Attention Based Glaucoma Detection: A Large-scale Database and CNN Model

Liu Li†, Mai Xu†‡*, Xiaofei Wang†, Lai Jiang†, Hanruo Liu§,
† School of Electronic and Information Engineering, Beihang University, Beijing, China
‡ Hangzhou Innovation InstituteBeihang University, Hangzhou, Zhejiang, China
§Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Beijing, China
†{liliu1995, maixu, xfwang, jianglai.china}@buaa.edu.cn

Abstract

Recently, the attention mechanism has been successfully applied in convolutional neural networks (CNNs), significantly boosting the performance of many computer vision tasks. Unfortunately, few medical image recognition approaches incorporate the attention mechanism in the CNNs. In particular, there exists high redundancy in fundus images for glaucoma detection, such that the attention mechanism has potential in improving the performance of CNN-based glaucoma detection. This paper proposes an attention-based CNN for glaucoma detection (AG-CNN). Specifically, we first establish a large-scale attention based glaucoma (LAG) database, which includes 5,824 fundus images labeled with either positive glaucoma (2,392) or negative glaucoma (3,432). The attention maps of the ophthalmologists are also collected in LAG database 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. Different from other attention-based CNN methods, the features are also visualized as the localized pathological area, which can advance the performance of glaucoma detection. Finally, the experiment results show that the proposed AG-CNN approach significantly advances state-of-the-art glaucoma detection.

1. Introduction

In recently years, the attention mechanism has been successfully applied in deep learning based computer vision tasks, i.e., object detection [3, 31, 28], image caption [35, 39, 2] and action recognition [30]. 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 [31, 28, 35, 30] are yielded in a self-learned pattern, with other information weakly supervising the attention maps, i.e., the classification labels. On the other hand, [39, 37] utilize the human attention information to guide the DNNs focusing on the region of interest (ROI).

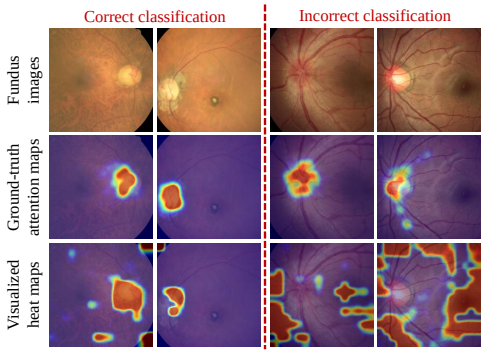

Figure 1. Examples of glaucoma fundus images, attention maps by ophthalmologists in glaucoma diagnosis and visualization results of a CNN model (Bottom) [15] by an occlusion experiment [40]. The Pearson Correlation Coefficient (CC) results between the visualized heat maps and ground-truth ophthalmologist attention maps are 0.33 and 0.14 for correct and incorrect glaucoma classification, respectively.

Redundancy also exists in medical image recognition, interfering the recognition results. In particular, there exists heavy redundancy in fundus images for disease recognition. For example, the pathological areas of fundus images are in the region of optic cup and disc, or its surrounding blood vessel and optic nerve area [25]; other regions such as the boundary of the eye ball are redundant for the medical diagnosis. As shown in Figure 1, glaucoma, an irreversible optic disease, can be correctly detected by a convolutional neural network (CNN) [15], when the visualized heat maps are consistent with the attention maps of ophthalmologists. Otherwise, glaucoma is mislabeled by the CNN model when the visualized heat maps focus on redundant

*Mai Xu is the corresponding author of this paper.

regions. Therefore, it is reasonable to combine the attention mechanism in the CNN model for using fundus images to detect ophthalmic disease.

However, to our best knowledge, there has been no works incorporating the human attention in medical image recognition. This is mainly because there lacks the doctor attention database, which needs the qualified doctors and a special technique of capturing the doctor attention in the diagnosis. As such, in this paper, we first collect a large-scale attention based fundus image database for glaucoma detection (LAG), including 5,824 images with diagnose labels and human attention maps. Based on the real human attention, we propose an attention based CNN method (called AG-CNN) for glaucoma detection based on fundus images.

Although human attention is able to reduce heavy redundancy in fundus images for disease recognition, it may also miss some of the pathological area which is helpful for disease detection. As a result, the existing CNN models have outperformed the doctors in medical image recognition [18, 27, 26]. Thus, we propose to refine the predicted attention maps by incorporating a feature visualization structure for glaucoma detection. As such, the gap between human attention and pathological area can be bridged. In fact, there have been several methods for automatically locating the pathological area [41, 12, 8, 11, 24], based on the class activation mapping model (CAM) [42]. However, these methods cannot locate the pathological area at a small region due to the limitation of its feature size. In this paper, we employ the guided back propagation (BP) method [33] 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 regions for glaucoma detection.

The main contributions of this paper are: (1) We establish a LAG database with 5,824 fundus images, along with their labels and attention maps. (2) We propose incorporating the attention maps in AG-CNN, such that the redundancy can be removed from fundus images for glaucoma detection. (3) We develop a new architecture of AG-CNN, which visualizes the CNN feature maps for locating pathological area and then classifies binary glaucoma.

2. Medical Background

The recent success of deep learning methods has benefitted medical diagnosis [7, 4, 38], especially for automatically detecting oculopathy in fundus images [13, 10, 34]. Specifically, [13, 10] worked on classification of diabetic retinopathy using the CNN models. [34] further proposed deep learning systems for multi-ophthalmological diseases detection. However, the above works all transfered some classic CNN model for nature image classification to medical image classification, regardless of the characteristic of fundus images.

Glaucoma detection methods can be basically divided into 2 categories, i.e., heuristic methods and deep learning methods. The heuristic glaucoma detection methods extract features based on some image processing techniques [1, 6, 17, 32]. Specifically, [1] extracted the texture features and higher order spectra features for glaucoma detection. [6] used the wavelet-based energy features for glaucoma detection. Both [1, 6] 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 [29, 43, 5, 22, 23]. Specifically, [29, 43] reported their deep learning work on glaucoma detection based on automatic segmentation of optic cup and disc. However, their work assume that only the optical cup and disc are related to glaucoma, lacking end-to-end training. On the other hand, [5] firstly proposed a CNN method for glaucoma detection in an end-to-end mannar. [22] 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 [5, 22] 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. Most recently, a deeper CNN structure has been proposed in [23]. However, the fundus images exist large redundancy irrelevant to glaucoma detection, leading to the low efficiency for [23].

3. Database

3.1. Establishment

In this work, we establish a large-scale attention based glaucoma detection database. Our LAG database contains 5,824 fundus images with 2,392 positive and 3,432 negative glaucoma samples obtained from Beijing Tongren Hospital [1]. Our work is conducted according to the tenets of Helsinki Declaration. As the retrospective nature and fully anonymized usage of color retinal fundus images, we are exempted by the medical ethics committee to inform the patients. Each fundus image is diagnosed by qualified glaucoma specialists, taking the consideration of both morphologic and functional analysis, i.e, intra-ocular pressure, visual field loss and manual optic disc assessment. As a result, the binary labels of positive or negative glaucoma of all fundus images are confirmed, seen as the gold standard.

Based on the above labelled fundus images, we further conduct an experiment to capture the attention regions of the ophthalmologists in glaucoma diagnosis. The experiment is based on an alternative method for eye tracking [19],

[1]The database is available for online access upon request.

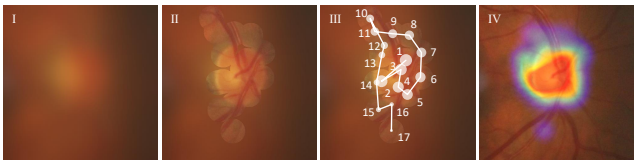

Figure 2. An example of capturing fixations of an ophthalmologist in glaucoma diagnosis. (Left): Original blurred fundus images. (Middle-left): Fixations of the ophthalmologist with cleared regions. (Middle-right): The order of clearing the blurred regions. Note that the size of the white circles represents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

Table 1. CC values of attention maps between one ophthalmologist and the mean of the rest 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}

in which mouse clicks are used by the ophthalmologists to explore 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 the radius of all circle regions is set to 40 pixels, while all fundus images are with 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 of screen height). The order of clearing the blurred regions represents the degree of attention by ophthalmologists, as the GT of the attention map. Once the ophthalmologist is able to diagnose glaucoma with the partly 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 Gaussian filter is set to 25, according to [36]. Figure 2 shows an example of the fixations of one ophthalmologist and attention map of all ophthalmologists for a fundus image.

3.2. Data analysis

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

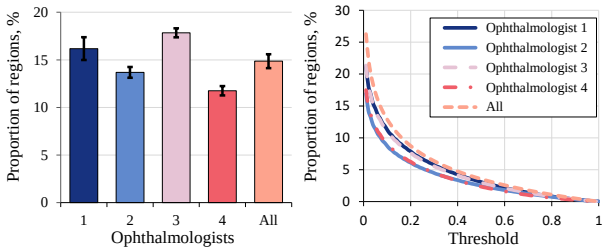

Figure 3. (Left): Proportion of regions in the fundus images cleared by different ophthalmologists for glaucoma diagnosis. (Right): Proportion of regions in attention maps with values being above a varying threshold. Note that the values of the attention maps range from 0 to 1.

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

Analysis: In this analysis, we calculate the Pearson correlation coefficients (CC) of attention maps between one ophthalmologist and the remaining three ophthalmologists. Table 1 reports the CC results averaged over all fundus images in our LAG database. 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 the Gaussian distribution. We can see from Table 1 that the CC values of attention maps between one and the remaining ophthalmologists are all above 0.55, significantly larger than those of the random baseline. This implies that attention exists consistency among ophthalmologists in glaucoma diagnosis. This completes the analysis of *Finding 2*.

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 circle regions for each ophthalmologist, which is averaged over all 5,824 fundus images of our LAG database. We can see that 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 fast 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: *Finding 2* shows that the ROI is small for glaucoma diagnosis, comparing with the whole fundus images. Here, although ROI is small, its scale is various across all the fundus images. Figure 4 visualizes the fixation maps of some fundus images, in which the ROI are with different scales. As shown in Figure 4, the sizes of the optic discs

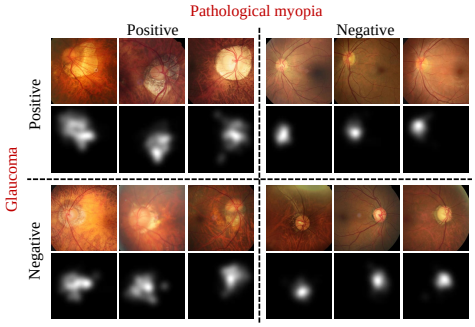

Figure 4. Fundus images with or without glaucoma for both positive and negative pathological myopia.

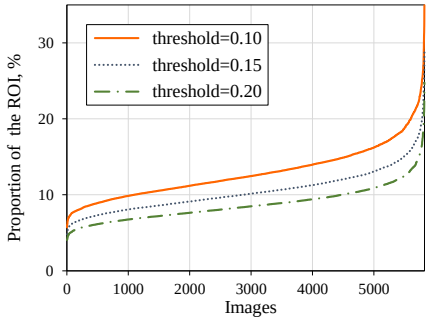

Figure 5. Proportion of ROI above the threshold of 0.10, 0.15 and 0.20, for all of the fundus images in LAG database.

for pathological myopia are considerably larger than others. As such, we use myopia and non-myopia to select samples with various scales of ROI (large or small optic cups). We further find that the images of both positive and negative glaucoma have various-scaled ROI, as demonstrated in Figure 4. 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. We can see that the ROI is at different scale for glaucoma diagnosis. Finally, the analysis of *Finding 3* can be accomplished.

4. Method

4.1. Framework

In this section, we discuss the proposed AG-CNN method. Since *Findings 1* and *2* show that glaucoma diagnosis is highly related to small ROI regions, the attention prediction subnet is developed in AG-CNN for reducing the redundancy of fundus images. In addition, we design a pathological area localization subnet, which is achieved by visualizing the CNN feature map, based on ROI regions of the attention prediction subnet. Based on the pathological area, the glaucoma classification subnet is developed for producing the binary labels of glaucoma, in which the multi-scale features are learned and extracted. The introduction of multi-scale features is according to *Finding 3*.

The framework of AG-CNN is shown in Figure 6, and its components, including multi-scale building block, deconvolutional module and feature normalization, are further demonstrated in Figure 7. 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. In addition, the located pathological area is obtained in our AG-CNN in two 2 stages. In the first stage, the ROI of glaucoma detection is learned from the attention prediction subnet, aiming to predict human attention on diagnosing glaucoma. 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. Finally, the located pathological area is further used to 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 [15], in which the basic module is 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 sped up. The process of training AG-CNN is in an end-to-end manner with three parts of supervision, i.e., attention prediction loss, pathological area localization loss and glaucoma classification loss.

4.2. 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 is represented by the tensor (size: $224 \times 224 \times 3$). Then, the input tensor is fed to one convolutional layer with 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 [15]. Afterwards, the features of 4 hieratical building blocks are processed by feature normalization (FN), the structure of which is shown in Figure 7 (Right). As a result, four $28 \times 28 \times 128$ features are obtained. They are concatenated to form $28 \times 28 \times 512$ deep multi-scale features. Given the deep multi-scale features, a deconvolutional module is applied to generate the gray attention map with the size of $112 \times 112 \times 1$. The structure of the deconvolutional module is also shown in Figure 7 (middle). As shown in this figure, the deconvolutional module is comprised by 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 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

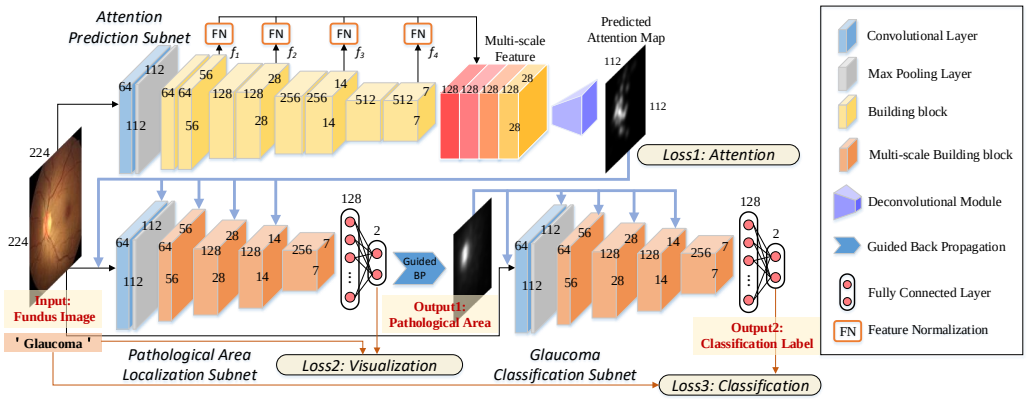

Figure 6. Architecture of our AG-CNN network for glaucoma detection. The sizes of the feature maps and convolutional kernels are shown in this figure.

section.

4.3. 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 useful for glaucoma classification. Moreover, it has been verified that the deep learning methods outperform human in the task of image classification both on nature images [14, 21] and medical images [18, 27, 26]. 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 4.4. Then, the visualization map of pathological area is yielded through guided BP [33] from the output of the fully connection layer to the input RGB channels fundus images. Finally, the visualization map is down-sampled to 112×112 with its values being normalized to $0-1$, as the output of the pathological area localization subnet.

4.4. Glaucoma classification subnet

In addition to the attention prediction subnet and pathological area localization subnet, 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, 4 multi-scale building blocks.

The multi-scale building blocks differ from the traditional building block of [15] from the following aspect. As shown in Figure 7 (Left), 4 channels of convolutional layers C_1, C_2, C_3 and C_4 with different kernel sizes are concatenated to extract multi scale features, comparing with the traditional building block which only has a single convolutional channel. Finally, 2 fully connected layers are applied to output the classification result.

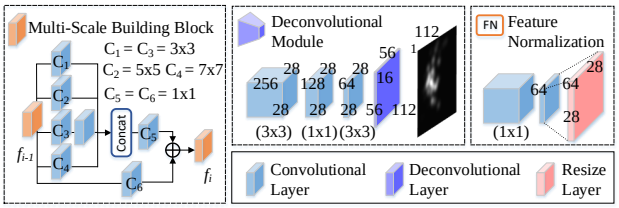

Figure 7. Components of the AG-CNN architecture.

The main difference between the glaucoma classification subnet and the conventional residual network [15] is that the visualization maps of pathological area weight both the input image and extracted features to focus on the ROI. Assume that the visualization map generated by the pathological area localization subnet is $\hat{\mathbf{V}}$. Mathematically, the features $\mathbf{F}$ in the glaucoma classification subnet can be masked by $\hat{\mathbf{V}}$ as follows,

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

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 the element-wise 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.

Table 2. Performance of three methods for glaucoma detection over the test set of our LAG database.

Method	Accuracy	Sensitivity	Specificity	AUC	$\mathrm{F}_2-\mathrm{score}$
Ours	**95.3%**	**95.4%**	**95.2%**	**0.975**	**0.951**
Chen et al.	89.2%	90.6%	88.2%	0.956	0.894
Li et al.	89.7%	91.4%	88.4%	0.960	0.901

Table 3. Performance of three methods for glaucoma detection over the RIM-ONE database.

Method	Accuracy	Sensitivity	Specificity	AUC	$\mathrm{F}_2-\mathrm{score}$
Ours	**85.2%**	**84.8%**	85.5%	**0.916**	**0.837**
Chen et al.	80.0%	69.6%	**87.0%**	0.831	0.711
Li et al.	66.1%	71.7%	62.3%	0.681	0.679

4.5. Loss function

In order to achieve end-to-end training, we supervise the training process of AG-CNN through attention prediction loss (denoted by $\mathrm{Loss_a}$) , feature visualization loss (denoted by $\mathrm{Loss_f}$) and glaucoma classification loss (denoted by $\mathrm{Loss_c}$), as shown in Figure 6. In our LAG database, both the glaucoma label l ($\in \{0, 1\}$) and the attention map $\mathbf{A}$ (with its elements $A_{i,j} \in [0, 1]$) are available for each fundus image, seen 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 [16], we utilize the Kullback-Leibler (KL) divergence function as the human-attention loss $\mathrm{Loss_a}$. Specifically, the human-attention loss is represented by

$$\mathrm{Loss}_a = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij} \log(\frac{A_{ij}}{\hat{A}_{ij}}), \tag{2}$$

where I and J are the length and width of attention maps.

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, $\mathrm{Loss_f}$ is calculated as follows,

$$\mathrm{Loss}_c = l \log(\frac{1}{1 + e^{-\hat{l}_c}}) + (1 - l) \log(1 - \frac{1}{1 + e^{-\hat{l}_c}}), \tag{3}$$

where $\hat{l}_c$ represents the predicted label from the glaucoma classification subnet. Similar way is used to calculate Lossf, which replaces $\hat{l}_c$ by $\hat{l}_f$ in 3.

Finally, the overall loss is the linear combination of $\mathrm{Loss_a}$, $\mathrm{Loss_f}$ and $\mathrm{Loss_c}$:

$$\mathrm{Loss} = \alpha \cdot \mathrm{Loss_a} + \beta \cdot \mathrm{Loss_f} + \gamma \cdot \mathrm{Loss_c}, \tag{4}$$

where α, β and γ are hyper-parameters for balancing the trade-off among attention loss, visualization loss and classification loss. At the begining of training AG-CNN, we choose to set $\alpha \gg \beta = \gamma$ to speed the convergence of attention prediction subnet. Then, we set $\alpha \ll \beta = \gamma$ to minimize the feature visualization loss and the classification loss, thus realizing the convergence of prediction. Given the loss function of (4), our AG-CNN model can be end-to-end trained for glaucoma detection and pathological location.

5. Experiments and Results

5.1. 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 5,824 fundus images in our LAG database are randomly divided into training (4,792 images), validation (200 images) and test (832 images) sets. To test the generalization ability of our AG-CNN, we further validate the performance of our method on another public database RIM-ONE [9]. Before inputting to AG-CNN, the RGB channels of fundus images are all resized to 224×224. In training AG-CNN, the gray attention maps are downsampled to 112×112 with their values normalized to be $0 \sim 1$. The loss function of (4) for training the AG-CNN model is minimized through the gradient descent algorithm with Adam optimizer [20]. The initial learning rate is 1×10^{-5}. We first set $\alpha = 20$ and $\beta = \gamma = 1$ in (4) until the loss of the attention prediction subnet converges, and then set $\alpha = 1$ and $\beta = \gamma = 10$ for focusing on the feature visualization loss and glaucoma classification loss. Additionally, 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 [5, 23], in terms of different metrics. Specifically, the metrics of sensitivity and specificity are defined as follows,

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{6}$$

where $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}$ and FN are the numbers of the true positive glaucoma, true negative glaucoma, false positive glaucoma and false negative glaucoma, respectively. Based on TP, FP and FN, the $\mathrm{F}_\beta-\mathrm{score}$ is calculated by

$$\mathrm{F}_\beta-\mathrm{score} = \frac{(1 + \beta^2) \cdot \mathrm{TP}}{(1 + \beta^2) \cdot \mathrm{TP} + \beta^2 \cdot \mathrm{FN} + \mathrm{FP}}. \tag{7}$$

In the above equation, β is the hyper-parameter balancing the trade-off between sensitivity and specificity, and it is set to 2 as the sensitivity is more important in medical diagnosis. In addition, 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 a computer with an Intel(R)

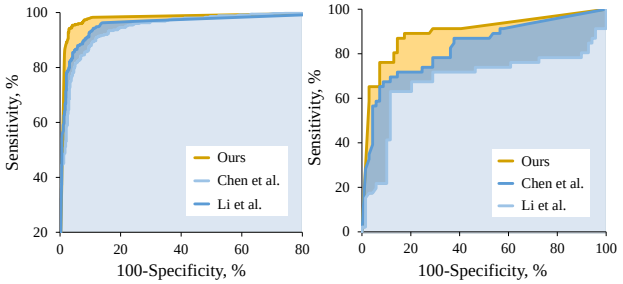

Figure 8. Comparison of ROC curves among different methods. (Left): Testing on our LAG testing set. (Right): Testing on RIM-ONE database.

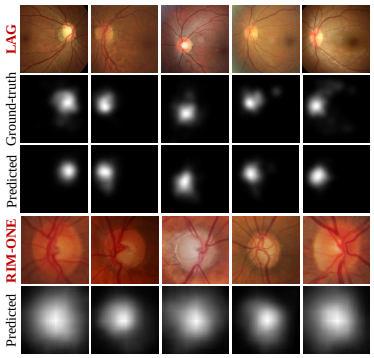

Figure 9. Attention maps predicted by AG-CNN ramdomly selected from the test dataset. The fundus images are from our LAG (upper) and RIM-ONE (lower) database. Note that the RIM-ONE database has not the GT of the attention map.

Core(TM) i7-4770 CPU@3.40GHz, 32GB RAM and a single Nvidia GeForce GTX 1080 GPU. Benefiting from the GPU, our method is able to detect glaucoma of 30 fundus images per second, and it is comparable to 83 and 21 fundus images per second for [5] and [23].

5.2. Evaluation on glaucoma detection

In this section, we compare the glaucoma detection performance of our AG-CNN method with two other methods [5, 23]. Note that the models of other methods are retrained over our LAG database for fair comparison. Table 2 lists the results of accuracy, sensitivity, specificity, $\mathrm{F}_2-\mathrm{score}$ and AUC. As seen in Table 2, our AG-CNN method achieves 95.3%, 95.4% and 95.2% in terms of accuracy, sensitivity and specificity, respectively, which are considerably better than other two methods. Then, the $\mathrm{F}_2-\mathrm{score}$ of our method is 0.951, while [5] and [23] only have $\mathrm{F}_2-\mathrm{scores}$ of 0.894 and 0.901. The above results indicate that our AG-CNN method significantly outperforms other two methods in all metrics.

In addition, Figure 8 (Left) plots the ROC curves of our and other methods, for visualizing the trade-off between sensitivity and specificity. We can see from this figure that the ROC curve of our method is closer to the upper-left corner, when comparing with other two methods. This means that the sensitivity of our method is always higher than those of [5, 23] at the same specificity. We further quantify ROC performance of three methods through AUC. The AUC results are also reported in Table 2. As shown in this table, our method has larger AUC than other two compared methods. In summary, we can conclude that our method performs better in all metrics than [5, 23] in glaucoma detection.

To evaluate the generalization ability, we further compare the performance of glaucoma detection by our method with other 2 methods [5, 23] on the RIM-ONE database [9]. To our best knowledge, there is no other public database of fundus images for glaucoma. The results are shown in Table 3 and Figure 8 (Right). As shown in Table 3, 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 specificity of [23]). It is worth mentioning that the metric of sensitivity is more important than that of specificity in glaucoma detection, as other indicators, e.g., intra-ocular pressure and the field of vision, can be further used for confirming the diagnosis of glaucoma. This implies that our method has high generalization ability.

More importantly, Table 3 and Figure 8 (Right) show that our AG-CNN method performs significantly better than other methods especially in terms of sensitivity. In particular, the performance of [23] severely degrades, as incurring the over-fitting issue. In a word, our AG-CNN method performs well in the generalization ability, considerably better than other state-of-the-art methods [5, 23].

5.3. Evaluation on attention prediction and pathological area localization

We first evaluate the accuracy of the attention model embedded in our AG-CNN model. Figure 9 visualizes the attention maps predicted by our AG-CNN method over the LAG database and RIM-ONE database. We can see from this figure that the predicted attention maps of AG-CNN are close to those of GT, when testing 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. This implies the attention prediction subnet of AG-CNN is able to predict attention maps with high accuracy. We can further see from Figure 9 that the attention maps locate the salient optic cup and disc for the RIM-ONE database, in which the scales of fundus images are totally different from those of LAG database. Thus, our method is robust to the scales of fundus images in predicting attention maps.

Then, we focus on the performance of pathological area localization. Figure 10 visualizes the located pathological area over the LAG database. Comparing the GT pathological area with our localization results, we can see from Figure 10 that our AG-CNN model can accurately located the

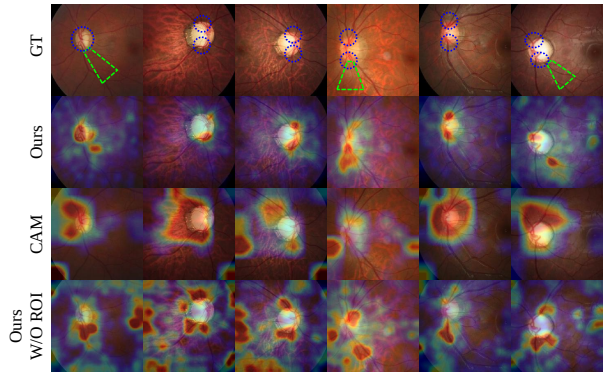

Figure 10. Comparison of pathological area localization results for glaucoma detection. (1^{st} row): The pathological areas located by ophthalmologists. Optic cup and disc are labeled in blue and the regions of retinal nerve fiber layer defect are labeled in green. (2^{nd} row): The result of our method. (3^{rd} row): The result of CAM based method. (4^{th} row): The result of ablation experiment.

areas of optic cup and disc and the region of retinal nerve fiber layer defect, especially for the pathological areas of the upper and lower optic disc edge.

Besides, we calculate the CC 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 also implies that (1) on one hand, the pathological area localization results are consistent with the attention maps of ophthalmologists; (2) on the other hand, the 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-art method [12], which is based on the CAM model [42]. The results of [12] are shown in the 3^{rd} row of Figure 10. We can see that it can roughly highlight the ROI but cannot pinpoint the tiny pathological area, e.g., the upper and lower edge of the optic disc boundary. In some cases, [12] 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, especially compared to the CAM based method that does not incorporate human attention.

5.4. Results of ablation experiments

In our ablation experiments, we first illustrate the impact of predicted attention maps for located pathological area. To this end, we simply remove the attention prediction subnet, and then compare the pathological localization results with and without predicted attention maps. The results are shown in Figure 10. We can see that 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.

Table 4. Ablation results over the test set of our LAG database. APS represents the attention prediction subnet. PAL represents the pathological area localization subnet.

Method	Accuracy	Sensitivity	Specificity	AUC	$\mathrm{F}_2-\mathrm{score}$
Full AG-CNN	**95.3%**	**95.4%**	**95.2%**	**0.975**	**0.951**
W APS W/O PAL	94.0%	94.0%	94.0%	0.973	0.936
W/O APS W PAL	87.1%	87.7%	86.7%	0.941	0.867
W/O APS W/O PAL	90.8%	91.1%	90.5%	0.966	0.904
W/O multi-scale block	92.2%	92.0%	92.3%	0.974	0.915

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 of AG-CNN, respectively, for classifying the binary labels of glaucoma. The results are shown in Table 4. As seen in this table, the introduction of both the predicted attention map and located pathological area can improve accuracy, sensitivity, specificity and $\mathrm{F}_2-\mathrm{score}$ by 4.5%, 4.3%, 4.7% and 4.7%. However, the performance of only embedding the pathological area localization subnet and without the attention prediction subnet is even worse than removing them both. It verifies the necessity of our attention prediction subnet for pathological area localization and glaucoma detection.

Hence, the attention prediction subnet and pathological area localization subnet are able to improve the performance of glaucoma detection in AG-CNN. Additionally, 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 [15]. The results are also shown in Table 4. We can see that the multi-scale block can also enhance the performance of glaucoma detection.

6. 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. 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 5,824 fundus images labeled with either positive or negative glaucoma, along with their attention maps on glaucoma detection. The experiment results showed that the predicted attention maps significantly improve the performance of glaucoma detection and pathological area localization in our AG-CNN method, far better than other state-of-the-art methods.

7. Acknowledgement

This work was supported by BMSTC under Grants Z181100001918035.

References

[1] U. R Acharya, S Dua, Xian Du, S Vinitha Sree, and Chua Kuang Chua. Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):449–455, 2011.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2015.

[4] Hao Chen, Qi Dou, Xi Wang, Jing Qin, and Pheng Ann Heng. Mitosis detection in breast cancer histology images via deep cascaded networks. In *The AAAI Conference on Artificial Intelligence*, pages 1160–1166, 2016.

[5] Xiangyu Chen, Yanwu Xu, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Glaucoma detection based on deep convolutional neural network. In *Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE*, page 715, 2015.

[6] S Dua, U. R. Acharya, P Chowriappa, and S. V. Sree. Wavelet-based energy features for glaucomatous image classification. *IEEE Transactions on Information Technology in Biomedicine*, 16(1):80–7, 2012.

[7] 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.

[8] Xinyang Feng, Jie Yang, Andrew F Laine, and Elsa D Angelini. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer, 2017.

[9] 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 *International Symposium on Computer-Based Medical Systems*, pages 1–6, 2011.

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

[11] Zongyuan Ge, Sergey Demyanov, Rajib Chakravorty, Adrian Bowling, and Rahil Garnavi. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–258. Springer, 2017.

[12] Waleed M Gondal, Jan M Köhler, René Grzeszick, Gernot A Fink, and Michael Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 2069–2073. IEEE, 2017.

[13] 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.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[17] Ashish Issac, M. Partha Sarathi, and Malay Kishore Dutta. *An adaptive threshold based image processing technique for improved glaucoma detection and classification*. Elsevier North-Holland, Inc., 2015.

[18] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex Mckeown, Ge Yang, Xiaokang Wu, and Fangbing Yan. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122C1131.e9, 2018.

[19] Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction*, 24(5):1–40, 2017.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[22] A. Li, J. Cheng, D. W. Wong, J. Liu, Annan Li, Jun Cheng, Damon Wing Kee Wong, Jiang Liu, J. Cheng, and D. W. Wong. Integrating holistic and local deep features for glaucoma classification. In *Engineering in Medicine and Biology Society (EMBC), 38th Annual International Conference of the IEEE*, page 1328, 2016.

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

[24] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Fei-Fei Li. Thoracic disease identification and localization with limited supervision. *arXiv preprint arXiv:1711.06373*, 2017.

[25] J. Liang, D. R. Williams, and D. T. Miller. Supernormal vision and high-resolution retinal imaging through adaptive optics. *Journal of the Optical Society of America A Optics Image Science & Vision*, 14(11):2884–92, 1997.

[26] Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. Mcconnell, Greg S. Corrado, Lily Peng, and Dale R. Webster. Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. *arXiv preprint arXiv:1708.09843*, 2017.

[27] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, pages 91–99, 2015.

[29] Sharath M Shankaranarayana, Keerthi Ram, Kaushik Mitra, and Mohanasankar 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.

[30] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2016.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Anushikha Singh, Malay Kishore Dutta, M. Parthasarathi, Vaclav Uher, and Radim Burget. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Computer Methods & Programs in Biomedicine*, 124(C):108, 2016.

[33] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[34] Dsw Ting, C. Y. Cheung, G. Lim, Gsw Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garciafranco, Iy Yeo 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.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[36] Mai Xu, Lai Jiang, Xiaoyan Sun, Zhaoting Ye, and Zulin Wang. Learning to detect video saliency with hevc features. *IEEE Transactions on Image Processing*, 26(1):369–385, 2017.

[37] Mai Xu, Chen Li, Yufan Liu, Xin Deng, and Jiaxin Lu. A subjective visual quality assessment method of panoramic videos. In *2017 IEEE International Conference on Multimedia and Expo*, pages 517–522. IEEE, 2017.

[38] Lequan Yu, Xin Yang, Chen Hao, Jing Qin, and Pheng Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *The AAAI Conference on Artificial Intelligence*, 2017.

[39] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2680–29, 2017.

[40] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. pages 818–833, 2014.

[41] Qiang Zhang, Abhir Bhalerao, and Charles Hutchinson. Weakly-supervised evidence pinpointing and description. In *International Conference on Information Processing in Medical Imaging*, pages 210–222. Springer, 2017.

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[43] Julian Zilly, Joachim M. Buhmann, and Dwarikanath Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics*, 55:28–41, 2017.