A Large-scale Database and a CNN Model for Attention-based Glaucoma Detection

Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang

***Abstract*—Glaucoma is one of the leading causes of irreversible vision loss. Many approaches have recently been proposed for automatic glaucoma detection based on fundus images. However, none of the existing approaches can efficiently remove high redundancy in fundus images for glaucoma detection, which may reduce the reliability and accuracy of glaucoma detection. To avoid this disadvantage, this paper proposes an attention-based convolutional neural network (CNN) for glaucoma detection, called AG-CNN. Specifically, we first establish a large-scale attention-based glaucoma (LAG) database, which includes 11,760 fundus images labeled as either positive glaucoma (4,878) or negative glaucoma (6,882). Among the 11,760 fundus images, the attention maps of 5,824 images are further obtained from ophthalmologists through a simulated eye-tracking experiment. Then, a new structure of AG-CNN is designed, including an attention prediction subnet, a pathological area localization subnet and a glaucoma classification subnet. The attention maps are predicted in the attention prediction subnet to highlight the salient regions for glaucoma detection, under a weakly supervised training manner. In contrast to other attention-based CNN methods, the features are also visualized as the localized pathological area, which are further added in our AG-CNN structure to enhance the glaucoma detection performance. Finally, the experiment results from testing over our LAG database and another public glaucoma database show that the proposed AG-CNN approach significantly advances the state-of-the-art in glaucoma detection.**

***Index Terms*—glaucoma detection, attention mechanism, pathological area detection, weakly supervised.**

I. Introduction

Glaucoma is one of the leading causes of irreversible blindness [3]. The incidence of serious glaucoma is reported to be 3.5% among people over 45 years of age, i.e., approximately 64.3 million individuals are suffering from glaucoma worldwide [45]. This number is predicted to increase to 80 million by 2020 and to 111.8 million by 2040 as a result of aging and population growth [45]. Most vision loss caused by glaucoma can be avoided through early detection and treatment [44]. Thus, it is important to detect glaucoma at an early stage. However, due to the lack of qualified ophthalmologists, it is hard to conduct manual glaucoma screening for all suspected patients. Therefore, developing an automatic method for glaucoma detection with high accuracy and efficiency is essential.

In recent years, several methods have been developed to detect glaucoma based on optical fundus images, which are the photographs of the back of the eyes. These methods can be divided into 2 categories: heuristic methods and deep learning methods. The heuristic methods employ the handcrafted features of the vertical cup-disc ratio (VCDR) based on the segmentation [41]. VCDR is one of the principles for ophthalmologists to diagnose glaucoma [17]. However, these methods are affected by the accuracy of segmenting the optic cup since the boundary of the optic cup remains ill-defined and fuzzy in glaucoma images, which is caused by the irreversible damage of the nerve fiber layer and the quality of fundus images. Another category of glaucoma detection algorithms is based on convolutional neural networks (CNNs) [29], [6], [27], [11]. Such methods achieve end-to-end training and testing, feeding the fundus images as input and directly outputting the binary labels of positive and negative glaucoma. A CNN automatically learns the extensive features for classification, without any segmentation of the fuzzy optical cups. Nevertheless, most of these methods lack sufficient training data, inevitably leading to the overfitting problem. Recently, [29] proposed a deeper CNN method for glaucoma detection, benefiting from the large-scale database established in their work. This work transfers the task of nature images classification [7], [43] to glaucoma detection on fundus images. Comparing with nature images, the fundus images contain large redundancy regions without any valuable information for glaucoma detection, e.g., the black-background of the fundus images and the edge regions of the eyeball. The redundancy regions may mislead the CNN to focus on the useless information. Thus, [29] is ineffective in dealing with the redundant information in the fundus images.

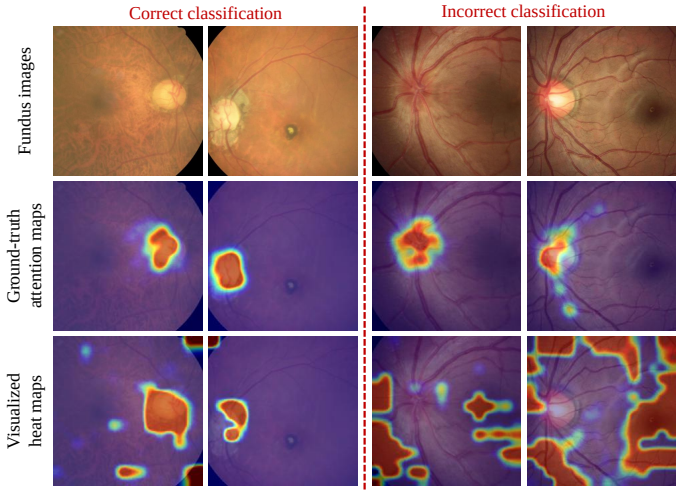

Fig. 1. Examples of glaucoma fundus images, attention maps by ophthalmologists in glaucoma diagnosis and visualization results of a CNN model [19] by an occlusion experiment [53]. The Pearson correlation coefficient (CC) between the visualized heat maps and ground-truth ophthalmologist attention maps are 0.33 and 0.14 for correct and incorrect glaucoma classification, respectively.

As shown in Figure 1, glaucoma can be correctly detect-

L. Li, M. Xu, X. Wang, L. Jiang and Z. Wang are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191 China; Y. Li is with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191 China; H. Liu, N. Wang are with the Beijing Institute of Ophthalmology, Beijing 100730 China; X. Fan is with the Department of Ophthalmology, Peking University Third Hospital, Beijing 100191 China. M. Xu is the corresponding author of this paper (E-mail: Maixu@buaa.edu.cn).

ed by a CNN method [19] when the visualized heat maps are consistent with the attention maps of ophthalmologists in glaucoma diagnosis. Otherwise, glaucoma is mislabeled by the CNN model. Therefore, it is reasonable to combine the attention mechanism in the CNN model for using fundus images to detect ophthalmic diseases. However, to our best knowledge, there has been no work incorporating human attention in fundus image recognition. This is mainly due to the lack a doctor attention database, which requires the qualified doctors and a special technique for capturing the doctor's attention during the diagnosis. In this paper, we first collect a large-scale attention-based fundus image database for glaucoma detection (LAG), which includes 11,760 images with diagnosis labels, and 5,824 of them are further labeled with human attention. Then, we develop an attention-based CNN method for glaucoma detection (AG-CNN), which is supervised by the labeled human attention. We further train the AG-CNN to predict attention maps in a weakly supervised manner because the ground truth (GT) attention maps are available for only part of the training images.

However, there also exist some insignificant pathological areas in the fundus images, which may not attract human attention. Consequently, the existing CNN models have already outperformed human experts in some medical image recognition tasks [23], [35], [34]. Thus, we propose refining the predicted attention maps by incorporating a feature visualization structure for glaucoma detection. In this way, the gap between human attention and pathological area can be bridged.In fact, several methods have been proposed for automatically locating the pathological area [54], [15], [10], [14], [30] based on the class activation mapping model (CAM) [55]. However, these methods cannot locate the pathological area in a small region due to the limitation of its feature size. In this paper, we employ the guided back-propagation (BP) method to locate the tiny pathological area based on the predicted attention maps. Consequently, the attention maps can be refined and then used to highlight the most critical region for glaucoma detection.

This paper extends our conference paper [28] from four aspects. First, we further review more related works, in particular the attention mechanism applied in deep learning methods. Second, we significantly enlarge our LAG database to 11,760 fundus images, including another 5,936 fundus images. Third, in addition to the supervised method in [28], this paper further proposes a weakly supervised learning for glaucoma detection. Finally, the additional experiments are presented for thoroughly evaluating the performance of our method. The main contributions of this paper are twofold. (1) We establish a LAG database that includes 11,760 glaucoma-labeled fundus images, among which 5,824 images are further labeled with attention maps. (2) We propose a new AG-CNN architecture for locating pathological areas and then classifying binary glaucoma, in which attention maps are incorporated in a weakly supervised manner to remove the redundancy from fundus images for glaucoma detection.

II. Related work

A. Automatic glaucoma detection

The recent success of deep learning methods has benefited medical diagnosis [9], [5], [51], particularly automatically detecting oculopathy in fundus images [16], [13], [46]. Specifically, [16], [13] worked on classifying of diabetic retinopathy using the CNN models. [46] further proposed deep learning systems for detecting multiple ophthalmological diseases. However, the above works all transferred some classic CNN model for nature image classification to medical image classification without consideration of the characteristics of fundus images.

Glaucoma detection methods can basically be divided into 2 categories: heuristic methods and deep learning methods. The heuristic glaucoma detection methods extract features based on some image processing techniques [1], [8], [41], [32], [21]. Specifically, glaucoma screening methods were proposed in [32], [21], based on the detection of retinal nerve fiber layer (RNFL) thickness. [1] extracted the texture features and higher order spectral features for glaucoma detection. [8] used the wavelet-based energy features for glaucoma detection. [1], [8] both applied support vector machine (SVM) and naive Bayesian classifier to classify the handcrafted features. However, the above heuristic methods only consider a handful of features on fundus images, leading to lower classification accuracy.

Another category of glaucoma detection methods is based on deep learning [37], [56], [33], [6], [27], [29], [11], [22], [39]. Specifically, [37], [56], [22] reported their deep learning work on glaucoma detection based on the automatic segmentation of the optic cup and disc. However, their work assumes that only the optic cup and disc are related to glaucoma, lacking end-to-end training. [11] further proposed a multistream CNN that combined the full optical images with the segmentation result. Panda *et al.* [33] proposed a deep learning method for glaucoma detection based on RNFL defect. On the other hand, [6] firstly proposed a CNN method for glaucoma detection in an end-to-end manner. [27] followed Chen's work and proposed an advanced CNN structure combining the holistic and local features for glaucoma classification. To regularize the input images, both [6] and [27] preprocessed the original fundus images to remove the redundant regions. However, due to the limited training data and simple structure of networks, the previous works did not achieve high sensitivity and specificity. Shibata *et al.* [39] proposed a CNN method based on ResNet, improving the performance in glaucoma detection. Recently, a deeper CNN structure has been proposed in [29]. However, the fundus images contain large redundancy regions that are irrelevant for glaucoma detection, which leads to the low efficiency of the method in [29]. Note that the efficiency here refers to the effort in extracting useful features for glaucoma detection.

B. Attention mechanism

In recent years, the attention mechanism has been successfully applied in deep-learning-based computer vision tasks, e.g., object detection [2], [36], image caption [48], [52] and action recognition [38]. The basic idea of the attention mechanism is to locate the most salient parts of the features in deep neural networks (DNNs) such that redundancy is removed for the vision tasks. In general, the attention mechanism is embedded in DNNs by leveraging the attention maps. Specifically, on the one hand, the attention maps in [40], [36], [48], [38] are yielded in a self-learned pattern, with other information weakly supervising the attention maps, e.g., the classification labels. On the other hand, [52] utilize the human attention information to guide the DNNs to focus on the region of interest (ROI).

TABLE I
SUMMARY OF OUR LAG DATABASE.

Source Database	Images No.	Positive No. (%)	Individuals No.	Age, Mean (SD)	Female No. (%)	Camera
CGSA	7,463	2,749 (36.8)	6,441	54.1 (14.5)	55.8	Topcon, Canon, Carl Zeiss
Beijing Tongren Hospital	4,297	2,129 (49.5)	3,706	52.8 (16.7)	49.7	Topcon, Canon
Full LAG	11,760	4,878 (41.5)	10,147	53.6 (15.3)	53.6	Topcon, Canon, Carl Zeiss

TABLE II
IMAGE NUMBERS IN OUR LAG DATABASE.

Dataset	Image Number, (With/Without Attention Groundtruth)		
	All	Positive Glucoma	Negative Glcuoam
LAG	11,760 (5,824/5,936)	4,878 (2,392/2,486)	6,882 (3,432/3,450)
Training	10,928 (4,992/5,936)	4,528 (2,042/2,486)	6,400 (2,950/3,450)
Validation	832 (832/0)	350 (350/0)	482 (482/0)

TABLE III
CC VALUES OF ATTENTION MAPS BETWEEN ONE OPHTHALMOLOGIST AND THE MEAN OF THE REMAINING OPHTHALMOLOGISTS.

Ophthalmologist	one v.s. others	one v.s. random
1^{st}	0.594	6.59×10^{-4}
2^{nd}	0.636	2.49×10^{-4}
3^{rd}	0.687	2.49×10^{-4}
4^{th}	0.585	8.44×10^{-4}

Redundancy also exists in medical image recognition, interfering with the recognition results. In particular, heavy redundancy exists in fundus images for disease recognition. For example, the pathological areas of fundus images are in the region of the optic cup and disc or its surrounding blood vessel and optic nerve area [31]; other regions, such as the boundary of the eyeball, are redundant for medical diagnosis. Therefore, it is reasonable to combine the attention mechanism in the CNN model for using fundus images to detect ophthalmic diseases.

III. DATABASE

A. Establishment

In this section, we establish the LAG database, which can be used for glaucoma detection.[1] Our LAG database contains 11,760 fundus images corresponding to 4,878 positive and 6,882 negative glaucoma samples. 10,861 individuals are involved in our LAG database, from which 10,147 individuals are collected with only one fundus image, i.e., one image per eye and subject. The remaining individuals refer to multiple images per subject. More details about the numbers of individuals with multiple images are reported in Table 1 of the supplementary materials.

The fundus images in our LAG database are obtained from Chinese Glaucoma Study Alliance (CGSA) and Beijing Tongren Hospital. The CGSA was established from 2009, progressively covering 89 hospitals across China, to further improve the diagnosis and treatment capacity of vision-threatening eye diseases, including glaucoma. As shown in Table I, detail information about our LAG database is list as follows. The number of female patients is 6,300 (53.6%) and the average age of our LAG database is 53.6 with a standard deviation (SD) of 15.3. The fundus images in our database are taken by 3 types of devices: Topcon, Canon and Carl Zeiss. Besides, the dimension of input fundus images ranges from 582×597 to 3456×5184, with an average of 1977×2594 and a standard deviation of 840×1417.

[1]The database is available at https://github.com/smilell/AG-CNN.

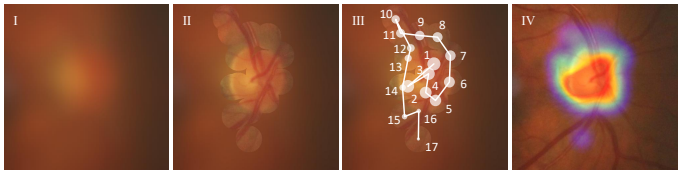

Fig. 2. An example of capturing fixations of an ophthalmologist in glaucoma diagnosis. (Left): Original blurred fundus images. (Middle-left): Fixations of the ophthalmologist with cleared regions. (Middle-right): The order of clearing the blurred regions. Note that the size of the white circles represents the order of fixations. (Right): The generated attention map based on the captured fixations.

Each sample in our LAG database is subject to a tiered grading system that consists of 3 layers of trained graders with increasing expertise. The first tier of graders consists of five trained medical students conducting initial quality control, i.e., the image field includes the entire optic nerve head and macula, the illumination is acceptable, the image is sufficiently focused for grading the optic nerve head and RNFL. The second tier of graders consists of five Chinese board-certified ophthalmologists or postgraduate ophthalmology trainees ($\geq$ 2 years of experience) who have passed a pretraining test. In the process of grading, each image is assigned randomly to two ophthalmologists for grading. Each grader independently grades and records each image according to the criteria of referable glaucomatous optic neuropathy [29]. In our grading system, the grading accuracy of the 5 experts from tier 2 is 88.4%, 87.7%, 90.0%, 87.0% and 92.7%, respectively. The third tier of grader is a senior independent glaucoma specialist ($>$ 10 years of experience in conducting glaucoma retinopathy diagnosis), who is consulted in the cases of disagreement in tier 2 grading. Consequently, all the samples in our LAG database are labeled with positive glaucoma or negative glaucoma. The details of the image number of our database is listed in Table II. Note that our database is constructed according to the tenets of the Declaration of Helsinki. Because of the retrospective nature and fully anonymized usage of color retinal fundus images in this work, we are exempted by the medical ethics committee to inform the patients.

Based on the above labeled fundus images, we further conduct an experiment to capture the attention regions of 4 ophthalmologists in glaucoma diagnosis, which includes 2 third tier graders and 2 second tier graders, respectively. Note that the 4 ophthalmologists are independent of the aforementioned tiered grading system. Table II shows that 5,824 fundus images are further labeled with attention regions, in which 2,392 are positive glaucoma and the rest 3,432 are negative glaucoma. The experiment is based on an alternative method for eye tracking [24], in which mouse clicks are used by the ophthalmologists to explore the ROI for glaucoma diagnosis. Specifically, all the fundus images are initially displayed blurred, and then the ophthalmologists use the mouse as an eraser to successively clear the circle regions for diagnosing glaucoma. Note that

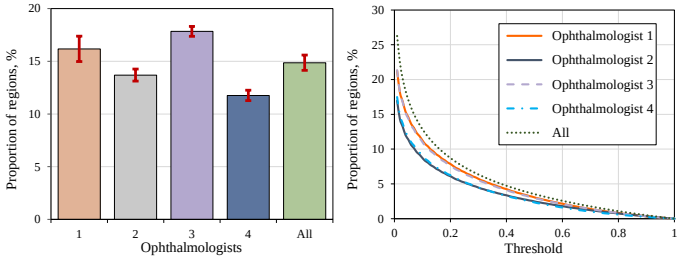

Fig. 3. (Left): Proportion of regions in the fundus images cleared by different ophthalmologists for glaucoma diagnosis. (Right): Proportion of regions in attention maps with values being above a varying threshold. Note that the values of the attention maps range from 0 to 1.

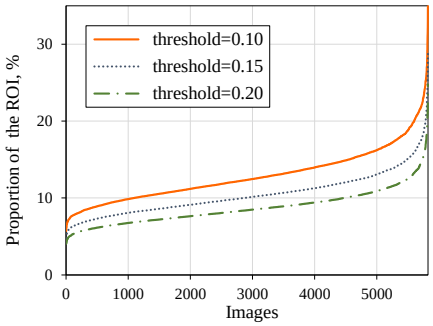

Fig. 5. Proportion of ROI above the thresholds of 0.10, 0.15 and 0.20 for all of the fundus images in the LAG database.

the radius of all circle regions is set to 40 pixels, while all fundus images are 500×500 pixels. This ensures that the circle regions are approximately equivalent to the fovea ($2^\circ - 3^\circ$) of the human vision system at a comfortable viewing distance (3-4 times the screen height). The order of clearing the blurred regions represents the degree of attention by ophthalmologists, as the GT of attention maps. Once the ophthalmologist is able to diagnose glaucoma with the partially cleared fundus image, the above region clearing process is terminated and the next fundus image is displayed for diagnosis.

In the above experiment, the fixations of ophthalmologists are represented by the center coordinate (x_i^j, y_i^j) of the cleared circle region for the i-th fixation of the j-th ophthalmologist. Then, the attention map $\mathbf{A}$ of one fundus image can be generated by convoluting all fixations $\{(x_i^j, y_i^j)\}_{i=1,j=1}^{I_j,J}$ with the 2D Gaussian filter at square decay according to the order of i, where J is the total number of ophthalmologists (=4 in our experiment) and I_j is the number of fixations from the j-th ophthalmologist on the fundus image. Here, the standard deviation of the 2D Gaussian filter is set to 25, according to [49]. Figure 2 shows an example of the fixations of one ophthalmologist and the attention map of all ophthalmologists for a fundus image.

In conclusion, our LAG database consists of 3 parts, i.e., fundus images, diagnosis labels and attention regions, and it requires permission, qualified glaucoma specialists and annotation software as follows. (1) The permission has been obtained from CGSA and Beijing Tongren Hospital for using the fundus images with the purpose of academic research. (2) Several qualified glaucoma specialists with an unified grading standard have been involved in our grading system. (3) An annotation software has been developed in this paper, in order to obtain the attention information from ophthalmologists and further generate the attention maps.

B. Data analysis

Now, we mine our LAG database to investigate the attention maps of 5,824 fundus images in glaucoma diagnosis. Specifically, we obtain the following findings.

Finding 1: *The ROI in fundus images is consistent across ophthalmologists for glaucoma diagnosis.*

Analysis: In this analysis, we calculate the Pearson correlation coefficients (CCs) of attention maps between one ophthalmologist and the remaining three ophthalmologists. We follow [50] to calculate the CC values at pixel-wise. Mathematically, it is calculated by

$$\mathbf{CC} = \frac{\sum_{i=1}^{W}\sum_{j=1}^{H}(\mathbf{A}_{ij} - \mu_a) \cdot (\bar{\mathbf{A}}_{ij} - \mu_{\bar{a}})}{\sqrt{\sigma_a^2 \cdot \sigma_{\bar{a}}^2}}, \quad (1)$$

where μ_a and σ_a represent the mean and standard deviation of $\mathbf{A}$, while $\mu_{\bar{a}}$ and $\sigma_{\bar{a}}$ denote the mean and standard deviation of $\bar{\mathbf{A}}$. Additionally, W and H are the width and height of $\mathbf{A}$ and $\bar{\mathbf{A}}$. Table III reports the CC results averaged over 5,824 fundus images. In this table, we also show the CC results of attention maps between one ophthalmologist and the random baseline. Note that the random baseline generates the attention maps by making their values follow a Gaussian distribution.As shown in Table III that the CC values of attention maps between one and the remaining ophthalmologists are all above 0.58, significantly larger than those of the random baseline. According to [4], the CC value of human attention on natural images is 0.52 among different subjects, which is lower than the CC values among ophthalmologists. This result implies that ophthalmologists are consistent in where they focus their attention during glaucoma diagnosis. This completes the analysis of *Finding 1*.

Finding 2: *The ROI in fundus images concentrates on small regions for glaucoma diagnosis.*

Analysis: In this analysis, we calculate the percentage of regions that ophthalmologists cleared for glaucoma diagnosis. Figure 3 (left) shows the percentage of the cleared regions for each ophthalmologist, which is averaged over the 5,824 fundus images in our LAG database. As shown, the average ROI accounts for 14.3% of the total area in the fundus images, with a maximum of 17.8% (the 3^{rd} ophthalmologist) and a minimum of 11.8% (the 4^{th} ophthalmologist). Moreover, we calculate the proportion of regions in attention maps, the values of which are above a varying threshold. The result is shown in Figure 3 (right). The rapidly decreasing curve shows that most attention only focuses on small regions of fundus images for glaucoma diagnosis. This completes the analysis of *Finding 2*.

Finding 3: *The ROI for glaucoma diagnosis is of different scales.*

Analysis: The above findings show that the ROI is consistent and small for glaucoma diagnosis. Here, we further analyze the size of the ROI in fundus images. To this end, Figure 4 visualizes the fixation maps of some fundus images, in which the ROI has different scales. As shown in this figure, the sizes of the optic discs for pathological myopia are considerably larger than others. Note that pathological myopia is an eye disorder that the patients of myopia see distant objects to be blurry while the close objects appear normal. It is caused by biomechanical forces related to axial elongation, resulting in larger larger optic

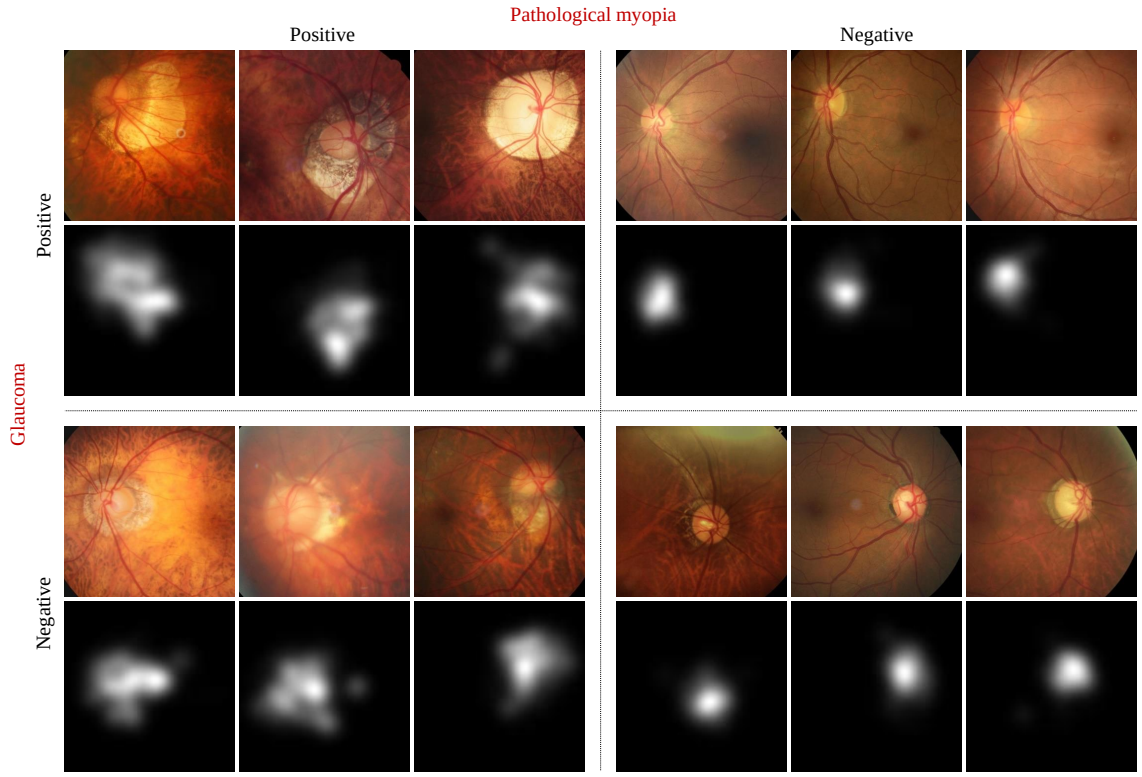

Fig. 4. Fundus images with or without glaucoma for both positive and negative pathological myopia.

disc area [47]. Accordingly, the ROI in fundus images (i.e., high-valued regions in fixation maps) is at multiple scales for glaucoma diagnosis. Note that when the ROI is either small or large, both positive and negative glaucoma results exists. For each image in our LAG database, Figure 5 further plots the proportion of the ROI in the fixation maps, the values of which are larger than a threshold. As shown, the ROI is at different scales for glaucoma diagnosis. Finally, the analysis of *Finding 3* can be accomplished.

IV. Method

A. Framework

In this section, we focus on the proposed AG-CNN method. The framework of AG-CNN is shown in Figure 6. As shown in Figure 6, the input to AG-CNN is the RGB channels of a fundus image, while the output is (1) the located pathological area and (2) the binary glaucoma label. Our AG-CNN has two 2 stages as follows.

- In the first stage, the ROI of glaucoma detection is learned from the attention prediction subnet, aiming to predict human attention on diagnosing glaucoma. It is because *Findings 1* and *2* show that glaucoma diagnosis is highly related to small ROI regions. In addition, the multiscale building block is also included in our AG-CNN model, because *Finding 3* shows that ROIs for glaucoma diagnosis are of different scales.
- In the second stage, the predicted attention map is embedded in the pathological area localization subnet, and then the feature map of this subnet is visualized to locate the pathological area. It is because the CNN may extract some pathological areas that are not obvious for the diagnosis by ophthalmologists [23], [35], [34]. Finally, the located pathological area combined with the predicted attention map is further used to mask the input and features of the glaucoma classification subnet, for outputting the binary labels of glaucoma.

The main structure of AG-CNN is based on residual networks [19], in which the basic module is a building block. Note that all convolutional layers in AG-CNN are followed by a batch normalization layer and a ReLU layer for increasing the nonlinearity of AG-CNN such that the convergence rate can be accelerated. The process of training AG-CNN is in an end-to-end manner with three parts of supervision: attention prediction loss, feature visualization loss and glaucoma classification loss.

B. Attention prediction subnet

In AG-CNN, an attention prediction subnet is designed to generate the attention maps of the fundus images, which are then used for pathological area localization and glaucoma detection. Specifically, the input of the attention prediction subnet is the RGB channels of a fundus image, which are represented by a tensor (size: $224 \times 224 \times 3$). Then, the input tensor is fed to one convolutional layer with a kernel size of 7×7, followed by one max-pooling layer. Subsequently, the features flow into 8 building blocks for extracting the hierarchical features. For more details about the building blocks, refer to [19]. Afterwards, the features of 4 hierarchical building blocks are processed by feature normalization (FN), the structure of which is shown in Figure 6 (lower-right). Consequently, four $28 \times 28 \times 128$ features are obtained. These features are concatenated to form $28 \times 28 \times 512$ deep multiscale features. Given the deep multiscale features, a deconvolutional module is applied to generate the gray attention map with the size of $112 \times 112 \times 1$. The structure

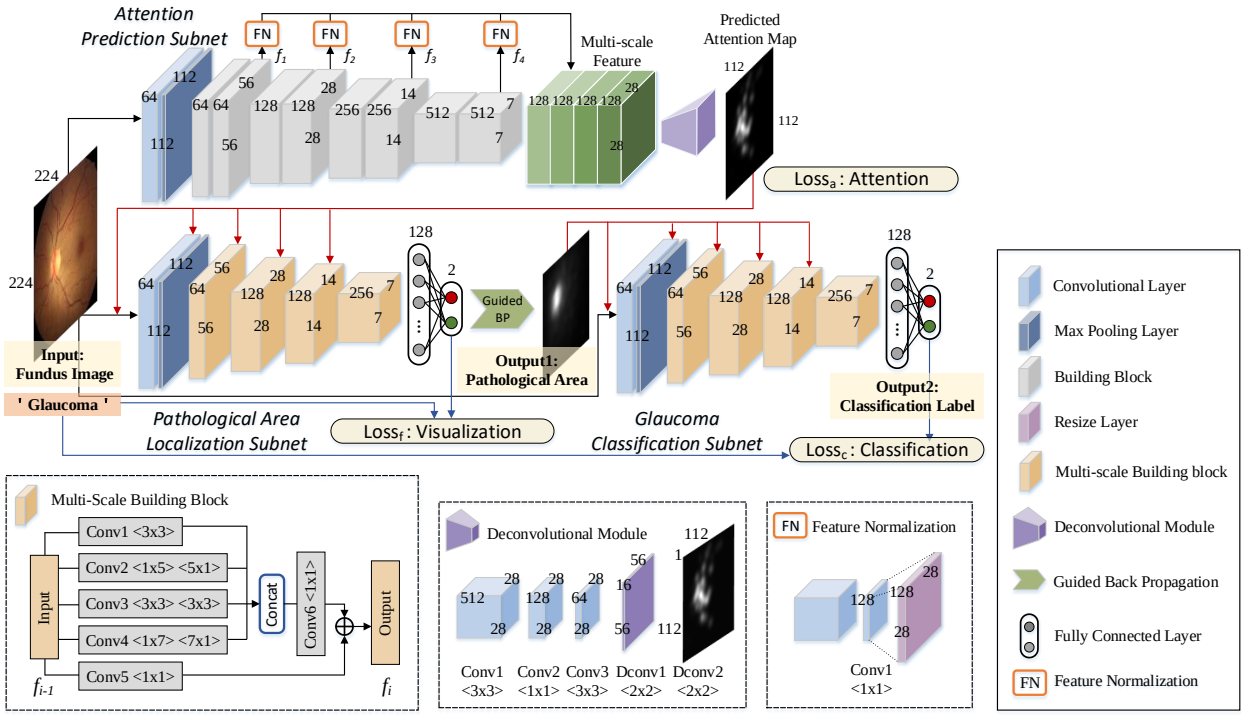

Fig. 6. Architecture of our AG-CNN network for glaucoma detection and its components, with the sizes of the feature maps and convolutional kernels.

of the deconvolutional module is also shown in Figure 6 (lower-middle). As shown in this figure, the deconvolutional module consists of 4 convolutional layers and 2 deconvolutional layers. Finally, a $112 \times 112 \times 1$ attention map can be yielded, the values of which range from 0 to 1. In AG-CNN, the yielded attention maps of glaucoma detection are used to weight the input fundus images and the extracted features of the pathological area localization subnet. This is to be discussed in the next section.

C. Pathological area localization subnet

After predicting the attention maps, we further design a pathological area localization subnet to visualize the CNN feature map in glaucoma classification. The predicted attention maps can effectively make the network focus on the salient region with reduced redundancy; however, the network may inevitably miss some potential features that are useful for glaucoma classification. Moreover, it has been verified that the deep learning methods outperform humans in the task of image classification both on nature images [18], [26] and medical images [23], [35], [34]. Therefore, we further design a subnet to visualize the CNN features for finding the pathological area.

Specifically, the pathological area localization subnet is mainly composed of convolutional layers and fully connected layers. In addition, the predicted attention maps are used to mask the input fundus images and the extracted feature maps at different layers of the pathological area localization subnet. The structure of this subnet is the same as the glaucoma classification subnet, which is to be discussed in Section IV-D. Then, the visualization map of the pathological area is obtained through guided back-propagation (BP) [42] from the output of the fully connected layer f_{out} to the input RGB channels fundus image I. The difference between guided BP and normal BP is the activation function of ReLU. In the process of forward propagation (FP), the input to a ReLU layer is defined as u^i, and its output is defined as u^{i+1}. In the process of BP, the input to this ReLU layer is denoted as R^{i+1}, and its output is denoted as R^i. Mathematically, we have

$$R^{i+1} = \frac{\partial f_{out}}{\partial u^{i+1}}, \tag{2}$$

for BP. Then, the guided BP of ReLU can be represented as follows,

$$R^i = H(R^{i+1}) \cdot H(u^i) \cdot R^{i+1}, \tag{3}$$

where

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{4}$$

Finally, the visualization map is downsampled to 112×112 with its values being normalized to $0 - 1$ as the output of the pathological area localization subnet.

D. Glaucoma classification subnet

In addition to the subnet of attention prediction and pathological area localization, we design a glaucoma classification subnet for the binary classification of positive or negative glaucoma. Similar to the attention prediction subnet, the glaucoma classification subnet is composed of one 7×7 convolutional layer, one max-pooling layer and 4 multiscale building blocks. The multiscale building block differs from the traditional building block of [19] from the following aspect. As shown in Figure 6 (lower-left), 5 channels of convolutional layers with different kernel sizes are concatenated to extract multiscale features compared with the traditional building block, which only has two convolutional channels. Finally, 2 fully connected layers are applied to output the classification result.

The main difference between the glaucoma classification subnet and the conventional residual network [19] is that the refined attention maps, combining the predicted attention maps

and the visualization maps of pathological area, weight both the input image and extracted features to focus on the ROI. Assume that the refined attention map is $\hat{\mathbf{S}}$. Mathematically, the features $\mathbf{F}$ in the glaucoma classification subnet can be masked by $\hat{\mathbf{S}}$ as follows:

$$\mathbf{F}' = \mathbf{F} \odot \left\{ (1-\theta) \cdot \hat{\mathbf{S}} \oplus \theta \right\}, \tag{5}$$

where θ (=0.5 in this paper) is a threshold to control the impact of the visualization map. In the above equation, $\odot$ and $\oplus$ represent elementwise multiplication and addition. In the glaucoma classification subnet, the input fundus image is masked with the visualization map in the same way. Finally, in our AG-CNN method, the redundant features irrelevant to glaucoma detection can be inhibited and the pathological area can be highlighted.

E. Weakly supervised loss function

To achieve end-to-end training, we supervise the training process of AG-CNN through attention prediction loss (denoted by Loss_a), feature visualization loss (denoted by Loss_f) and glaucoma classification loss (denoted by Loss_c), as shown in Figure 6. In our LAG database, the glaucoma label l ($\in \{0,1\}$) is available for each of all 11,760 samples, while 5,824 samples are further labeled with the attention maps $\mathbf{A}$ (with its elements $A(i,j) \in [0,1]$), viewed as the GT in the loss function. We assume that $\hat{l}$ ($\in \{0,1\}$) and $\hat{\mathbf{A}}$ (with its elements $\hat{A}(i,j) \in [0,1]$) are the predicted glaucoma label and attention map, respectively.

Following [20], we utilize the Kullback-Leibler (KL) divergence function to measure the difference between two attention maps. The attention prediction subnet is trained in a weakly supervised manner with the attention prediction loss Loss_a. Here, Loss_a is composed of 2 parts, Loss_as and Loss_an, which stand for the loss with and without the supervision of GT attention map $\mathbf{A}$, respectively. The supervised attention prediction loss is represented by

$$\text{Loss}_\text{as} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij} \log\left(\frac{A_{ij}}{\hat{A}_{ij}}\right), \tag{6}$$

where I and J are the length and width of attention maps. Note that Loss_as is calculated by the samples labeled with GT attention maps.

Moreover, the unsupervised attention prediction loss Loss_an is applied to encourage the invariance of symmetry and cropping such that the accuracy of both attention prediction and glaucoma classification can be improved. We first transfer the original fundus images by flipping and cropping. Specifically, the operation of flipping means flipping the original image along horizontal axis. Mathematically, $\mathbf{I}_f$ is calculated by

$$\mathbf{I}_f = T_f(\mathbf{I}), \tag{7}$$

$$T_f(I_{i,j}) = I_{W-i,j}, \tag{8}$$

where $T_f(\cdot)$ represents the functions of flipping, $I_{i,j}$ is the element of $\mathbf{I}$, and W is the width of $\mathbf{I}$. The operation of cropping means cropping the center region of the original fundus image symmetrically and resizing the cropped image back to the original size by nearest neighbor interpolation. Mathematically, $\mathbf{I}_c$ is calculated by

$$\mathbf{I}_c = T_c(\mathbf{I}), \tag{9}$$

$$T_c(I) = R_{W,H}\{I_{\lfloor \frac{(W-p\cdot W)}{2} \rfloor : \lfloor \frac{(W+p\cdot W)}{2} \rfloor, \lfloor \frac{(H-p\cdot H)}{2} \rfloor : \lfloor \frac{(H+p\cdot H)}{2} \rfloor}\}, \tag{10}$$

where $T_c(\cdot)$ represents the function of cropping; $R_{W,H}\{\cdot\}$ is the function of resizing an image to the dimension of $W \times H$. In addition, p is the cropping ratio; W and H are the width and height of image $\mathbf{I}$. Then, two forms of fundus images, i.e., the original fundus image ($\mathbf{I}$) and flipped or cropped fundus images ($\mathbf{I}_f$ or $\mathbf{I}_c$), are input into the attention prediction subnet, and the output predicted attention map of $\mathbf{I}$ is denoted as $\hat{\mathbf{A}}$. Similarly, the attention maps of $\mathbf{I}_f$ and $\mathbf{I}_c$ are $\hat{\mathbf{A}}_f$ ($\hat{A}_f(i,j) \in [0,1]$) and $\hat{\mathbf{A}}_c$ ($\hat{A}_c(i,j) \in [0,1]$). According to the invariance of symmetry and cropping, $T_f(\hat{\mathbf{A}})$ ($\hat{A}'_f(i,j) \in [0,1]$) and $T_c(\hat{\mathbf{A}})$ ($\hat{A}'_c(i,j) \in [0,1]$) should be similar to $\hat{\mathbf{A}}_f$ and $\hat{\mathbf{A}}_c$. To encourage this invariance, Loss_an is calculated as follows,

$$\begin{aligned} \text{Loss}_\text{an} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} [\hat{A}'_f(i,j) \log\left(\frac{\hat{A}'_f(i,j)}{\hat{A}_f(i,j)}\right) + \\ \hat{A}'_c(i,j) \log\left(\frac{\hat{A}'_c(i,j)}{\hat{A}_c(i,j)}\right)]. \end{aligned} \tag{11}$$

Note that Loss_an does not require the GT attention map, and therefore all the samples in our LAG database can be applied for optimizing Loss_an. Finally, the weakly supervised attention prediction loss is composed of

$$\text{Loss}_\text{a} = \text{Loss}_\text{as} + \text{Loss}_\text{an}. \tag{12}$$

Furthermore, the pathological area localization subnet and glaucoma classification subnet are all supervised by the glaucoma label l based on the cross-entropy function, which measures the distance between the predicted label $\hat{l}$ and its corresponding GT label l. Mathematically, Loss_c is calculated as follows:

$$\text{Loss}_c = -l \log\left(\frac{1}{1+e^{-\hat{l}_c}}\right) - (1-l) \log\left(1 - \frac{1}{1+e^{-\hat{l}_c}}\right), \tag{13}$$

where $\hat{l}_c$ represents the predicted label from the glaucoma classification subnet. A similar approach is used to calculate Loss_f, which replaces $\hat{l}_c$ by $\hat{l}_f$ in (13). Finally, the overall loss is the linear combination of Loss_a, Loss_f and Loss_c:

$$\text{Loss} = \alpha \cdot \text{Loss}_\text{a} + \beta \cdot \text{Loss}_\text{f} + \gamma \cdot \text{Loss}_\text{c}, \tag{14}$$

where α, β and γ are the hyper-parameters for balancing the trade-off among attention loss, visualization loss and classification loss. At the beginning of training AG-CNN, we set $\alpha \gg \beta = \gamma$ to accelerate the convergence of attention prediction subnet. Then, we set $\alpha \ll \beta = \gamma$ to minimize the feature visualization loss and the classification loss, thereby achieving the convergence of prediction. Given the loss function of (14), our AG-CNN model can be end-to-end trained for glaucoma detection and pathological area localization.

V. Experiments and Results

A. Settings

In this section, the experiment results are presented to validate the performance of our method in glaucoma detection and pathological area localization. In our experiment, the 11,760 fundus images in our LAG database are randomly divided into training (10,928 images) and validation (832 images) sets. The training set of 10,928 images is further augmented by 3 times, via cropping each fundus image into 3 sizes, i.e., 30%, 50%

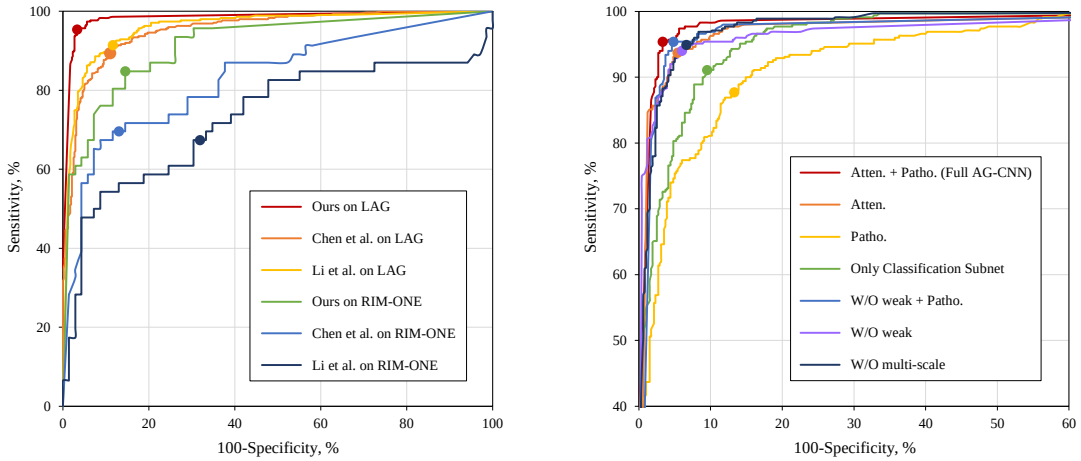

Fig. 7. Comparison of ROC curves among different methods. (Left): Testing on the LAG and RIM-ONE database. (Right): The result of the ablation experiment.

and 75% of the initial dimension. Note that there is no overlap for either subject or eye both in the training and validation sets. To test the generalization ability of our AG-CNN, we further validate the performance of our method on another public database, RIM-ONE [12]. Before inputting to AG-CNN, the RGB channels of the fundus images are all resized to 224×224, following [40], [43], [19] to save computational complexity. Note that we found that higher resolution input may incur under-fitting problem. In training AG-CNN, the gray attention maps are downsampled to 112×112 with their values normalized to be $0 \sim 1$. The cropping ratio p in equation 10 is set to be 0.5. The loss function of (14) for training the AG-CNN model is minimized through the gradient descent algorithm with the Adam optimizer [25]. The initial learning rate is 1×10^{-5}. The learning rate is tuned over the training set in order to obtain appropriate accuracy at a fast convergence speed. The hyper-parameters of α, β and γ in (10) were tuned over our training set, in order to obtain appropriate accuracy at fast convergence speed. Specifically, We first set $\alpha = 20$ and $\beta = \gamma = 1$ in (14) until the loss of the attention prediction subnet converges, and then we set $\alpha = 1$ and $\beta = \gamma = 10$ for focusing on the feature visualization loss and glaucoma classification loss. Additionally, the batch size is set to be 8.

Given the trained AG-CNN model, our method is evaluated and compared with two other state-of-the-art glaucoma detection methods [6] and [29] in terms of different metrics. Specifically, the metrics of accuracy, sensitivity and specificity are measured according to [6] , and the $\mathrm{F}_\beta\mathrm{-score}$ is calculated by

$$\mathrm{F}_\beta\mathrm{-score} = \frac{(1+\beta^2)\cdot \mathrm{TP}}{(1+\beta^2)\cdot \mathrm{TP} + \beta^2 \cdot \mathrm{FN} + \mathrm{FP}}, \qquad (15)$$

where TP, FP and FN are the numbers of true positive glaucoma, false positive glaucoma and false negative glaucoma, respectively. In the above equation, β is the hyper-parameter balancing the trade-off between sensitivity and specificity, and it is set to 2 since the sensitivity is more important in medical diagnosis. In addition, the receiver operating characteristic curve (ROC) and area under ROC (AUC) are also evaluated for comparing the performance of glaucoma detection. All experiments are conducted on the same computer with an Intel(R) Core(TM) i7-4770 CPU@3.40GHz, 32 GB RAM and a single Nvidia GTX 1080 GPU. Benefiting from the GPU, it takes around 15 hours to train our AG-CNN model with 25 epochs. Also, our method is able to detect glaucoma in 30 fundus images per second, comparing to 83 and 21 images per second for [6] and [29].

TABLE IV

PERFORMANCE OF THREE METHODS FOR GLAUCOMA DETECTION OVER OUR LAG VALIDATION SET AND THE TEST SET OF RIM-ONE DATABASE.

Database	Method	Accuracy	Sensitivity	Specificity	AUC	$\mathrm{F}_2\mathrm{-score}$
LAG	Chen et al.[6]	89.2%	89.4%	89.0%	0.953	0.886
	Li et al.[29]	89.7%	91.4%	88.4%	0.960	0.901
	Ours	**96.2%**	**95.4%**	**96.7%**	**0.983**	**0.954**
RIM-ONE	Chen et al.[6]	80.0%	69.6%	**87.0%**	0.831	0.711
	Li et al.[29]	67.8%	67.4%	68.1%	0.731	0.654
	Ours	**85.2%**	**84.8%**	85.5%	**0.916**	**0.837**

B. Evaluation on glaucoma detection

In this section, we compare the glaucoma detection performance of our AG-CNN method with two other methods [6], [29]. Note that the models of other methods are retrained over the whole training set (10,928) of our LAG database for a fair comparison. Table IV lists the results of accuracy, sensitivity, specificity, $\mathrm{F}_2\mathrm{-score}$ and AUC. As shown in Table IV, our AG-CNN method achieves 96.2%, 95.4% and 96.7% in terms of accuracy, sensitivity and specificity, respectively, which are considerably better than the other two methods. Then, the $\mathrm{F}_2\mathrm{-score}$ of our method is 0.954, while [6] and [29] only have $\mathrm{F}_2\mathrm{-scores}$ of 0.886 and 0.901. The above results indicate that our AG-CNN method significantly outperforms other two methods in terms of accuracy, sensitivity, specificity and $\mathrm{F}_2\mathrm{-score}$.

Then, Figure 7 (left) plots the ROC curves of our and other methods for visualizing the trade-off between sensitivity and specificity. As shown in this figure that the ROC curve of our method is closer to the upper-left corner when compared with the other two methods. This result means that the sensitivity of our method is always higher than those of [6], [29] at the same specificity. We further quantify the ROC performance of the three methods through AUC. The AUC results are also reported in Table IV. As shown in this table, our method has a larger AUC than the other two compared methods. In summary, we can conclude that our method performs better in all metrics than [6], [29] in glaucoma detection.

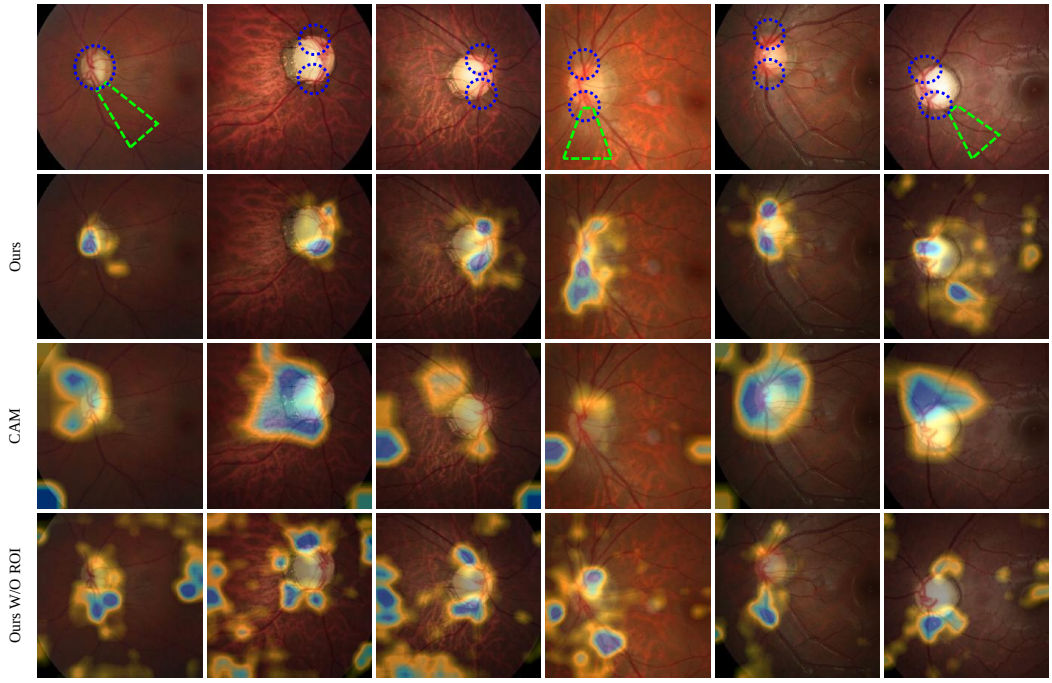

Fig. 8. Comparison of pathological area localization results for glaucoma detection. (1^{st} row): The pathological areas located by ophthalmologists. Optic cup and disc are labeled in blue and the regions of retinal nerve fiber layer defect are labeled in green. (2^{nd} row): The result of our method. (3^{rd} row): The result of the CAM-based method. (4^{th} row): The result of the ablation experiment.

TABLE V
TESTING RESULTS ON THE MULTI-DISEASE SET.

Category	Myopia	PLC	Other disease	All validation set
Accuracy	92.4%	90.0%	93.8%	96.2%
Sensitivity	91.1%	/	92.3%	95.4%
Specificity	94.4%	90.0%	94.7%	96.7%

To evaluate the generalization ability, we further compare the performance of glaucoma detection by our method with those of the other 2 methods [6], [29] on the RIM-ONE database [12]. Note that we fine-tuned the models of AG-CNN, [6] and [29] on the training images of the RIM-ONE database. To our best knowledge, there is no other public database of fundus images for glaucoma. The results are also shown in Table IV and Figure 7 (left). As shown in Table IV, all metrics of our AG-CNN method over the RIM-ONE database are above 0.83, despite slightly smaller than the results over our LAG database. The performance of our method is considerably better than other two methods (except the specificity of [6]). Note that the metric of sensitivity is more important than that of specificity in glaucoma detection, because other indicators, e.g., intra-ocular pressure and the field of vision, can be further used for confirming the diagnosis of glaucoma. This result implies that our method has a high generalization ability.

More importantly, Table IV and Figure 7 (left) show that our AG-CNN method performs significantly better than other methods especially in terms of sensitivity. In particular, the performance of [29] severely degrades, as it incurs the over-fitting issue. To summarize, our AG-CNN method performs well in terms of generalization ability, considerably better than other state-of-the-art methods [6], [29].

In order to validate the influence of other disease on glaucoma detection, we further collect the labels of pathologic myopia, physiologic large cupping (PLC), and other fundus disease. Specifically, 5,824 fundus images in our LAG database are further annotated with multiple disease labels, in which 657 fundus images are with pathologic myopia, 66 fundus images are with physiologic large cupping and 212 fundus images have other fundus diseases. We further test the accuracy, sensitivity and specificity of glaucoma detection over the validation set of these multi-disease images. The results are shown in Table V. The glaucoma detection accuracy on the sets of myopia, physiologic large cupping and other disease decreases 3.8%, 6.0% and 2.4%, respectively, when comparing the test result over the multi-disease sets with the result over the whole validation set. This indicates that other fundus diseases slightly influence the accuracy of glaucoma detection.

C. Evaluation on attention prediction and pathological area localization

We first evaluate the accuracy of the attention model embedded in our AG-CNN model. The attention maps predicted by our AG-CNN method over the LAG database and RIM-ONE database are visualized in Figure 1 of the supplementary materials. Note that all fundus images with or without GT attention maps are directly input to the attention prediction subnet, for outputting their attention maps. As shown in this figure that the predicted attention maps are close to those of the GT, when validating on our LAG database. The CC between the predicted attention maps and the GT is 0.934 on average, with a variance of 0.0032. The result implies that the attention prediction subnet is able to predict attention maps with high accuracy. This figure also shows that the attention maps can locate the salient optic cup and disc for the RIM-ONE database, in which the scales of fundus images are completely different from those of the LAG database. Thus, our method is robust to the scales of fundus images in predicting attention maps.

TABLE VI
ABLATION RESULTS OVER THE VALIDATION SET OF OUR LAG DATABASE.

Method		Accuracy	Sensitivity	Specificity	AUC	$F_2-score$
Atten.	Patho.					
✓	✓	**96.2%**	**95.4%**	**96.7%**	**0.983**	**0.954**
✓	×	94.2%	93.7%	94.6%	0.976	0.935
×	✓	87.1%	87.7%	86.7%	0.941	0.867
×	×	90.8%	91.1%	90.5%	0.966	0.904
W/O weak	Patho.					
✓	✓	95.3%	95.4%	95.2%	0.975	0.951
✓	×	94.0%	94.0%	94.0%	0.973	0.936
W/O multi-scale		94.0%	94.9%	93.4%	0.981	0.941

* Atten.: attention prediction subnet; Patho.: pathological area localization subnet; W/O weak: training the attention prediction subnet without weakly supervised loss.

In this part, we focus on the performance of pathological area localization. Figure 8 visualizes the located pathological area over the LAG database. Comparing the pathological area with our localization results, Figure 8 shows that our AG-CNN model can accurately locate the areas of the optic cup and disc and the region of RNFL defect, especially for the pathological areas of the upper and lower optic disc edge.

Besides, we calculate the CC values between the located pathological area and the GT attention maps of ophthalmologists, with an average of 0.581 and a variance of 0.028. This result also implies that (1) on the one hand, the pathological area localization results are consistent with the attention maps of ophthalmologists; (2) on the other hand, the located pathological area cannot be completely covered by the attention maps. Moreover, we also compare our attention-based pathological area localization results with a state-of-the-art method [15], which is based on the CAM model [55]. The results of [15] are shown in the 3^{rd} row of Figure 8. As shown, it can roughly highlight the ROI but cannot pinpoint the tiny pathological area, e.g., the upper and lower edges of the optic disc boundary. In some cases, [15] highlight the boundary of the eyeball, indicating that the CAM-based methods extracted some unuseful features (i.e., redundancy) for classification. Therefore, the pathological area localization in our approach is effective and reliable, particularly compared to the CAM-based method that does not incorporate human attention.

D. Results of ablation experiments

In our ablation experiments, we first investigate the impact of predicted attention maps for pathological area localization. To this end, we simply remove the attention prediction subnet, and then we compare the pathological localization results with and without the predicted attention maps. The results are shown in Figure 8. As shown, the pathological area can be effectively localized by using the attention maps. In contrast, the located pathological area distributes over the whole fundus image once the attention maps are not incorporated. Therefore, the above results verify the effectiveness and necessity of predicting the attention maps for pathological area localization in our AG-CNN approach.

Next, we assess the impact of the predicted attention map and the located pathological area on the performance of glaucoma detection. To this end, we simply remove the attention prediction subnet and pathological area localization subnet, respectively, for classifying the binary labels of glaucoma. The ablation results are reported in Table VI. As shown in this table, the introduction of the predicted attention map and located pathological area can improve the accuracy, sensitivity, specificity and $F_2-score$ by 5.4%, 4.3%, 6.2% and 5.0%, respectively. However, when the attention prediction subnet is removed, the performance of only embedding the pathological area localization subnet is even worse, with an AUC reduction of 0.025. This result indicates the necessity of our attention prediction subnet for pathological area localization and glaucoma detection.

We further evaluate the impact of weakly supervised training manner of the attention prediction subnet. To this end, we remove the unsupervised Loss_{an} from Equation (12). Consequently, the training process of the attention prediction subnet is fully supervised. The results are also shown in Table VI. As shown, the introduction of the weakly supervised training manner can improve the performance of glaucoma classification in terms of accuracy by 0.9% and 0.2%, with and without the pathological area localization subnet, respectively. Similar results can be found for specificity and AUC. Finally, we show the effectiveness of the proposed multi-scale block in AG-CNN, via replacing it by the default conventional shortcut connection in residual network [19]. The results are also tabulated in Table VI. We can see that the multiscale block can enhance the performance of glaucoma detection. In summary, our ablation experiments show that the attention prediction subnet, pathological area localization subnet, weakly supervised training manner and multiscale block are effective for glaucoma detection.

VI. CONCLUSION

In this paper, we have proposed a new deep learning method, named AG-CNN, for automatic glaucoma detection and pathological area localization upon fundus images. Our AG-CNN model is composed of the subnets of attention prediction, pathological area localization and glaucoma classification and is trained in weakly supervised manner. As such, glaucoma could be detected using the deep features highlighted by the visualized maps of pathological areas, based on the predicted attention maps. For training the AG-CNN model, we established the LAG database with 11,760 fundus images labeled as positive or negative glaucoma. A total of 5,824 images in our LAG database have the attention map on glaucoma detection obtained from 4 ophthalmologists. The experiment results show that the predicted attention maps improve the performance of glaucoma detection and pathological area localization in our AG-CNN method, far better than other state-of-the-art methods.

REFERENCES

[1] U. R. Acharya, S. Dua, X. Du, S. Vinitha Sree, and C. K. Chua. Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE TITB*, 15(3):449–455, 2011.
[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2015.
[3] R. R. A. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, and K. Naidoo. Causes of vision loss worldwide, 1990c2010: a systematic analysis. *Lancet Glob Health*, 1(6):e339–e349, 2013.
[4] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark.
[5] H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng. Mitosis detection in breast cancer histology images via deep cascaded networks. In *The AAAI Conference on Artificial Intelligence*, pages 1160–1166, 2016.

[6] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu. Glaucoma detection based on deep convolutional neural network. In *IEEE EMBC*, page 715, 2015.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. Ieee, 2009.

[8] S. Dua, U. R. Acharya, P. Chowriappa, and S. V. Sree. Wavelet-based energy features for glaucomatous image classification. *IEEE TITB*, 16(1):80–7, 2012.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[10] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In *MICCAI*, pages 568–576. Springer, 2017.

[11] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE TMI*, 2018.

[12] F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *IEEE CBMS*, pages 1–6, 2011.

[13] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[14] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *MICCAI*, pages 250–258. Springer, 2017.

[15] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In *IEEE ICIP*, pages 2069–2073. IEEE, 2017.

[16] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402, 2016.

[17] N. Harizman, C. Oliveira, A. Chiang, C. Tello, M. Marmor, R. Ritch, and J. M. Liebmann. The isnt rule and differentiation of normal from glaucomatous eyes. *Archives of Ophthalmology*, 124(11):1579, 2006.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE ICCV*, pages 1026–1034, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.

[20] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE ICCV*, pages 262–270, 2015.

[21] G. D. Joshi, J. Sivaswamy, R. Prashanth, and S. Krishnadas. Detection of peri-papillary atrophy and rnfl defect from retinal images. In *International Conference Image Analysis and Recognition*, pages 400–407. Springer, 2012.

[22] C. Jun, L. Jiang, X. Yanwu, Y. Fengshou, D. W. K. Wong, T. Ngan-Meng, T. Dacheng, C. Ching-Yu, A. Tin, and W. Tien Yin. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE TMI*, 32(6):1019–1032, 2013.

[23] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. Mckeown, G. Yang, X. Wu, and F. Yan. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122C1131.e9, 2018.

[24] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM TOCHI*, 24(5):1–40, 2017.

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[26] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, page 436, 2015.

[27] A. Li, J. Cheng, D. W. Wong, J. Liu, A. Li, J. Cheng, D. W. K. Wong, J. Liu, J. Cheng, and D. W. Wong. Integrating holistic and local deep features for glaucoma classification. In *IEEE EMBC*, page 1328, 2016.

[28] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu. Attention based glaucoma detection: A large-scale database and cnn model. *arXiv:1903.10831*, 2019.

[29] Z. Li, Y. He, S. Keel, W. Meng, R. Chang, and M. He. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 2018.

[30] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and F.-F. Li. Thoracic disease identification and localization with limited supervision. *arXiv preprint arXiv:1711.06373*, 2017.

[31] J. Liang, D. R. Williams, and D. T. Miller. Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSAA*, pages 2884–92, 1997.

[32] J. Odstrcilik, R. Kolar, R.-P. Tornow, J. Jan, A. Budai, M. Mayer, M. Vodakova, R. Laemmer, M. Lamos, Z. Kuna, et al. Thickness related textural properties of retinal nerve fiber layer in color fundus images. *Computerized Medical Imaging and Graphics*, 38(6):508–516, 2014.

[33] R. Panda, N. B. Puhan, A. Rao, B. Mandal, D. Padhy, and G. Panda. Deep convolutional neural network-based patch classification for retinal nerve fiber layer defect detection in early glaucoma. *Journal of Medical Imaging*, 5(4):044003, 2018.

[34] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. Mcconnell, G. S. Corrado, L. Peng, and D. R. Webster. Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. *arXiv preprint arXiv:1708.09843*, 2017.

[35] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[37] S. M. Shankaranarayana, K. Ram, K. Mitra, and M. Sivaprakasam. Joint optic disc and cup segmentation using fully convolutional and adversarial networks. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 168–176. Springer, 2017.

[38] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2016.

[39] N. Shibata, M. Tanito, K. Mitsuhashi, Y. Fujino, M. Matsuura, H. Murata, and R. Asaoka. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific reports*, 8(1):14665, 2018.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] A. Singh, M. K. Dutta, M. Parthasarathi, V. Uher, and R. Burget. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *CMPB*, 124(C):108, 2016.

[42] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, 2016.

[44] A. J. Tatham, F. A. Medeiros, L. M. Zangwill, and R. N. Weinreb. Strategies to improve early diagnosis in glaucoma. *Progress in Brain Research*, 221:103, 2015.

[45] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040 : A systematic review and meta-analysis. *Ophthalmology*, 121(11):2081, 2014.

[46] D. Ting, C. Y. Cheung, G. Lim, G. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garciafranco, I. Y. San, and S. Y. Lee. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211, 2017.

[47] T. Y. Wong, F. Alberto, H. Rowena, C. Gemma, and M. Paul. Epidemiology and disease burden of pathologic myopia and myopic choroidal neovascularization: an evidence-based systematic review. *Ophthalmology*, 157(1):9–25.e12, 2014.

[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[49] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang. Learning to detect video saliency with hevc features. *IEEE TIP*, 26(1):369–385, 2017.

[50] M. Xu, L. Jiang, Z. Ye, and Z. Wang. Bottom-up saliency detection with sparse representation of learnt texture atoms. *Pattern Recognition*, 60:348–360, 2016.

[51] L. Yu, X. Yang, C. Hao, J. Qin, and P. A. Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *The AAAI Conference on Artificial Intelligence*, 2017.

[52] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim. Supervising neural attention models for video captioning by human gaze data. In *IEEE CVPR*, pages 2680–29, 2017.

[53] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *IEEE ECCV*, pages 818–833. Springer, 2014.

[54] Q. Zhang, A. Bhalerao, and C. Hutchinson. Weakly-supervised evidence pinpointing and description. In *IPMI*, pages 210–222. Springer, 2017.

[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE CVPR*, pages 2921–2929, 2016.

[56] J. Zilly, J. M. Buhmann, and D. Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *CMIG*, 55:28–41, 2017.