments, such as neuro-ophthalmic

diseases. Previous studies have shown that a larger vertical CDR is closely associated with the progression of glaucoma.[7–9] However, calculations for the CDR often vary among ophthalmologists and are relatively subjective, because they require a comprehensive judgment of shapes and structures of optic disc (OD) and optic cup (OC).[10,11] As such, several tools incorporating computer vision and machine learning techniques have been developed to perform automated OD and OC segmentation for large-scale data analysis.

Recently, deep learning techniques have been shown to perform very well in a wide variety of medical imaging tasks,[12,13] including diabetic retinopathy screening,[14–16] and age-related macular degeneration detection.[17–19] Automated glaucoma detection from fundus photographs has also received increasing attention.[20–23] However, most of the studies have focused on predicting glaucoma directly from the fundus photographs, without any visualization result. By contrast, OD and OC segmentation could be helpful for calculating the risk factors (e.g. CDR, rim-to-disc ratio[24]), and providing a segmentation visualization result. Moreover, although some automated segmentation methods appear to perform well on the small datasets,[25–28] they have not been compared to performance by the practicing ophthalmologists.

In this study, we developed a deep learning system for automated OD and OC segmentation in fundus photographs and evaluated its performances compared against seven ophthalmologists for OD and OC segmentation and glaucoma discriminating based on CDR calculations.

Methods

Data Acquisition

The fundus photographs were collected from Zhongshan Ophthalmic Center, Sun Yat-sen University, China. The fundus photographs were captured by using Zeiss Visucam 500 and Canon CR-2 machines. We included 1496 fundus photos from 748 subjects. As long as the diagnosis of both eyes are determined, both eyes of the same subjects were included. After a quality assessment, the low-quality fundus photographs (e.g. low-contrast and blurring) are excluded. Finally, a total of 1200 fundus photographs are selected in our study with 120 glaucoma and 1080 nonglaucoma cases (inclusion criteria: 1. age $\geq$ 18 years old; 2. clear images without artifacts or overexposure; and 3. definite diagnoses acquired). Diagnoses were based on the comprehensive evaluation of the subjects' medical records, including fundus photographs, IOP measurements, optical coherence tomography (OCT) images, VFs. The fundus photographs came from previous clinical studies,[29] and all the participants signed informed consent before enrollment. Institutional review board/Ethics Committee ruled that approval was not required for this study.

The dataset was split into a training set (400 photographs with 40 glaucoma cases), a validation set (400 photographs with 40 glaucoma cases, women: 52%, mean age: 25.3 $\pm$ 11.5 years), and a test set (400 photographs with 40 glaucoma cases, women: 55%, mean age: 23.7 $\pm$ 9.0 years), following the REFUGE challenge.[29] The photographs from the same patient were assigned to the same set. The training set was used to learn the algorithm parameters, the validation set was used to choose the model, and the test set was used to evaluate the algorithm, as well as the ophthalmologists.

Diagnostic Criteria for Glaucoma

Patients with glaucomatous damage in the ONH area and reproducible glaucomatous VF defects were included in our study. A glaucomatous VF defect is defined as a reproducible reduction in sensitivity compared to the normative dataset, in reliable tests, at: (1) two or more contiguous locations with P value < 0.01, (2) three or more contiguous locations with P value < 0.05. ONH damage is defined as CDR > 0.7, thinning of RNFL (an RNFL defect in the optic nerve head shown on the OCT reports), or both, without a retinal or neurological cause for VF loss. Specifically, first, the diagnostic criteria were based on the trial in glaucoma (i.e. UKGTS).[30] Second, if the points exist on the rim, there could be false-positive cases. However, as mentioned in our manuscript, the included subjects in our study received repeated VF tests to ensure reliability. If the defects exist all the time, we consider them as glaucomatous defects.

All OD and OC annotations were manually labeled by seven licensed ophthalmologists (average experience: 8 years, range: 5–10 years). All ophthalmologists independently reviewed and marked OD and OC in each photograph as the tilted ellipses using a free image labeling tool with capabilities for image review, zoom, and ellipse fitting. Ophthalmologists did not have access to any patient information or knowledge of disease prevalence in the data. The final standard reference labels of OD and OC were created by merging the annotations from multiple ophthalmologists using majority voting. Specifically, a senior specialist with $>$ 10 years of experience in glaucoma performed a quality check afterward, analyzing the resulting masks to account for potential mistakes. When errors in

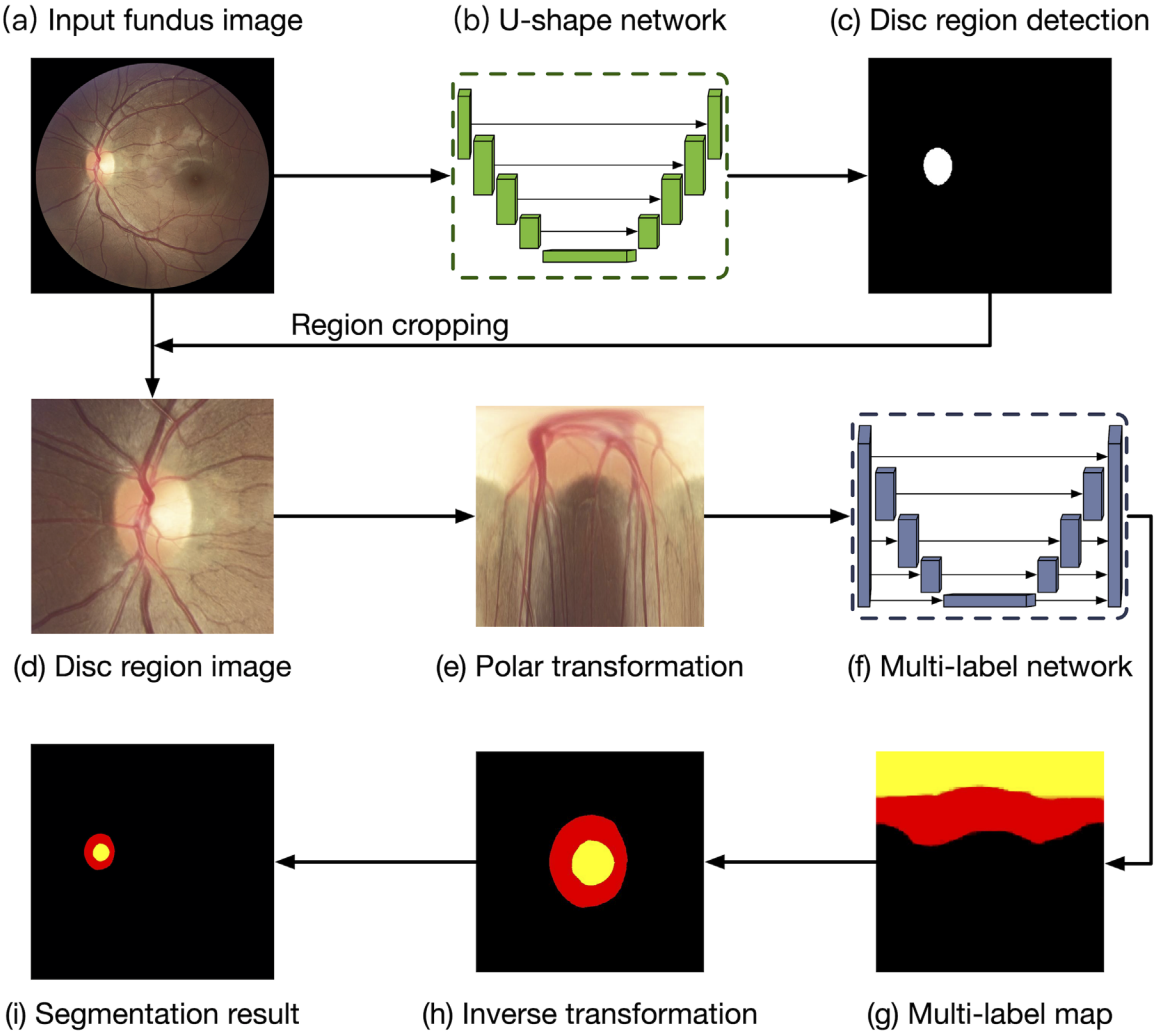

Figure 1. Deep learning system for automated segmentation in fundus images. The algorithm included two stages: optic disc (OD) region detection and OD and optic cup (OC) segmentation. For a given fundus image (**a**), the U-Net network (**b**) was utilized to detect the OD region (**c**). With the cropped OD region (**d**), a polar transformation was used to map the image into polar coordinate (**e**). A multilabel network (**f**) segmented the OD and OC jointly, and an inverse transformation returned the output map (**g**) back to original coordinates (**h**).

the annotations were observed, this additional reader analyzed each of the seven segmentations, removed those that were considered failed in his/her opinion, and repeated the majority voting process with the remaining ones. Only a few cases had to be corrected using this protocol.

Algorithm Development

In this study, we proposed a deep learning system for automated OD and OC segmentation in fundus photographs (Fig. 1). The proposed system included two main stages: (1) OD region detection, which first localized the OD center within the whole fundus photograph, and then cropped the OD region to remove the background; and (2) OD and OC segmentation, which segmented the OD and OC jointly via a multilabel deep network in the cropped OD image. We used a U-Net network for OD region detection, which was based on encoder–decoder architecture to achieve satisfactory performances in many biomedical image tasks.[31] The encoder path consisted of the multiple convolutional layers with various filter banks to produce a set of feature representations for the inputs, whereas the decoder path aggregated the feature representations to predict the probability map of the OD region in the fundus photograph. Additionally, skip connections were used to concatenate the feature representations from the encoder path to the corresponding decoder path. The final output of the U-Net network was a probability map, indicating the OD region and background for each pixel in the fundus image, as

shown as Figure 1c. The implementation details of the U-Net network for OD detection were given in Supplementary Fig. 1A. With the probability map of OD localization, we used a thresholding of 0.5 to obtain the mask for the OD region, and cropped a local image around the OD for the following OD and OC segmentation stage.

In the second stage of our algorithm, a multilabel network was utilized to segment OD and OC simultaneously in the cropped OD region image.[26] Similar to the U-Net network, the multilabel network also consisted of an encoder and a decoder path based on convolutional layers. The difference is that the multilabel network used the average pooling layers to naturally down-sample the images as multiscale inputs to the corresponding encoder path, whereas the multiscale outputs from each scale of decoder path were fused together as the final probability map. Additionally, the multilabel loss function was used to learn the binary classifier of each class (i.e. OD and OC), and assign multiple labels to each pixel for segmentation of OD and OC jointly. The implementation details of the multilabel network for OD and OC segmentation were given in Supplementary Fig. 1B. In the fundus photograph, the size ratio of the OC region is less than the OD and background, which could lead overfitting of deep model during training. To address this, we map the OD region image into the polar coordinates, before being fed into the multilabel network. Polar transformations were carried out using the OD center as the origin and the local image width as the radius (see Fig. 1e). The implementation details of polar transformations were given in Supplementary Fig. 1C. After passing through the multilabel network, an inverse polar transformation reverted the predicted map back to the original coordinates.

The U-Net network for OD detection and multilabel network for OD and OC segmentation were trained separately. The U-Net network was trained based on the whole fundus images resized to 800 by 800 pixels, with the OD reference label, whereas the multilabel network was trained based on the OD region images resized to 400 by 400 pixels, with the OD and OC reference labels. Random flips and rotations were applied to all training photographs before they were fed into the networks for data augmentation. These two networks were implemented with Python (version 3.6) based on Keras (version 2.2) with a Tensorflow (version 1.12) backend. All network parameters of the networks were optimized by using stochastic gradient descent with a learning rate of 0.0001 and a momentum of 0.9. In order to prevent the networks from overfitting, early stopping was performed, which saved the network model after each epoch and chose the final model with the lowest loss on the validation set. Each stage of training required around 2 hours for completion, on a single NVIDIA Titan XP.

Statistical Analysis and Evaluation

For segmentation evaluation, we reported three performance metrics, namely, OD dice, OC dice, and CDR mean absolute error (MAE). The dice scores measured the overlap ratio between the target regions of the reference label and segmented result, whereas the CDR MAE was the mean absolute error between the calculated CDR values from the reference label and segmented result. We also determined the SD and 95% Bayesian confidence interval (CI)[32] for each segmentation metric.

In addition to evaluating the segmentation performance, we also compared the algorithm against ophthalmologists for discriminating glaucoma from nonglaucoma photographs based on CDR calculations. The performances across different diagnostic thresholds of CDR were assessed in terms of the area under receiver operator characteristic curve (AUC). To convert the CDR to a binary prediction, we chose the highest point on the receiver operator characteristic (ROC) curve, which offers minimal trade-off between sensitivity and specificity, as the final discriminating threshold. Moreover, the 95% bootstrapping CI[33] was provided for each discriminating metric as: computing 10,000 bootstrap replicates from the set, and each metric was computed for algorithm and reference label on the same bootstrap replicate. The P values were reported by comparing the AUC with the algorithm and ophthalmologist predictions. All statistical analyses were performed using Python (version 3.6) with SciPy (version 1.2) and Scikit-learn (version 2.20). Figures were created using Matplotlib (version 3.0) and Seaborn (version 0.9).

Results

The segmentation performances of our algorithm and annotations of the seven ophthalmologists, for the test set, are listed in Table 1. For glaucoma data, the algorithm obtained an OD dice of 0.941 (SD = 0.057; 95% CI = 0.926–0.956), OC dice of 0.864 (SD = 0.089; 95% CI = 0.841–0.887), and CDR MAE of 0.065 (SD = 0.056; 95% CI = 0.051–0.080). For nonglaucoma data, the algorithm predicted an OD dice of 0.937 (SD = 0.040; 95% CI = 0.934–0.941), OC dice of 0.794 (SD = 0.096; 95% CI = 0.786–0.803), and CDR MAE of 0.079 (SD = 0.050; 95%

Table 1. Segmentation Performances of Ophthalmologists and Algorithm on Test Set

	OD Dice (SD, 95% CI)	OC Dice (SD, 95% CI)	CDR MAE (SD, 95% CI)
Glaucoma data (40 images)			
Ophthalmologist 1	0.963 (0.026, 0.957-0.970)	0.880 (0.090, 0.857-0.904)	0.057 (0.055, 0.043-0.071)
Ophthalmologist 2	0.945 (0.034, 0.936-0.954)	0.867 (0.090, 0.843-0.890)	0.057 (0.045, 0.045-0.069)
Ophthalmologist 3	0.944 (0.031, 0.936-0.952)	0.888 (0.072, 0.869-0.907)	0.039 (0.035, 0.030-0.048)
Ophthalmologist 4	0.953 (0.027, 0.946-0.960)	0.884 (0.071, 0.865-0.902)	0.052 (0.040, 0.042-0.063)
Ophthalmologist 5	0.947 (0.030, 0.939-0.955)	0.865 (0.089, 0.842-0.889)	0.055 (0.054, 0.041-0.069)
Ophthalmologist 6	0.954 (0.030, 0.946-0.962)	0.904 (0.064, 0.887-0.921)	0.048 (0.042, 0.037-0.059)
Ophthalmologist 7	0.954 (0.030, 0.947-0.962)	0.719 (0.138, 0.683-0.755)	0.149 (0.091, 0.125-0.173)
Algorithm	0.941 (0.057, 0.926-0.956)	0.864 (0.089, 0.841-0.887)	0.065 (0.056, 0.051-0.080)
Nonglaucoma data (360 images)			
Ophthalmologist 1	0.956 (0.028, 0.953-0.958)	0.686 (0.107, 0.677-0.695)	0.157 (0.071, 0.151-0.163)
Ophthalmologist 2	0.926 (0.047, 0.922-0.931)	0.842 (0.086, 0.834-0.849)	0.053 (0.040, 0.049-0.056)
Ophthalmologist 3	0.922 (0.040, 0.918-0.925)	0.837 (0.078, 0.831-0.844)	0.041 (0.032, 0.038-0.044)
Ophthalmologist 4	0.949 (0.033, 0.946-0.952)	0.801 (0.117, 0.791-0.811)	0.056 (0.043, 0.053-0.060)
Ophthalmologist 5	0.945 (0.035, 0.942-0.948)	0.870 (0.082, 0.863-0.877)	0.041 (0.036, 0.038-0.044)
Ophthalmologist 6	0.953 (0.034, 0.950-0.956)	0.903 (0.064, 0.898-0.909)	0.034 (0.030, 0.031-0.037)
Ophthalmologist 7	0.955 (0.032, 0.952-0.958)	0.664 (0.138, 0.652-0.676)	0.130 (0.063, 0.125-0.136)
Algorithm	0.937 (0.040, 0.934-0.941)	0.794 (0.096, 0.786-0.803)	0.079 (0.050, 0.074-0.083)
All data (400 images)			
Ophthalmologist 1	0.956 (0.028, 0.954-0.959)	0.705 (0.121, 0.695-0.715)	0.147 (0.075, 0.141-0.153)
Ophthalmologist 2	0.928 (0.046, 0.925-0.932)	0.844 (0.086, 0.837-0.851)	0.053 (0.040, 0.050-0.056)
Ophthalmologist 3	0.924 (0.039, 0.921-0.927)	0.843 (0.079, 0.836-0.849)	0.041 (0.032, 0.038-0.044)
Ophthalmologist 4	0.949 (0.032, 0.947-0.952)	0.809 (0.116, 0.800-0.819)	0.056 (0.042, 0.052-0.059)
Ophthalmologist 5	0.945 (0.035, 0.942-0.948)	0.870 (0.082, 0.863-0.876)	0.043 (0.038, 0.040-0.046)
Ophthalmologist 6	0.953 (0.034, 0.950-0.956)	0.903 (0.064, 0.898-0.909)	0.035 (0.032, 0.033-0.038)
Ophthalmologist 7	0.955 (0.031, 0.952-0.957)	0.670 (0.138, 0.658-0.681)	0.132 (0.066, 0.127-0.138)
Algorithm	0.938 (0.041, 0.934-0.941)	0.801 (0.097, 0.793-0.809)	0.077 (0.051, 0.073-0.082)

OD, optic disc; OC, optic cup; CDR, cup-to-disc ratio; MAE, mean absolute error; CI, confidence interval.

CI = 0.074–0.083). The segmentation performances of the algorithm on the whole test set achieved an OD dice of 0.938 (SD = 0.041; 95% CI = 0.934–0.941), OC dice of 0.801 (SD = 0.097; 95% CI = 0.793–0.809), and CDR MAE of 0.077 (SD = 0.051; 95% CI = 0.073–0.082). For OD segmentation, the algorithm performed better than ophthalmologist 2, who reported an OD dice of 0.928 (SD = 0.046; 95% CI = 0.925–0.932), and ophthalmologist 3, who determined the OD dice to be 0.924 (SD = 0.039; 95% CI = 0.921–0.927). For OC segmentation, the algorithm performed better than ophthalmologist 1, who obtained an OC dice of 0.705 (SD = 0.121; 95% CI = 0.695–0.715), and ophthalmologist 7, who got an OC dice of 0.670 (SD = 0.138; 95% CI = 0.658–0.681). The OD and OC dice scores of inter-agreement for seven ophthalmologists are given in Figure 2. Boxplots for the calculated CDRs of the reference label, the ophthalmologist annotations and the algorithm outputs, for glaucoma and nonglaucoma data on test set, were plotted in Figure 3. The average CDRs of the reference labels for the glaucoma and normal cases were 0.656 and 0.453, respectively, for the test set.

Figure 4 shows the visual results of automated OD and OC segmentation, for both glaucoma and nonglaucoma data. Several failure cases were also provided in Figures 4E, 4F. One common failure case for OD segmentation was confusion when peripapillary atrophy (PPA) was present, because this looks similar to the OD (green arrow in Fig. 4E). Failure cases also occurred due to the low-quality of the fundus photographs, where poor illumination and low-contrast often made it difficult to determine the boundary of the OC (green arrow in Fig. 4F). However, this could be relieved using additional image enhancement pre-processing.

The performances of discriminating glaucoma from nonglaucoma subjects based on CDR, for the test set,

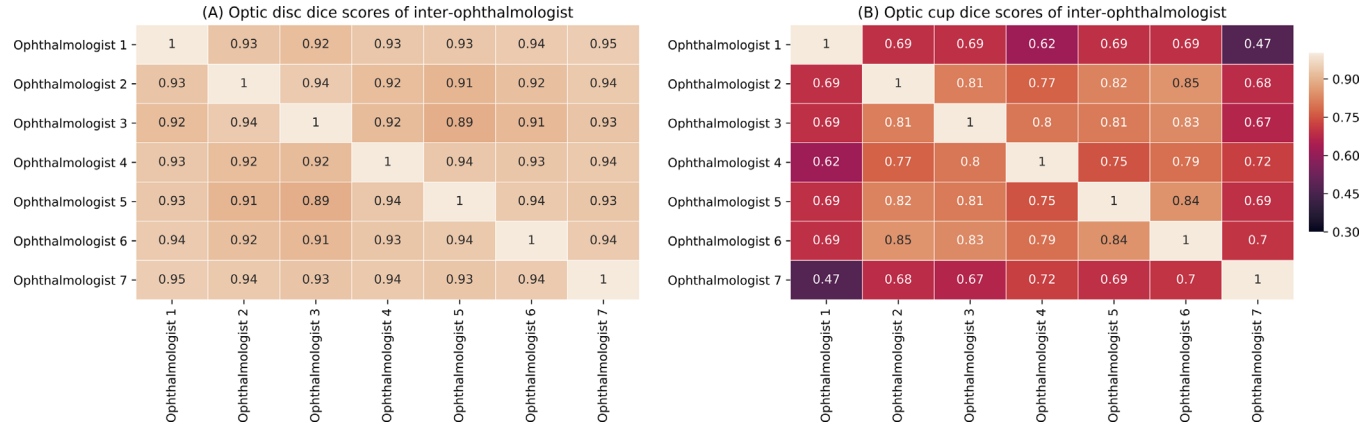

Figure 2. Dice scores of inter-agreement for seven ophthalmologists on the test set for (**A**) optic disc, and (**B**) optic cup.

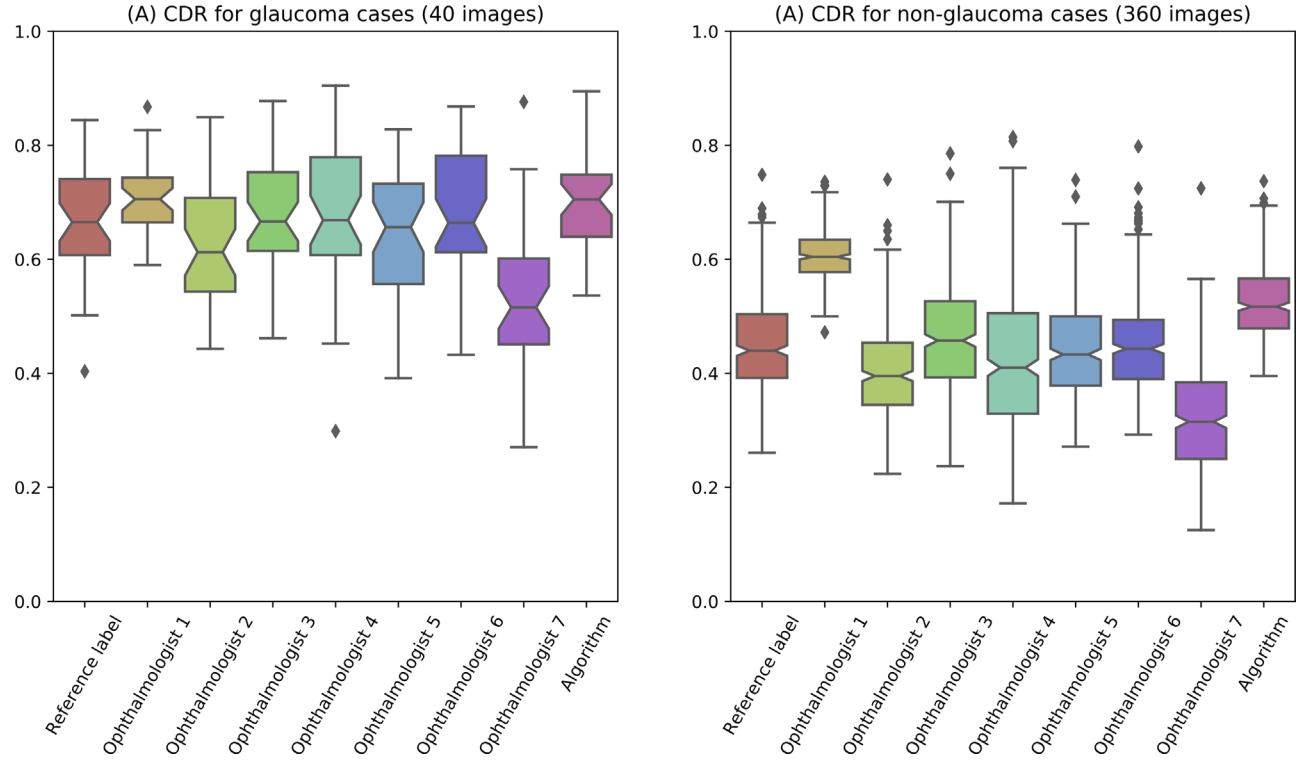

Figure 3. Boxplots of the calculated cup-to-disc ratio (CDR) from segmentation results on test set.

are shown in Figure 5 and Table 2. The algorithm obtained an AUC of 0.948 (95% CI = 0.920–0.973), with a sensitivity of 0.850 (95% CI = 0.794–0.923) and specificity of 0.853 (95% CI = 0.798–0.918). The algorithm obtained the rank 2 discriminating performance, only lower than ophthalmologist 2, who got an AUC of 0.956 (95% CI = 0.933–0.975, P value < 0.0001). Moreover, Figures 4E, 4F showed the false negative and false positive samples, respectively.

Discussion

The purpose of this study was to develop a deep learning algorithm for automated OD and OC segmentation in fundus photographs and compare its performance to ophthalmologist annotations. The results demonstrated that the proposed deep learning algorithm achieved satisfactory performances on

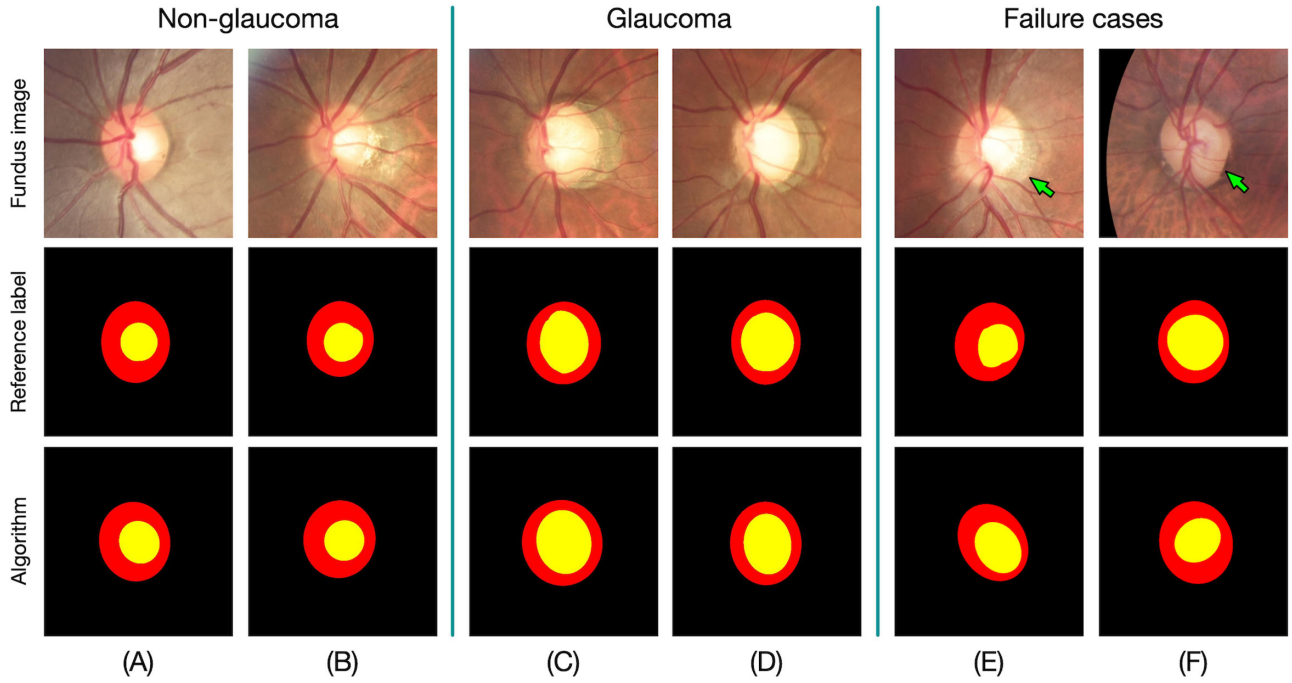

Figure 4. The visual results of segmentation. The segmented optic disc and optic cup regions were labeled by *red* and *yellow* colors, respectively. (**a, b**) Glaucoma cases, (**c, d**) nonglaucoma cases, and (**e, f**) failure cases.

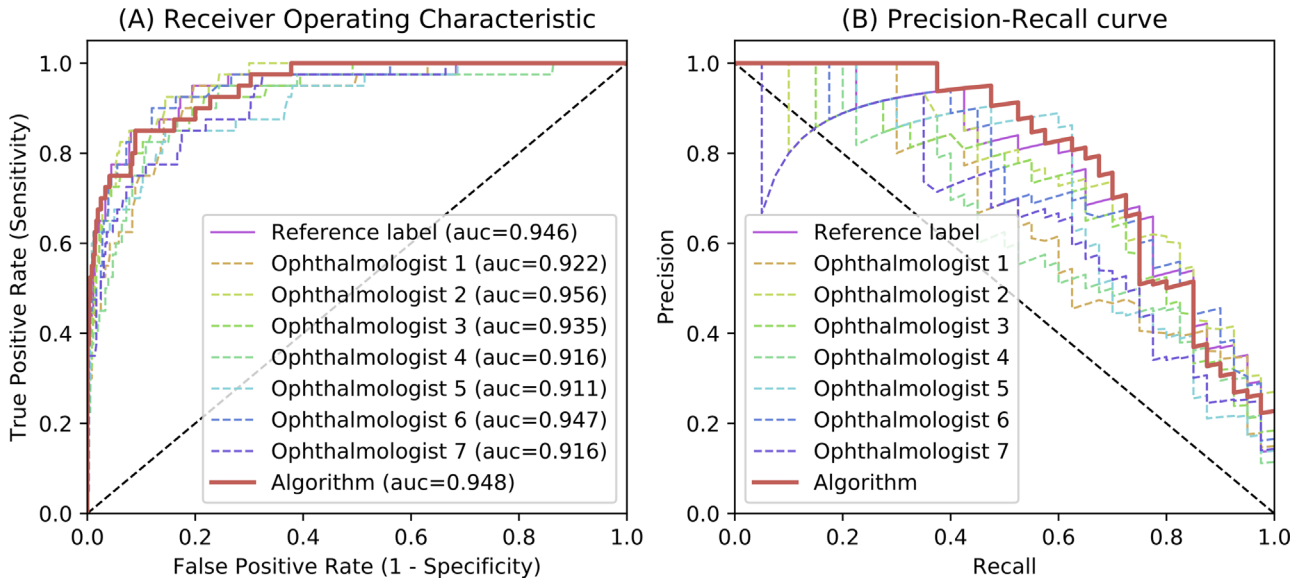

Figure 5. (**A**) The average receiver operating characteristic curves (AUC) for glaucoma diagnosis based on cup to disc ratio (CDR) on test set. (**B**) The precision-recall curves for glaucoma diagnosis based on cup to disc ratio (CDR) on test set.

the OD and OC segmentation task and the glaucoma discriminating task based on CDR calculations.

OD and OC segmentation are fundamental for fundus analysis, especially for CDR calculations during discriminating glaucoma from nonglaucoma subjects. Developing an automated system for this task is crucial. First, as briefly mentioned, manual fundus photograph labeling is highly time-consuming, with the average ophthalmologist requiring 40 seconds to annotate a single photograph. Because our algorithm

Table 2. Diagnosis Performances of Experts and Algorithm on Test Set

	AUC (95% CI)	*P* Value	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
Reference label	0.946 (0.911-0.974)	< 0.0001	0.875 (0.811-0.927)	0.867 (0.819-0.923)	0.700 (0.588-0.791)
Ophthalmologist 1	0.922 (0.884-0.955)	< 0.0001	0.875 (0.800-0.897)	0.858 (0.812-0.887)	0.585 (0.464-0.694)
Ophthalmologist 2	0.956 (0.933-0.975)	< 0.0001	0.875 (0.826-0.930)	0.869 (0.834-0.928)	0.700 (0.577-0.786)
Ophthalmologist 3	0.935 (0.902-0.964)	< 0.0001	0.850 (0.789-0.913)	0.850 (0.795-0.908)	0.650 (0.556-0.771)
Ophthalmologist 4	0.916 (0.871-0.954)	< 0.0001	0.850 (0.791-0.903)	0.847 (0.796-0.898)	0.575 (0.438-0.667)
Ophthalmologist 5	0.911 (0.866-0.951)	< 0.0001	0.850 (0.756-0.892)	0.856 (0.755-0.886)	0.625 (0.525-0.755)
Ophthalmologist 6	0.947 (0.916-0.972)	< 0.0001	0.900 (0.824-0.930)	0.881 (0.832-0.922)	0.659 (0.556-0.773)
Ophthalmologist 7	0.916 (0.876-0.951)	< 0.0001	0.825 (0.765-0.882)	0.825 (0.772-0.878)	0.625 (0.469-0.720)
Algorithm	0.948 (0.920-0.973)	–	0.850 (0.794-0.923)	0.853 (0.798-0.918)	0.700 (0.600-0.800)

AUC, area under the receiver operating characteristic curve; CI, confidence interval.

could reduce this time to 2 seconds, it would be highly beneficial for accelerating processing time and analyzing large-scale datasets. Second, manual annotations are highly subjective. In fact, the segmentations carried out by the ophthalmologists were easily affected by both fundus resolution and image quality. The inter-agreement rating between the various ophthalmologists, for both OD and OC dice scores on the test set, are shown in Figure 2. As can be seen, there was slight variation between the OD segmentation results, with inter-agreement scores ranging from 0.89 to 0.95. However, the OC segmentation task suffered a larger variability, with inter-agreement scores ranging from 0.47 to 0.85. The boundary of OD was clear and definite enough to determine in fundus photograph, which produced a high inter-agreement score between the ophthalmologists, as shown in Figure 2A. Different from the OD, the boundary of OC was more difficult to identify, which was influenced by many factors, such as tilted disc, illumination, and low contrast, etc. These factors may result in the clinical uncertainty during different ophthalmologists and a variable OC segmentation. Moreover, OC segmentation by an ophthalmologist was a highly subjective task, which was related to individual bias and clinical experiences. This also led a low inter-agreement score (see Fig. 2B). By contrast, the automated algorithm provided a consistent result for the same photograph with freezing the trained parameters and model. Moreover, due to limited GPU memory capabilities and parameter size constraints, input fundus photographs had to be down-sampled for training, thus removing the requirement for high-resolution photographs. Another observation is that the performances of algorithm on glaucoma cases (OD dice of 0.941, cup dice of 0.864, and CDR MAE of 0.065) was better than its on nonglaucoma cases (OD dice of 0.937, cup dice of 0.794, and CDR MAE of 0.079). One reason is that the advanced glaucoma cases with severe cupping usually present more clear interfaces between the OD and OC.

Over the decades, many automated deep learning algorithms have been proposed for glaucoma diagnosis in fundus photographs,[22,34] OCT,[35,36] and anterior segment OCT (AS-OCT).[37,38] However, although many of these produce diagnostic results from fundus photographs directly, they lacked clinical interpretability and analyticity. By contrast, segmentation-based algorithms generate a visible segmentation result and have more potential for clinical assistant and analysis. Some automated algorithms based on various visual features and machine learning techniques have been developed for segmenting OD and OC.[28,39,40] Cheng et al.[25] classified each superpixel in the fundus image with various hand-crafted features as OD and OC segmentation and reported an OD dice of 0.905 and OC dice of 0.759. Zheng et al.[41] integrated the OD and OC segmentation within a graph-cut framework. However, they only utilized hand-crafted features, which were affected by the low quality of fundus photographs. In our study, a multilabel deep network was used to obtain highly discriminative representations and segment the OD and OC jointly with the multilabel loss. The results demonstrated that the proposed method enabled automated OD and OC segmentation with a comparable performance to ophthalmologists. We also evaluated the model for discriminating glaucoma from nonglaucoma subjects based on CDR calculations, which were calculated based on the segmentation results as an important glaucoma indicator. The proposed algorithm performed extremely well in comparison to ophthalmologists for glaucoma discriminating.

One limitation of this study was a specific Chinese population that was evaluated and the results may not apply to other ethnic groups. Another potential limitation of our study was that the fundus photographs were only taken using Zeiss Visucam 500 and Canon CR-2 cameras. This could possibly have a negative effect on the quality and performance when the algorithm was applied to images from other fundus acquisition devices. Third, in our study, we added CDR calculations as one of the clues for glaucoma diagnosis. However, some patients were shown to have a small CDR despite significant VF loss, whereas others displayed a large CDR without reporting any VF loss.[7] The dataset contained 10% of glaucoma subjects and most of these glaucoma subjects were at moderate or advanced stages, the difficulty of discriminating glaucoma from nonglaucoma subjects based on CDR calculation was relatively lower. The performance of the algorithm may go down in another larger dataset. Future studies were needed to explore whether other annotations, such as RNFL defects, would further enhance the performance of the algorithm. Besides, early-stage glaucoma is very hard to diagnose through fundus photographs. It would be interesting to add more photographs from early-stage patients and train the algorithm to make a diagnosis. We may try to find new clues other than CDR or RNFL defects in glaucoma discriminating based on fundus photographs.

In summary, we developed and investigated a deep learning system for OD and OC segmentation in fundus images. Deep learning technique was shown to be a promising technology for helping clinicians to reliably and rapidly identify OD and OC regions. Moreover, we also evaluated discriminating glaucoma from nonglaucoma subjects based on the CDR calculations, where the proposed algorithm performed extremely well in comparison to ophthalmologists, obtaining an AUC of 0.946. As such, our technique showed high potential for assisting ophthalmologists in fundus analysis and glaucoma screening.

Acknowledgments

The sponsor or funding organization had no role in the design or conduct of this research.

Disclosure: **H. Fu**, None; **F. Li**, None; **Y. Xu**, None; **J. Liao**, None; **J. Xiong**, None; **J. Shen**, None; **J. Liu**, None; **X. Zhang**, None

* HF and FL share first authorship and contributed equally to this work.

iChallenge-GON study group includes:

Chunman Yang: The 2nd Affiliated Hospital of Guizhou Medical University, Kaili, Guizhou, China.

Fengbin Lin: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.

Huang Luo: Guangzhou Hospital of TCM, Guangzhou, Guangdong, China.

Hao Li: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.

Huixin Che: Aier Eye Hospital, Jinzhou, Liaoning, China.

Nuhui Li: Zhongshan Ophthalmic Center, Guangzhou, Guangdong, China.

Yazhi Fan: The 2nd Affiliated Hospital of Xi'an Jiaotong University, Shaanxi, China.

References

1. Jonas JB, Aung T, Bourne RR, Bron AM, Ritch R, Panda-Jonas S. Glaucoma. *Lancet*. 2017;390:2183–2193.
2. Medeiros FA, Zangwill LM, Bowd C, Vessani RM, Susanna R, Weinreb RN. Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *Am J Ophthalmol*. 2005;139:44–55.
3. Bussel II, Wollstein G, Schuman JS. OCT for glaucoma diagnosis, screening and detection of glaucoma progression. *Br J Ophthalmol*. 2014;98(Suppl 2):ii15–ii19.
4. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma. *JAMA*. 2014;311:1901–1911.
5. Hayreh SS. Optic disc changes in glaucoma. *Br J Ophthalmol*. 1972;56:175–185.
6. Hitchings RA, Spaeth GL. The optic disc in glaucoma. I: Classification. *Br J Ophthalmol*. 1976;60:778–785.
7. Garway-Heath DF, Ruben ST, Viswanathan A, Hitchings RA. Vertical cup/disc ratio in relation to optic disc size: its value in the assessment of the glaucoma suspect. *Br J Ophthalmol*. 1998;82:1118–1124.
8. Foster PJ. The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol*. 2002;86:238–242.
9. Gordon MO, Beiser JA, Brandt JD, et al. The ocular hypertension treatment study: baseline factors

that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120:714–720.
10. Tielsch JM, Katz J, Quigley HA, Miller NR, Sommer A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology*. 1988;95:350–356.
11. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215–221.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
13. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1–29.
14. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402.
15. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211.
16. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803.
17. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135:1170.
18. Burlina PM, Joshi N, Pacheco KD, Liu TYA, Bressler NM. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol*. 2019;137:258.
19. Liu H, Wong DWK, Fu H, Xu Y, Liu J. Deep AMD: detect early age-related macular degeneration by applying deep learning in a multiple instance learning framework. In: Jawahar C, Li H, Mori G, Schindler K, eds. Computer Vision – ACCV 2018. Asian Conference on Computer Vision. Lecture Notes in Computer Science, vol. 11365. Cham: Springer: 2019:625–640.
20. Zhao H, Li H, Maurer-Stroh S, Cheng L. Synthesizing retinal and neuronal images with generative adversarial nets. *Med Image Anal*. 2018;49:14–26.
21. Keel S, Wu J, Lee PY, Scheetz J, He M. Visualizing deep learning models for the detection of referable diabetic retinopathy and glaucoma. *JAMA Ophthalmol*. 2019;137:288.
22. Fu H, Cheng J, Xu Y, et al. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Transactions on Medical Imaging*. 2018;37:2493–2501.
23. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353.
24. Kumar JRH, Seelamantula CS, Kamath YS, Jampala R. Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. *Sci Rep*. 2019;9:7099.
25. Cheng J, Liu J, Xu Y, et al. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Trans Med Imaging*. 2013;32:1019–1032.
26. 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;37:1597–1605.
27. Jiang Y, Duan L, Cheng J, et al. JointRCNN: a region-based convolutional neural network for optic disc and cup segmentation. *IEEE Trans Biomed Eng*. 2020;67:335–343.
28. Gu Z, Cheng J, Fu H, et al. CE-Net: context encoder network for 2D medical image segmentation. *IEEE Trans Med Imaging*. 2019;38:2281–2292.
29. Orlando JI, Fu H, Barbosa Breda J, et al. REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal*. 2020;59:101570.
30. Garway-Heath DF, Crabb DP, Bunce C, et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. *Lancet*. 2015;385:1295–1304.
31. Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods*. 2019;16:67–70.
32. Weerahandi S. Generalized Confidence Intervals. In: *Exact Statistical Methods for Data Analysis*. New York, NY: Springer New York; 1995:143–168.
33. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986;1:54–75.
34. Asaoka R, Murata H, Hirasawa K, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. 2019;198:136–145.
35. Thompson AC, Jammal AA, Berchuck SI, Mariottoni EB, Medeiros FA. Assessment of

a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmol*. 2020;27710:1–7.
36. Maetschke S, Antony B, Ishikawa H, Wollstein G, Schuman J, Garnavi R. A feature agnostic approach for glaucoma detection in OCT volumes. Grulkowski I, ed. *PLoS One*. 2019;14:e0219126.
37. Fu H, Baskaran M, Xu Y, et al. A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *Am J Ophthalmol*. 2019;203:37–45.
38. Fu H, Xu Y, Lin S, et al. Angle-closure detection in anterior segment OCT based on multilevel deep network. *IEEE Trans Cybern*. 2019:1–9.
39. Wang L, Nie D, Li G, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 Challenge. *IEEE Trans Med Imaging*. 2019;38:2219–2230.
40. Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Comput Med Imaging Graph*. 2019;74:61–71.
41. Zheng Y, Stambolian D, O'Brien J, Gee J. Optic disc and cup segmentation from color fundus photograph using graph cut with priors. *Med Image Comput Comput Assist Interv*. 2013;16(Pt 2):75–82.