IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2019, accepted May 11, 2019, date of publication May 17, 2019, date of current version May 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917508

Optic Disc and Cup Segmentation Based on Deep Convolutional Generative Adversarial Networks

YUN JIANG, NING TAN, AND TINGTING PENG

College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China

Corresponding author: Ning Tan (tanning2315@126.com)

This work was supported in part by the National Natural Science Foundation Funding Project of China under Grant 61163036, in part by the Program of NSFC Financing for Natural Science Fund in 2016 under Grant 1606RJZA047, in part by the Special Fund Project of Basic Scientific Research Operating Expenses of Institutes and Universities in Gansu in 2012, in part by the Institutes and Universities Graduate Tutor Project in Gansu under Grant 1201-16, and in part by the Third Period of the Key Scientific Research Project of Knowledge and Innovation Engineering of the Northwest Normal University under Grant nwnu-kjcxgc-03-67.

ABSTRACT Glaucoma is a chronic eye disease that causes loss of vision and it is irreversible. Accurate segmentation of optic disc and optic cup is a basic step in screening glaucoma. The most existing deep convolutional neural network (DCNN) methods have insufficient feature information extraction, and hence they are susceptible to pathological regions and low-quality images, with have poor ability to restore context information. Finally, the accuracy of the model segmentation is low. In this paper, we propose GL-Net, a multi-label DCNN model that combines the generative adversarial networks. GL-Net consists of two network structures including a Generator and a Discriminator. In the Generator, we use skip connection to promote the fusion of low-level feature information and high-level feature information, which alleviates the difficulty of restoring detailed feature information during upsampling, and reduces the downsampling factor, effectively alleviating excessive feature information loss. In the loss function, we add the L_1 distance function and the cross-entropy function to prevent the mode collapse when the model is trained, which makes the segmentation result more accurate. We use transfer learning and data augmentation to alleviate the problem of insufficient data and over-fitting of the model during training. Finally, GL-Net was verified on DRISHTI-GS1 dataset. The experimental results show that GL-Net outperforms some state-of-the-art method, such as M-Net, Stack-U-Net, RACE-net, and BCRF in terms of $F1$ and boundary distance localization error (BLE). Particularly, in the optic cup segmentation, GL-Net outperforms RACE-net by 3.5 % and 4.16 pixels in terms of $F1$ and BLE, respectively.

INDEX TERMS Deep learning, optic disc segmentation, optic cup segmentation, deep convolutional neural networks, generative adversarial networks.

I. INTRODUCTION

Glaucoma is one of the major causes of blindness in the eye, and the loss of vision caused by glaucoma cannot be reversed. In clinical practice, it takes a lot of human power and cost for the clinician to perform manual evaluation, and it is also not suitable for batch screening. Therefore, exploring an automatic screening method has been widely studied [1]–[18], [20]. The main segmentation techniques include template-based methods [2], [3], variation level set [4], [5], boundary detection [6], [7], hand-crafted visual feature approach [8]–[11], and deep learning segmentation method [12]–[18].

In the segmentation of optic disc and optic cup, the template-based method [2], [3] is first proposed to obtain the boundary. In [2], a circular Hough transform is performed using morphological and edge detection techniques to obtain an approximate boundary of optic disc. In [3], a circular transformation is used to simultaneously capture the circular shape of optic disc and optic cup and image changes on the boundary, and to further accurately locate the center and boundary by evaluating the pixel of the maximum variation along all radial line segment of the retinal pixel. In these methods, precise boundary of optic disc and optic cup cannot be obtained. In anatomical, glaucoma experts considered that vessel bends

The associate editor coordinating the review of this manuscript and approving it for publication was Ke Gu.

at the boundary of the cup are related to optic cup. Therefore, a similar theory was used in [4], proposing a multi-stage strategy to derive a reliable subset called *r*-bend vessels, followed by local spline fitting to derive the desired optic cup boundary. The morphological features known as kink in optic disc are automatically detected in [5] and the non-stereoscopic retinal image is used to determine the optic cup boundary. However, the boundaries between optic disc and optic cup are not obvious, and it is still impossible to obtain accurate boundaries. In [6], [7], boundary detection based methods are proposed. The method proposed in [6] uses a boundary-based formula, but it requires the marking of 72 landmark points. The method proposed in [7] requires a direct edge in the 15-pixel neighborhood for each pixel, and does not consider depth information. The method of making visual features by hand-crafted method [8]–[11] converted the boundary problem into a pixel classification problem, which obtained satisfactory results. In [8], each superpixel is classified into optic disc or non-optic disc using histogram and center surround statistics, and position information is also included in the feature space to improve performance. In [9], a sliding window-based machine learning framework is used to acquire candidate optic disc and optic cup regions through a sliding window, and then to learn new histogram-based features. In [10], an unsupervised method that follows the superpixel frame and domain is used to segment optic disc and optic cup without using any additional training images. In [11], pixel features are extracted using stereo pairs for evaluation and classification. However, these methods of using hand-crafted features cannot obtain deeper feature information, and are susceptible to mis-segmentation due to noise, such as noise caused by pathological reason.

Due to convolutional neural networks(CNN) have received more and more attention in the medical field, and have achieved good results in the segmentation of optic disc and optic cup. In [12], the convolutional neural network based method is proposed for the segmentation of optic disc and optic cup, and an entropy based sampling technique is used to reduce computational complexity. The work in [13] proposes a deep convolutional neural network with feature learning, which embeds data in the receiving field by embedding a network with a more complex structure. In [14], by using the powerful classification ability of convolutional neural networks, the characteristics of high discrimination are learned from the original pixel intensity for feasibility analysis. In [15], an Ensemble Learning method based on architecture is proposed for extracting optic cup and optic disc segmentation. The entropy sampling technique is used to select the information points, and then the unsupervised graph cut algorithm is used to obtain the final segmentation result. In [16] a modified U-Net deep network was introduced for optic disc and optic cup segmentation. However, it still separates optic disc and optic cup segmentation in a sequential manner, without considering the interrelationship between the two. In [17], a new quadratic divergence regularized support vector machine (QDSVM) is used to segment optic disc and optic cup, but the image feature extraction is insufficient. In [18], a multi-label deep network is proposed to joint segmentation optic disc and optic cup, which alleviates the difficulty of information restoration. In [19], two U-Net structures are stacked to segmentation optic disc and optic cup, but the parameters of the model are increased and themodel training is difficult. In [20], a recurrent neural network is used to enhance the dependencies between the points on the optic disc and optic cup boundary, but the segmentation effect on the optic disc is not ideal.

In these studies, the visual features are extracted using a learning classifier, by which the pixels or patches of the retinal image are determined as the background, optic disc, and optic cup region. However, the features extracted lack sufficient discriminatory representation and are susceptible to low quality images and pathological regions. Furthermore, the interrelationship between optic disc and optic cup is ignored by these features. At present, Deep Convolutional Neural Networks(DCNN) has achieved breakthrough results in the medical image analysis. However, in the tasks of optic disc and optic cup segmentation, the existing DCNN method shows weak feature extraction ability, and insufficient learning of feature information in the dataset. It is unable to learn deeper information, and the generalization ability of the model is weak. When upsampling, some rough low-level information is difficult to restore. In order to effectively alleviate these problems, and improve the accuracy of optic disc and optic cup segmentation. In this paper, we use an end-to-end multi-label DCNN model with Generative Adversarial Nets(GAN) to segment optic disc and optic cup simultaneously. The model uses DCNN to learn various visual features and hierarchical information of retinal images, thereby exploring more image-related information and having more representation capabilities than hand-crafted features. The main contributions of our work include:

1) We propose an end-to-end multi-label DCNN model (GL-Net) that enables automatic segmentation of both optic disc and optic cup, and combined with the idea of GAN, improving the segmentation performance of the model. In generator, the decoding layer combines the coarse low-level segmentation information with the fine-layered appearance information of the coding layer. We have also reduced the downsampling factor to alleviate the problem of excessive loss of feature information and difficulty in recovery.
2) In order to prevent the mode collapse problem when GL-Net segment large-resolution images, we add the L_1 distance function, the cross-entropy function, and balance the influence of the three function on the loss function.
3) In order to alleviate the problem of insufficient data, we use data augmentation to augment the training data and use the generalization ability of transfer learning in DCNN, which the problem of insufficient training data is alleviated to improve the segmentation performance.

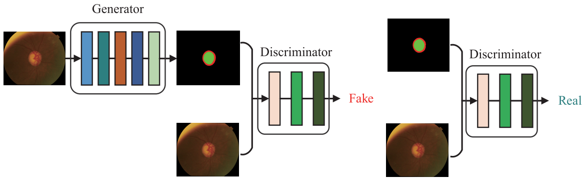

FIGURE 1. Optic disc and cup segmentation model.

II. METHODS

In this paper, we simultaneously segment optic disc and optic cup of the retinal image by using an end-to-end multi-label DCNN model, which is formed by deep convolutional generative adversarial networks (DCGAN) [21]. In the GL-Net, the training set is $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^{H\times W\times C_1}$, $y_i \in \mathbb{R}^{H\times W\times C_2}$. X, Y are two different domains. We define the retinal image set is $T = \{x_i\}_{i=1}^{N} \in X$, optic disc and optic cup label sets is $L = \{y_i\}_{i=1}^{N} \in Y$. We need to train G to generate mapping relationship $X \to Y$.

The model of this paper is shown in Figure.1. By using the idea of Generative Adversarial Nets(GAN) [22], which is the Nash equilibrium in game theory. The main structure of th model consists of two parts, G and D. G learns the mapping relationship between the retinal image x and the corresponding optic disc and optic cup. Finally, the optic disc and optic cup in the input retinal image x are segmented such that the representation $D(G(x))$ of the segmentation result $G(x)$ on D coincides with the representation $D(y)$ of ground truth y on D. D learns the difference between y and $G(x)$, and correctly discriminates the source of the input optic disc and optic cup label (ground truth or G), then guides G to reduce this difference, for making the segmented optic disc and optic cup more accurate. In the model training, G and D need to be continuously optimized to improve their segmentation and discriminating ability, and find the Nash equilibrium between the G and D. when G and D reach the Nash equilibrium, that the output of D equals to 1/2, and the source of the optic disc and optic cup label cannot be correctly discriminated. It is considered that the training is completed, and G can accurately segmentation optic disc and optic cup.

During the training of the model, the goal of G is defined as: when inputing x, $x \in T$, the output is $G(x)$), ideally, $G(x) = y$, $y \in L$. There are two goals for D, which are: (1) When the input is $\{x, G(x)\}$, the output value $D(x, G(x)) \approx 0$. (2) When the input is x, y, the output value is $D(x, y) \approx 1$. The ultimate goal of the model is:

$$\underset{G}{min}\,\underset{D}{max}\{\mathbb{E}_{x\in T,y\in L}[\log D(x,y)] + \mathbb{E}_{x\in T,y\in L}[\log(1 - D(x, G(x)))]\} \quad (1)$$

A. GENERATOR NETWORK ARCHITECTURE

The G is a 19-layer full convolutional network model which network structure is shown in Figure.2. This network takes the retinal image x as input. In this paper, we set $H = W = 512$, $C_1 = 3$. The network structure includes an encoder (left side) and a decoder (right side). The encoder extracts features of the retinal image by using the VGG16 network structure. The downsampling factor of the original VGG16 network is 32, but excessive downsampling will cause some feature information to be lost, and it is difficult to restore this information. For the optic disc region that is smaller than 32 pixels, when downsampling is set to 32, the information of these regions is completely lost. Furthermore, increasing the number of upsampling will increases the amount of calculation and parameters. Therefore, we removed the last two downsampling layer and changed the downsampling factor to 16, which alleviated the loss of information and reduced the parameters and calculations of the model. The feature map for each convolutional layer output is activated using the ReLU [23] activation function. In order to alleviate some of the low-level information is difficult to recover, the model can obtain complete context information, learn more accurate segmentation information, and avoid a large amount of low-level information directly through the entire network layer. When the decoder uses the deconvolution to upsample and restore the feature information, the feature map input by the pool layer in the encoder is directly transmitted to the decoder by using skip connection. When the decoder uses the deconvolution to perform upsampling to restore the feature information, the output segmentation feature map is spliced with the corresponding size feature map input by the pooling layer in the encoder in the dimension of the channel. After 4 times of downsampling at the decoding layer, the encoder uses 4 times of upsamplings to restore the feature map to the same height and width as the input image. The final classifier in the encoder utilizes 1×1 convolutional layer with softmax activation as the pixel-wise classification to produce the probability map. For multi-label segmentation, the output $G(x)$ is a K channel probability map, where K is the class number (in our work, $K = 3$ for optic disc, optic cup, and background). The predicted

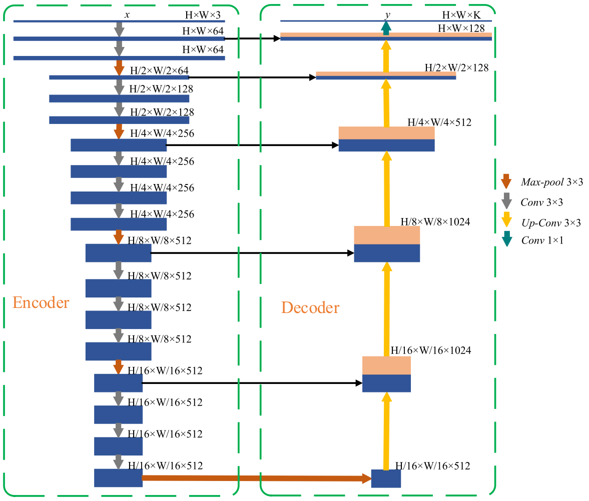

FIGURE 2. Generator network structure.

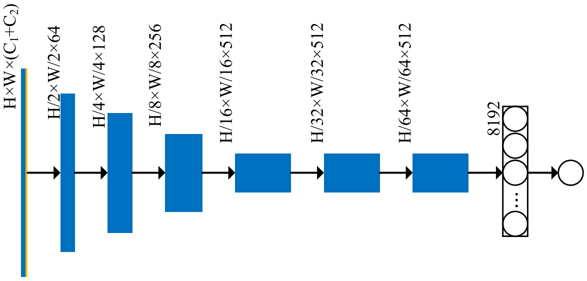

FIGURE 3. **Discriminator network structure.**

probability map corresponds to the category with maximum probability for each pixel, and optic disc and optic cup is segmented at the same time. In order to prevent the model from overfitting and improve the generalization ability of the model, we set *dropout* $= 0.8$ on the last two layers of the encoder for each iteration, and set the *dropout* $= 0.8$ for each layer of the decoder.

B. DISCRIMINATOR NETWORK ARCHITECTURE

The network structure of the discriminator is an 8-layer network model, which is shown in Figure.3. Retinal image x and ground truth y, or $G(x)$, are concatenated as input data. The size of input data for D is $H \times W \times (C_1 + C_2)$. D output y is the probability($D(x, y)$) of optic disc and optic cup label of the retinal image x, so as to optimize the parameters in the network structure. A strided convolution (strided $= 2$) instead of a pool layer is used in each layer. The size of the feature map is reduced to 1/4 of the original one after each convolution. In order to speed up the convergence and mitigate the impact of the initialization weight on the network model, we normalize the input of the current layer using batch normalization in the convolutional layer (mean $s = 0$, variance $\delta = 1$). In each layer, Leaky-ReLU is used as the activation function, with $\alpha = 0.2$. After the input image is downsampled 6 times the resolution of the feature image becomes $\frac{H}{64} \times \frac{W}{64}$. Finally the fully connected layer is used for connection. D output the discriminant result: 0 or 1. According to the error between D result and the label result, the parameters θ_g of the G and the parameter θ_d of D are tuned according to the gradient descent method, so as to achieve the best effect of model optimization.

C. LOSS

The loss function of G: G improves the ability to deceive D by learning the distribution of the ground truth, so that D cannot correctly distinguish the source of the data. Let x denotes original retinal image, y denotes ground truth, the loss function is defined as (2):

$$\mathcal{L}_G = \mathbb{E}_{x\sim T}[(1 - \log D(x, G(x)))] \tag{2}$$

Here, $G(x)$ denotes optic disc and optic cup segmentation by G, and $D(x, G(x))$ denotes the probability that optic disc and optic cup is segmented by the G.

The loss function of D: D is to recognition the source of optic disc and optic cup label as much as possible from the retinal image. Taking the ground truth as input, the ideal probability value of D output is $D(x, y) \approx 1$. Taking optic disc and optic cup segmentation by the G is input, the target of D as that the probability value of $D(x, G(x)) \approx 0$, and the loss function is defined as following(3):

$$\mathcal{L}_D = \mathbb{E}_{x\sim T,y\sim L}[\log(1 - D(x, y))] + \mathbb{E}_{x\sim T}[\log D(x, G(x))] \tag{3}$$

Here, in eq.(3), the first part is the input of the ground truth. The second part is the input of optic disc and optic cup segmentation by the G.

According to the theory of GAN, the performance of D and G continuously optimizes parameters by iterative training. Finally, when D cannot determine the source of optic disc and optic cup label map, we think that the G can segment the close to ground truth, and the objective function is described as follows(4):

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x\sim T,y\sim L}[\log D(x, y)] + \mathbb{E}_{x\sim T}[\log(1 - D(x, G(x)))] \tag{4}$$

Here, G attempts to minimize the objective function, and D attempts to maximize the objective function. $G^* = \underset{G}{min}\,\underset{D}{max}\,\mathcal{L}_{cGAN}(G, D)$

In order to achieve better segmentation effect, cross entropy (5) is utilized to compare optic disc and optic cup segmented with ground truth, and tune the weight of the G to minimize its value $\mathcal{L}_{CrossEntropy}(G)$ by iterative training.

$$\mathcal{L}_{CrossEntropy}(G) = -\frac{1}{K}\sum_{i=1}^{K}[y^{(i)} log\sigma(x^{(i)}) + (1 - y^{(i)})log(1 - \sigma(x^{(i)}))] \tag{5}$$

Here, K is the number of categories, $y^{(i)}$ is the *i-th* category, $\sigma(x^{(i)})$ is the probability that x belongs to the *i-th* category, and $\sigma(\cdot)$ is the *softmax* activation function.

In our pilot experiment, we have found that when eq.(4) is combined with traditional loss functions (such as the L_1 distance function (6)), the segmented optic disc and optic cup by G is closer to the ground truth. The role of D remains the same. The task of G not only deceives D, but also minimizes the L1 distance between the generated optic disc and optic cup label maps and ground truth.

$$\mathcal{L}_{L_1}(G) = E_{x\sim T,y\sim L}[\| y - G(x) \|] \tag{6}$$

Combine the objective function of the GAN, the L_1 distance function, and the cross-entropy function, the final total objective function of the model is defined as (7):

$$G^* = \arg \underset{G}{min}\,\underset{D}{max}\,\mathcal{L}_{cGAN}(G, D) + \lambda_1\mathcal{L}_{L1}(G) + \lambda_2\mathcal{L}_{CrossEntropy}(G) \tag{7}$$

Here, λ_1 and λ_2 are used to balance the three loss functions. Increasing the effect of the cross entropy and the L_1 distance

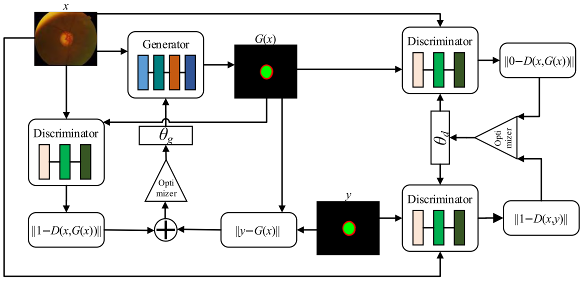

FIGURE 4. Model training process.

function on the total loss function makes the weight of the G tuned better.

III. TRAIN

Adam [24] is used as the optimizer to train through stochastic gradient descent(SGD), with the parameters of the optimizer being $\beta_2 = 0.999$, $\beta_1 = 0.5$, $\varepsilon = 10^{-8}$. Learning rate lr is initialized to 10^{-4}, and after 400 iterations, the learning rate decreases to 10^{-6}. To reduce the in-stability of the stochastic gradient during training, we set *mini-batch* $= 2$.

In the training process of GAN, we need to train G to minimize $\log(D(x)) + \log(1 - D(G(x)))$. At the same time, it is necessary to continuously improve D to distinguish the ground truth and G for optic disc and optic cup segmentation, so that $\log(D(x)) + \log(1 - D(G(x)))$ is maximized. D and the G are optimized mainly by means of alternate optimization. Firstly, G is fixed, and the accuracy of D is maximized by optimizing G. Then, D is fixed, and G is optimized to minimize the accuracy of D, and the source of optic disc and optic cup label maps cannot be correctly distinguished. When the model performs adversarial training, D can outweigh G with a slight advantage. In order to maintain balance, in the same iteration, the parameter of G is updated by *steps* times, and then the parameters of D are updated once. In this paper, *steps* $= 2$.

The training process of the model is shown in Figure.4. The training process of G including:

1) G segment optic disc and optic cup region in the input retinal image x, and the error between the segmentation result and the ground truth is measured by $\|y - G(x)\|$. The weight(θ_g) of G is optimized according to the error back propagation, which makes the error close to 0.
2) Input the retinal image x, and $G(x)$, the segmentation of G into D, and optimize the weight(θ_g) of G to minimize the error $\|1 - D(x, G(x))\|$.

The training process of D is:

1) When the input image pair is $\{x, y\}$, the weight(θ_d) of D is optimized according to the error between the result $D(x, y)$ of D and 1, so as to minimize the error $\|1 - D(x, y)\|$.
2) When the input image pair is $\{x, G(x)\}$, D weight(θ_d) is optimized according to the error between $D(x, G(x))$ output value of D and the 0, so as to achieve the minimum error $\|0 - D(x, G(x))\|$.

The two networks G and D optimize their corresponding weights through continuous iteration, and finally the value of the loss function reaches the global minimum. The detailed process is shown in Algorithm 1.

A. TRANSFER LEARNING

Due to the expensive cost and complicated acquisition procedures of medical data collection, in most situations, the training data for medical images lacks accurate annotation. Previous studies have demonstrated that transfer learning(TL) [25] in DCNN can alleviate the problem of insufficient training data [26], [27]. The convolution filter weights learned at the lower layers of the network are general, while the convolution filter weights in the higher layers are applicable to different tasks [28]. By transferring the rich feature knowledge learned from the relevant dataset to the target learning task, we could reduce the problem of over-fitting caused by training on limited medical dataset, and further improve the performance of the network model.

Therefore, in the feature extraction of the retinal image, we utilized the off-the-shelf trained weights in the DeepLab [29] model to initialize our model. DeepLab model used the PASCAL VOC 2012 dataset [30] for training. Compared to the small dataset (101 images) of DRISHTI-GS1 [31], the PASCAL VOC dataset contains more than 17,000 images with pixel level annotations. In order to take advantage of the effective generalization ability of transfer learning in DCNN, we use the trained weights in the DeepLab model to initialize the convolution filter in the downsample network structure, while the remaining convolutional layers are randomly initialized with a Gaussian distribution. Then, we implement end-to-end training in a mini-batch gradient to tune the weight in the entire network structure. In the experiment, we observed that the priori knowledge model learned from other datasets using transfer learning converges faster in the training process than the models trained with convolution filter weights initialized randomly.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Our framework was implemented under the open-source deep learning library TensorFlow [32] on a server with Intel(R) Xeon(R) E5-2620 v3 2.40GHz CPU, Tesla K80 GPU, and Ubuntu64 as OS.

A. DATASET AND PREPROCESSING

There are totally 101 retinal images in the Drishti-GS1 dataset [31], with 31 normal images and 70 lesion images. There are 50 images in the training set and 51 images in the testing set. For each image, optic disc and optic cup are marked manually by 4 ophthalmologists with different clinical experience, and the averaged optic disc and optic cup region of 4 expert markers is taken as the ground truth, which is shown in Figure.5.

Algorithm 1 Model Training Process

```
Input:
        Number of samples per input(minibatch)  $m$.
        Hyperparameters  $steps$.
        Total number of iterations of training  $trainiterations$.
        Retinal image set  $\{x_i\}_{i=1}^{N}, x_i \in X$.
        Ground True  $\{y_i\}_{i=1}^{N}, y_i \in Y$.
Output:
        $\{G(x_i)\}_{i=1}^{N}$  segmentation result of G.
        $\{D(x_i, G(x_i))\}_{i=1}^{N}$  The output of the discriminator when the input label is $\{G(x_i)\}_{i=1}^{N}$.
        $\{D(x_i, y_i)\}_{i=1}^{m}$.  The output of the discriminator when the input label is $\{y_i\}_{i=1}^{N}$.
for $k$ in trainiterations do
    for $q$ in steps do
        Randomly select $m$ samples from the sample set to get $\{x_i\}_{i=1}^{m}$.
        The convolution kernel parameter $\theta_g$ in G is optimized using a SGD:
        $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} [(y_i - G(x_i)) + log(1 - D(x_i, G(x_i)))]$.
    end for
    Randomly select $m$ samples from the sample set to get $\{x_i\}_{i=1}^{m}$.
    Input $\{x_i\}_{i=1}^{m}$ into G to generate $m$ results $\{G(x_i)\}_{i=1}^{m}$.
    Get the corresponding $m$ tag sample $\{y_i\}_{i=1}^{m}$.
    Optimize the convolution kernel parameter $\theta_d$ in the discriminator using SGD:
    if input== $\{x_i, G(x_i)\}_{i=1}^{m}$:
        $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} log(D(x_i, G(x_i)))$.
    if input== $\{x_i, y_i\}_{i=1}^{m}$:
        $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} log(1 - D(x_i, y_i)))$.
end for
```

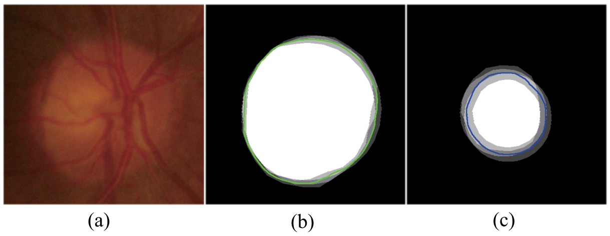

FIGURE 5. Sample markings and optic disc, optic cup segmentation maps. (a) retina image (b) 4 expert markings for optic disc and averaged optic disc (green) region. (c) 4 expert markings for optic cup and averaged optic cup (blue) region.

As there is only a small size of training data, in order to prevent model overfitting, and improve the accuracy and robustness of the model, we perform data augmentation to the training set, each training image is randomly sampled by one of the following options:

1) Using the entire original input image.
2) Shifting the input image up and down randomly within a range of 300 pixels.
3) Rotating the input images randomly within the range of [0, 360°].
4) Scaling the input image randomly by a factor of [0.8, 1.5].

Then a patch is sampled randomly, with the size of each sampled patch is [0.6, 1] of the original image size. After the aforementioned step, each sampled patch is resized to fixed size (512 × 512) and is flipped horizontally, and inverted vertically with probability of 50 %.

B. EVALUATION METRIC

We evaluate the performance of different methods by the boundary distance localization error(BLE) [31] and $F1$, which is widely used by the research community. According to the combination of the real category label and the classifier prediction, there are 4 cases in the model segmentation result, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

$F1$ is defined as:

$$F1 = 2 \times \frac{Precison \times Recall}{Precision + Recall} \tag{8}$$

Here, $Precision$(9) and $Recall$(10) is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

However, since $F1$ is not suitable for evaluating the segmentation performance at the local (boundary) level,

TABLE 1. The model segmentation results with and without GAN and TL.

Method	DA	optic disc(mean/std)		optic cup(mean/std)		Parameters(10^6)	Times(s)	
		F1	BLE(px)	F1	BLE(px)		train	test
Baseline		0.950/0.019	8.95/3.61	0.831/0.10	20.48/11.02	**24.55**	**1.0**	1.0
Baseline	✓	0.955/0.018	8.25/3.54	0.845/0.11	19.28/10.77	24.55	1.0	1.0
Baseline+TL		0.956/0.015	7.01/3.53	0.860/0.11	15.04/8.71	24.55	1.0	1.0
Baseline+TL	✓	0.960/0.015	7.50/3.51	0.867/0.11	15.63/8.51	24.55	1.0	1.0
Baseline+GAN		0.961/0.012	6.85/3.48	0.889/0.09	13.11/6.87	30.85	1.3	1.0
Baseline+GAN	✓	0.968/0.012	6.38/3.43	0.896/0.09	12.54/6.62	30.85	1.3	1.0
GL-Net(Baseline+GAN+TL)		0.970/0.012	6.33/3.30	0.897/0.08	12.20/6.33	30.85	1.3	1.0
GL-Net(Baseline+GAN+TL)	✓	**0.971/0.012**	**6.23/3.34**	**0.905/0.08**	**11.97/6.12**	30.85	1.3	**1.0**

the average BLE is employed to measure the boundary distance (in pixels) between the model segmentation result (C_o) and ground truth (C_g). *BLE* is defined as(11):

$$BLE = \frac{1}{N}\sum_{\theta=0}^{N-1}\sqrt{(d_g^\theta)^2 - (d_o^\theta)^2} \tag{11}$$

where, d_g^θ and d_o^θ are the distance from disk center to points on C_g and C_o, respectively in the angular direction indexed by θ. N is taken to be 24 in our evaluation. The desirable value for *BLE* is 0.

C. MODEL PERFORMANCE IMPROVEMENT

In order to verify the effectiveness of using GAN and TL, we compared the performance with and without GAN and TL in the model. The combination is as follows:

1) Baseline Model: The base model consists only of G(Figure.2), and the loss function consists of the cross entropy formula (eq.(5)) and the L_1 distance function (eq.(6)).
2) Baseline Model + TL: On the basis of 1, the convolution kernel in G is initialized using the trained weights in DeepLab.
3) Baseline Model + GAN: The network structure is shown in Figure.1. It consists of two network structures, G and D, with Eq(7) as the loss function. The weight of the convolution kernel is randomly initialized using a Gaussian distribution.
4) Baseline Model + GAN + TL: On the basis of 3, the convolution kernel in G is initialized using the trained weights in DeepLab.
5) For the above four combinations, the effects of the training set with and without data augmentation(DA) processing on the performance of the model were tested.

The experimental results are summarized in Table.1. It is shown that utilize GAN and TL both can effectively improve the performance of the model. When using Baseline + TL, the segmentation performance of the model has been significantly improved. The performance of optic disc and optic cup on $F1$ is 0.5 % and 2.2 % higher than Baseline, and the BLE is reduced by 1.25px and 3.65px respectively. It proves the effectiveness of TL.

When Baseline + GAN is used, the parameters of the model are increased by 1/4 due to the addition of D. D only supervises each other with G during training, and optimizes each other. When the training is completed, only G is used to segment the retinal image. In optic disc and optic cup segmentation results, $F1$ is 1.3 % and 5.1 % higher than Baseline, respectively. and 1.97px, 6.74px are reduced on BLE, respectively, and the variance is smaller, especially the effect on optic cup is obviously improved. Through the experimental results, it can be seen that G and D in the GAN model learn from each other through adversarial training, and constantly optimize their corresponding weights so that G can more accurately segment optic disc and optic cup.

When Baseline + GAN + TL is used, the segmentation performance of the model is further improved. When TL is used, on F1, optic disc and optic cup segmentation results were increased by 0.3 % and 0.9 %, respectively. On BLE, optic disc and optic cup segmentation results were reduced by 0.05px and 0.57px, respectively, and the standard deviation is further reduced. By taking advantage of transfer learning, the overfitting caused by the limited medical dataset could be reduced by learning the rich feature level information from the related dataset, and the generalization ability of the model can be effectively improved

It can be seen from the experimental results that in the model training, with the data augmentation processing on the training dataset, the performance of the model can be effectively improved.

Figure.6 shows a visual example of the segmentation results. The first two rows are glaucoma cases, and the remaining two rows are normal eyes. The segmentation performance of the three method is compared. As can be seen from Figure.6, compared with baseline, with GAN, it can help to significantly improve the accuracy of optic disc and optic cup segmentation results. The segmentation results are basically consistent with ground truth, and the edges of the segmented regions are smoother. When model with TL, the performance of the model has been improved. In the last retinal image, the optic cup region segmented using baseline is larger than the ground truth region, which makes the CDR value too large and is easily misjudged into glaucoma. After joining GAN, the segmented optic cup region is close to the ground truth region, which can effectively reduce this misjudgment.

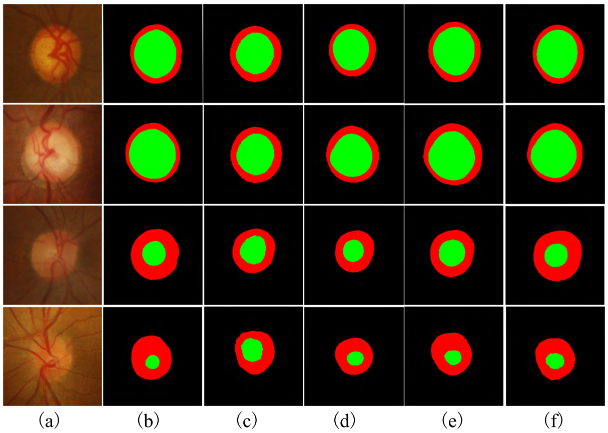

FIGURE 6. The visual examples of optic disc and cup segmentation, where the green and red region denote the cup and disc segmentations, respectively (a) cropped region around optic disc. (b) Ground Truth. (c) Segmented result by Baseline. (d) Segmented result by Baseline + TL. (e) Segmented result by Baseline + GAN. (f) Segmented result by Baseline + GAN + TL.

D. QUANTITATIVE ANALYSIS OF SEGMENTATION RESULTS OF OTHER METHODS AND DISCUSSION

In order to validate the performance of GL-Net, in this group of experiments, we compared the $F1$ and BLE of GL-Net with some state-of-art method, such as Vessel Bend [4], BCRF [6], Graph cut prior [7], Superpixel [8], Boosting CNN [12], M-Net [18], Stack-U-Net [19], RACE-net [20], and Multiview [33].

The segmentation results of these methods on the testset are shown in Table.2. It can be seen that the proposed method achieves the best performance on these two evaluation metric.

Vessel Bend [4] integrated the local information around the points of interest in the feature space to improve the robustness of the method, and used the parametric technique of curved vessels to fit optic disc region. Graph cut prior [7] uses the energy function with prior knowledge and optimization algorithms to capture the shape and position of optic disc and optic cup. The interrelationship between optic disc and optic cup was not considered in this bottom-up method, and it does not performance well on the optic cup. Multiview [33] shifts focus from optic disc and optic cup regions to optic disc and optic cup boundaries, and find the final boundary by considering the points on different sectors by way of confidence. A better performance was obtained by Multiview method than the existing approaches [4], [7], but it requires two retinal images for each eye. Superpixel [8] detects the boundary between optic disc and optic cup by classifying the pixels, and location information is added to the feature space information to improve model performance. However, it is based on various hand-crafted visual features, which lack sufficient discriminatory representation and are susceptible to pathological region. Boosting CNN [12] reduces computational complexity by using entropy-based sampling techniques to achieve better results than uniform sampling. However, the feature extraction ability of the model is weak, and it is unable to learn deeper semantic information, so the performance on optic disc and optic cup is relatively poor. BCRF [6] has joint extraction optic disc and optic cup boundary by Conditional Random Field formulation, and obtained the state-of-art performance(0.97) on optic disc, but the performance of optic cup is not good than deep learning method. M-Net [18] uses multi-input and multi-output methods to obtain multi-scale feature information, but the main structure is relatively simple, and unable to learn more feature information, resulting in relatively low segmentation performance of the model. The $F1$ of the optic disc and optic cup is only 0.959 and 0.866 respectively. Stack-U-Net [19] by stack two U-Nets, although it has achieved good results(0.89) on optic cup, but it hasmany parameters and complicated structure. RACE-net [20] uses the recurrent neural network method to segment optic disc and optic cup. It sharing a large amount of weight, which making feature information extraction insufficient, resulting in the segmentation result on the optic cup is 2 % and 1.74px worse than Stack-U-Net in term of $F1$ and BLE, respectively.

In this paper, we proposed a DCNN model, GL-Net, which joint segmentation optic disc and optic cup. It achieves state-of-art segmentation performance on the Drishti-GS1 dataset. Particularly in the optic cup segmentation results, GL-Net outperforms state-of-art method RACE-net by 3.5 % and 4.16px in terms of $F1$ and BLE, respectively.

Figure.7 shows the edge curves of optic disc and optic cup region segmentation by the state-of-art methods, such as Vessel Bend [4], BCRF [6], Superpixel [8], Boosting CNN [12],

TABLE 2. Drishti-GS1 dataset optic disc and optic cup segmentation result.

Method	optic disc(mean/std)		optic cup(mean/std)		Times(s)
	F1	BLE(px)	F1	BLE(px)	test
Vessel Bend [4]	0.960/0.02	8.93/2.96	0.770/0.20	30.51/24.80	4.3
Superpixel [8]	0.950/0.02	9.38/5.75	0.800/0.14	22.04/12.57	3.2
Multiview [33]	0.960/0.02	8.93/2.96	0.790/0.18	25.28/18.00	5.0
Graph cut prior [7]	0.940/0.06	14.74/15.66	0.770/0.16	26.70/16.67	4.5
Boosting CNN [12]	0.947/0.03	9.10/3.10	0.83/0.140	16.50/11.01	1.9
BCRF [6]	0.970/0.02	6.61/3.55	0.830/0.15	18.61/13.02	2.0
Modification-U-Net [16]	0.950/-	-/-	0.850/-	-/-	-
M-Net [18]	0.959/0.04	7.97/8.29	0.866/0.11	17.05/12.76	1.8
Stack-U-Net [19]	0.970/0.02	6.47/3.51	0.890/0.09	14.39/7.18	2.2
RACE-net [20]	0.970/0.02	6.06/3.84	0.870/0.09	16.13/7.63	1.5
Baseline+GAN+TL	**0.971/0.012**	**6.23/3.34**	**0.905/0.08**	**11.97/6.12**	**1.3**

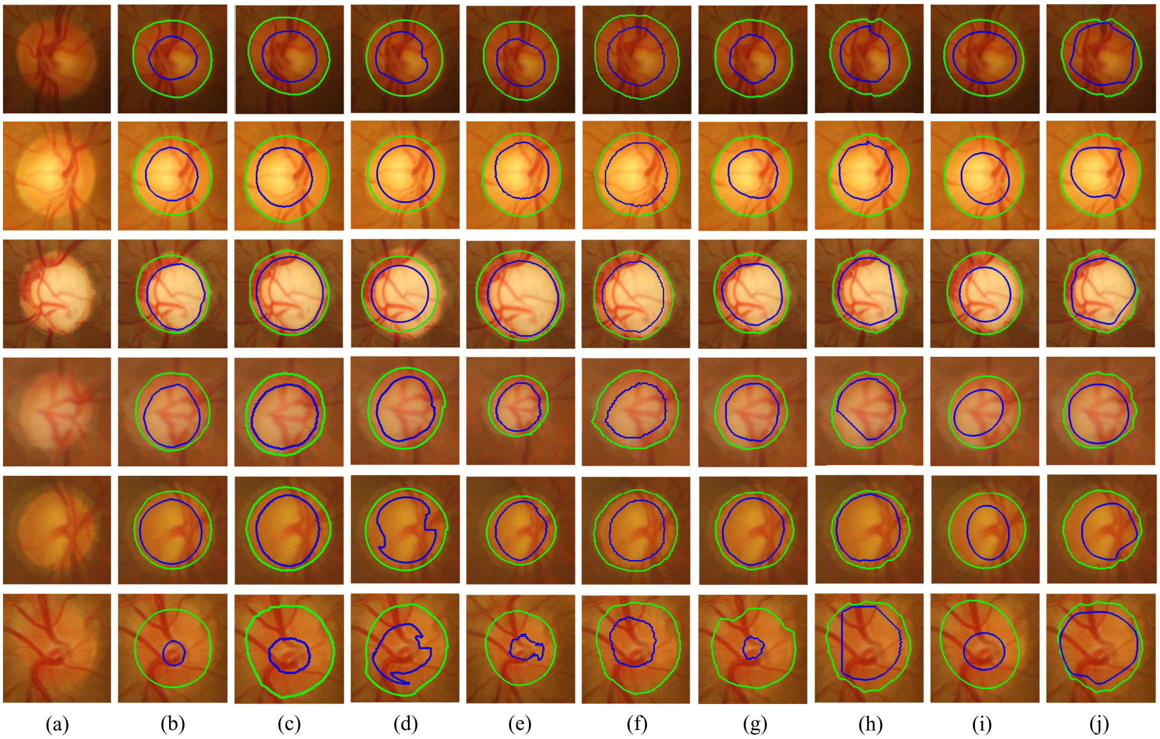

FIGURE 7. Qualitative results on some challenging cases. optic disc and optic cup boundaries are depicted in green and blue respectively. (a) cropped region around optic disc. (b) Ground Truth. Results of (c) GL-Net(ours). (d) Segmented result by M-Net [18]. (e) Segmented result by RACE-net [20]. (f) Segmented result by Stack-U-Net [19]. (g) Segmented result by BCRF [6]. (h) Segmented result by Multiview [33]. (i) Segmented result by Superpixel [8]. (j) Segmented result by Vessel Bend [4].

M-Net [18], Stack-U-Net [19], RACE-net [20] and Multiview [33] on samples drishtiGS_006, drishtiGS_007, drishtiGS_019, drishtiGS_055, drishtiGS_074, and drishtiGS_100. For the Superpixel method [8](i), in the case of glaucoma, the edge of the segmented region is smoother, but the segmented optic cup region is smaller than the standard optic cup region, which may result in a smaller CDR value. The boundary of optic disc region segmented by the Multiview [33](h) method is close to ground truth, but the segmented optic cup region is generally too large, which causes the value of the CDR to be too large, and it is easy to misjudged the normal eye as glaucoma. Vessel Bend [4](j) also has the same problem as Method Multiview. The segmentation effect of optic cup is not ideal, and normal eyes and glaucoma cannot be correctly distinguished. M-Net [18](d) easily generates a smaller optic disc. For the last row, where the image is blurred and has low-contrast for identifying the optic cup boundary, M-Net fail to produce accurate optic cup segmentation. The Stack-U-Net [19](f) method is susceptible to the shrinking arc around the optic disc, resulting in a large segmentation area. RACE-net [20](e) does not accurately process retinal images with poor contrast and the lack of a strong gradients at the optic disc boundary. BCRF method [6](g) is better than other methods for segmentation of optic disc and optic cup, especially for optic disc segmentation. Although the segmentation effect on the optic cup region and the boundary level is better than other methods, however, there is still a large room for improvement, and the error between the segmentation result and the ground truth is relatively large. In contrast, the GL-Net model proposed in this paper segment the boundaries of optic disc and optic cup better than other methods. The average error and variance between the edge curve of the GL-Net segmentation results and the ground truth are small, which makes the CDR value more accurate. It effectively reduces the misjudgment of glaucoma, fully proved the generalization ability and effectiveness of GL-Net.

E. DISCUSSION

GL-Net is a multi-label DCNN model with the GAN ideas. During the training process, G and D adversarial training each other and iteratively optimize their corresponding weight, so that the performance of the model is continuously improved. When the decoder is upsampled, we utilize the skip layer connection to promote the fusion of low-level information with the high-level information. It makes the segmentation map obtain complete context information, effectively alleviating the problem that some low-level information is difficult to recover. We use the rich semantic information learned from other related datasets by transfer learning, which effectively accelerates the convergence speed and performance of the model. Finally, the effectiveness and generalization of the proposed GL-Net is evaluated on the DRISHTI-GS1 dataset. When segmenting retinal images in the testset, GL-Net achieved optimal results on both the mean and variance of the two evaluation metric. GL-Net can segment better results for normal retinal images and lesion retinal images, and has strong anti-lesion ability and noise interference ability, which proves that it has better robustness and generalization. The average F1 for OD and OC segmentation is 0.971 and 0.905, with the corresponding average boundary distance localization error(BLE) being 6.23 pixels and 11.97 pixels, respectively. In the testing phase, averagely, GL-Net requires the least time for one retinal image.

V. CONCLUSION

Accurate segmentation of the optic disc and optic cup has great practical significance for helping doctors to screen for glaucoma. In this paper, we propose GL-Net, a multi-label DCNN model, for joint segmentation of optic disc and optic cup. In the model, we use VGG16 for feature extraction and reduce the downsampling factor from 32 to 16, which allowing the model to extract more useful feature information. When upsampling, we use skip connections to encourage the fusion of feature information. The L_1 distance function and the cross entropy function are added to the loss function to supervised the model better converge. Due to the small number of training samples in the medical dataset, we use transfer learning and data augmentation to improve the generalization and segmentation performance of the model. Finally, the GL-Net is verified on the DRISHTI-GS1 dataset. The experimental results show that the performance of this method is better than the existing M-Net, Stack-U-Net, RACE-net and other state-of-art methods, which suggests that GL-Net is more suitable for optic disc and optic cup segmentation.

However, further processing of high-level feature information can improve the segmentation performance of the model. In the future work, we will explore a more powerful backbone network and feature processing structure, and different upsampling methods to improve the segmentation performance of the model.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their constructive suggestions.

REFERENCES

[1] A. Almazroa, R. Burman, K. Raahemifar, and V. Lakshminarayanan, “Optic disc and optic cup segmentation methodologies for glaucoma image detection: A survey,” *J. Ophthalmol.*, vol. 2015, no. 7, Sep. 2015, Art. no. 180972.

[2] A. Aquino, M. Emilio, E. Gegundez-Arias, and D. Marin, “Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques,” *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1860–1869, Nov. 2010.

[3] S. Lu, “Accurate and efficient optic disc detection and segmentation by a circular transformation,” *IEEE Trans. Med. Imag.*, vol. 30, no. 12, pp. 2126–2133, Dec. 2011.

[4] G. D. Joshi, J. Sivaswamy, and S. R. Krishnadas, “Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment,” *IEEE Trans. Med. Imag.*, vol. 30, no. 6, pp. 1192–1205, Jun. 2011.

[5] D. W. K. Wong, J. Liu, J. H. Lim, H. Li, and T. Y. Wong, “Automated detection of kinks from blood vessels for optic cup segmentation in retinal images,” *Proc. SPIE*, vol. 7260, no. 6, 2009, Art. no. 72601J.

[6] A. Chakravarty and J. Sivaswamy, “Joint optic disc and cup boundary extraction from monocular fundus images,” *Comput. Methods Programs Biomed.*, vol. 147, pp. 51–61, Aug. 2017.

[7] Y. Zheng, D. Stambolian, J. O’Brien, and J. C. Gee, “Optic disc and cup segmentation from color fundus photograph using graph cut with priors,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* , vol. 16. Berlin, Germany: Springer, 2013, no. 2, pp. 75–82.

[8] J. Cheng *et al.*, “Superpixel classification based optic disc and optic cup segmentation for glaucoma screening,” *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013.

[9] Y. Xu *et al.*, “Sliding window and regression based cup detection in digital fundus images for glaucoma diagnosis,” in *Proc. MICCAI*, 2011, pp. 1–8.

[10] Y. Xu *et al.*, “Optic cup segmentation for glaucoma detection using low-rank superpixel representation,” in *Proc. MICCAI*, 2014, pp. 788–795.

[11] M. D. Abràmoff *et al.*, “Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features,” *Investigative Ophthalmol. Vis. Sci.*, vol. 48, no. 4, pp. 1665–1673, 2007.

[12] J. G. Zilly, J. M. Buhmann, and D. Mahapatra, “Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images,” in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 136–143.

[13] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu, “Automatic feature learning for glaucoma detection based on deep learning,” in *Proc. MICCAI*, 2015, pp. 669–677.

[14] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, “Convolutional neural network transfer for automated glaucoma identification,” in *Proc. Int. Symp. Med. Inf. Process. Anal.*, vol. 10160, 2016, pp. 1–10. doi: 10.1117/12.2255740.

[15] J. Zilly, J. M. Buhmann, and D. Mahapatra, “Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation,” *Comput. Med. Imag. Graph.*, vol. 55, pp. 28–41, Jan. 2017.

[16] A. Sevastopolsky, “Optic disc and cup segmentation methods for glaucoma detection with modification of U-net convolutional neural network,” *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 618–624, Jul. 2017.

[17] J. Cheng, D. Tao, D. W. K. Wong, and J. Liu, “Quadratic divergence regularized SVM for optic disc segmentation,” *Biomed. Opt. Express*, vol. 8, no. 5, pp. 2687–2696, 2017.

[18] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, “Joint optic disc and cup segmentation based on multi-label deep network and polar transformation,” *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[19] A. Sevastopolsky, S. Drapak, K. Kiselev, B. M. Snyder, J. D. Keenan, and A. Georgievskaya. (2018). “Stack-U-net: Refinement network for image segmentation on the example of optic disc and cup.” [Online]. Available: https://arxiv.org/abs/1804.11294

[20] A. Chakravarty and J. Sivaswamy, “RACE-net: A recurrent neural network for biomedical image segmentation,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1151–1162, May 2019.

[21] A. Radford, L. Metz, and S. Chintala. (2015). “Unsupervised representation learning with deep convolutional generative adversarial networks.” [Online]. Available: https://arxiv.org/abs/1511.06434

[22] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[23] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[24] D. P. Kingma and J. Ba. (2014). “Adam: A method for stochastic optimization.” [Online]. Available: https://arxiv.org/abs/1412.6980

[25] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[26] H. Chen *et al.*, “Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks,” in *Proc. MICCAI*, vol. 9349, 2015, pp. 507–514.

[27] H.-C. Shin *et al.*, “Deep convolutional neural networks for computer-aided detection: CNN Architectures, dataset characteristics and transfer learning,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Proc. NIPS*, vol. 27, 2014, pp. 3320–3328.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[31] J. Sivaswamy, S. R. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, "Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation," in *Proc. IEEE 11th Int. Symp. Biomed. Imag. (ISBI)*, Apr./May 2014, pp. 53–56.

[32] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467

[33] G. D. Joshi, J. Sivaswamy, and S. R. Krishnadas, "Depth discontinuity-based cup segmentation from multiview color retinal images," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1523–1531, Jun. 2012.

YUN JIANG was born in Zhejiang, China, in 1970. She received the Ph.D. degree in Northwestern Polytechnical University, Xian, China, in 2007. She is currently a Professor with College of Computer Science and Engineering, Northwest Normal University. Her research interests include data mining, rough set theory and application, and medical image processing.

NING TAN was born in Hunan, China, in 1995. He is currently pursuing the M.S. degree with the College of Computer Science and Engineering, Northwest Normal University. His research interests include deep learning and medical image processing.

TINGTING PENG was born in Gansu, China, in 1995. She is currently pursuing the M.S. degree with the College of Computer Science and Engineering, Northwest Normal University. Her research interests include deep learning and medical image processing.