Weakly-Supervised Simultaneous Evidence Identification and Segmentation for Automated Glaucoma Diagnosis

Rongchang Zhao,[1,2] **Wangmin Liao,**[1,2] **Beiji Zou,**[1,2] **Zailiang Chen,**[1,2] **Shuo Li**[3*]
[1] School of Information Science and Engineering, Central South University, Changsha, China
[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha, China
[3] University of Western Ontario, London, ON, Canada
zhaorc@csu.edu.cn, slishuo@gmail.com

Abstract

Evidence identification, optic disc segmentation and automated glaucoma diagnosis are the most clinically significant tasks for clinicians to assess fundus images. However, delivering the three tasks simultaneously is extremely challenging due to the high variability of fundus structure and lack of datasets with complete annotations. In this paper, we propose an innovative Weakly-Supervised Multi-Task Learning method (WSMTL) for accurate evidence identification, optic disc segmentation and automated glaucoma diagnosis. The WSMTL method only uses weak-label data with binary diagnostic labels (normal/glaucoma) for training, while obtains pixel-level segmentation mask and diagnosis for testing. The WSMTL is constituted by a skip and densely connected CNN to capture multi-scale discriminative representation of fundus structure; a well-designed pyramid integration structure to generate high-resolution evidence map for evidence identification, in which the pixels with higher value represent higher confidence to highlight the abnormalities; a constrained clustering branch for optic disc segmentation; and a fully-connected discriminator for automated glaucoma diagnosis. Experimental results show that our proposed WSMTL effectively and simultaneously delivers evidence identification, optic disc segmentation (89.6% TP Dice), and accurate glaucoma diagnosis (92.4% AUC). This endows our WSMTL a great potential for the effective clinical assessment of glaucoma.

Introduction

Evidence identification, optic disc segmentation and glaucoma diagnosis are the indispensable parts of clinical practice since they provide quantitative evaluation and precise diagnosing for glaucoma assessment. In practice, a clinician often inspects fundus images for identifying the suspicious regions, then zooms-in to manually contour the optic disc, evaluates its appearance and strives the diagnosis of glaucoma. In the process, evidence regions indicate where the abnormalities appear and the segmented optic disc provides quantitative evaluation for discrimination between the normal and glaucomatous cases. Therefore, the tasks of identification, segmentation and diagnosis are the three key parts for clinical assessment of fundus image.

Although most previous works are devoted to segmentation of optic disc as evident for glaucoma diagnosis(Cheng et al. 2013; Chen et al. 2015; Almazroa et al. 2015), designing the well-performing model to simultaneously identify evidence, segment optic disc and diagnose glaucoma is still challenging due to some limitations. (1) Lack of large-scale segmentation ground truth even though plenty of training data with binary diagnostic labels (normal/glaucoma). Since diagnostic labels are readily available and efficiently obtained from diagnostic reports, it would be great to design the algorithm taking the report directly for training. (2) Lack of effective framework to model the relationships between evidence identification, optic disc segmentation and glaucoma diagnosis since they are heterogeneous tasks for mutually exclusive objectives. (3) Lack of effective method to deal with the high variability and extreme inhomogeneity of optic disc structure from fundus image across subjects.

The weakly-supervised method has great potential since it can learn from large-scale weak-label data to discover the evidence and optic disc, further enhance the diagnosis confidence of glaucoma. Recently, it has been demonstrated that convolutional neural networks (CNN) trained with diagnostic labels have the remarkable capability in abnormalities localization (Wang et al. 2017; Zhang, Bhalerao, and Hutchinson 2017). In those models, the standard global max/average-pooling (Zhou et al. 2016) retains the spatial structure of pixels that can be exploited to discover discriminative local regions. However, those models usually identify some sparse regions to deal with the single task, which deviate from the requirement of optic disc segmentation and glaucoma diagnosis that needs pixel-wise inference. Therefore, existing models are unsuitable to be used directly for simultaneous identification, segmentation and diagnosis.

In this paper, we propose the Weakly-Supervised Multi-Task Learning method (WSMTL, shown in Fig.1) to deliver evidence identification, optic disc segmentation and glaucoma diagnosis simultaneously. Our basic assumption is that the information extracted from high-level diagnosis task can act as helpful supervision for low-level evidence identification and segmentation tasks, when the low-level label information is insufficient. Specifically, we first built model to construct pyramid evidence maps based on the multi-scale features representation extracted from the diagnosis network when training with the binary diagnosis labels. Take the ev-

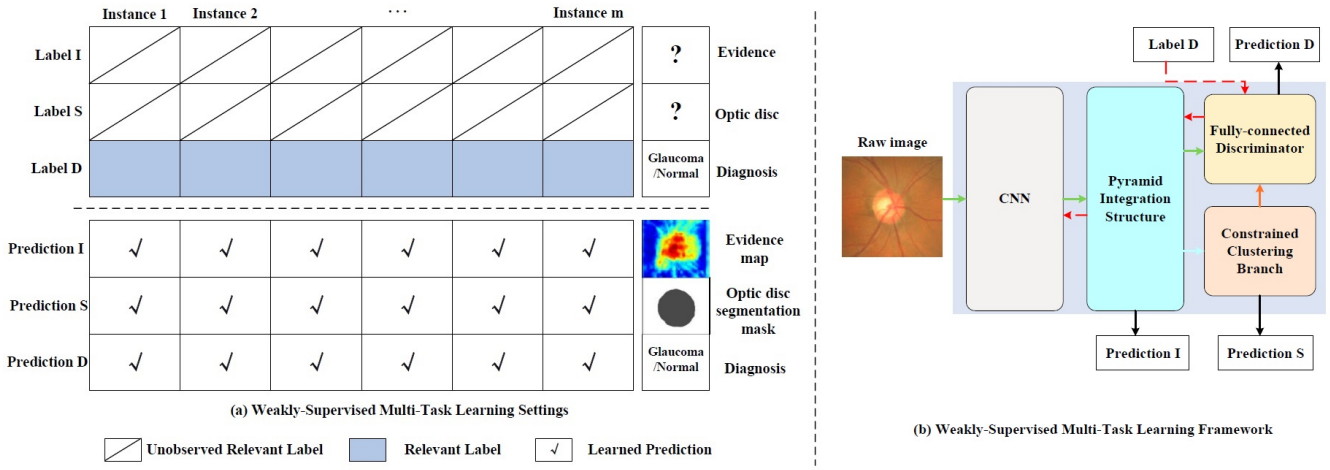

Figure 1: Weakly-Supervised Multi-Task Learning. (a) Weakly-Supervised Multi-Task Learning Settings. Three indispensable tasks are denoted as evidence identification (I), optic disc segmentation (S) and glaucoma diagnosis (D). By learning from the weak-label data, the WSMTL achieves the three tasks simultaneously. (b) Overview of the Weakly-Supervised Multi-Task Learning framework (WSMTL), which is trained with only the binary diagnostic labels.

idence maps as the bridge, then we leverage the evidence identification task by a pyramid integration structure and optic disc segmentation task via the constrained clustering branch, respectively. Finally, the diagnosis confidence is enhanced by use of the evidence and segmented optic disc information. The use of weakly-supervised multi-task framework is advantageous to evidence identification and optic disc segmentation, and learns relationships among multiple tasks from high level (diagnosis) to low level (segmentation) due to the fact that it provides an effective tool to model the correspondence from the diagnostic labels (image-level) to the spatial pixels. Benefit from those advantages, the WSMTL is capable of delivering effective evidence identification, optic disc segmentation and accurate glaucoma diagnosis simultaneously with one unified framework.

In summary, our main contributions are three-fold:

- For the first time, a unified framework (WSMTL) is proposed to simultaneously deliver three indispensable parts of clinical practice: evidence identification, optic disc segmentation and automated glaucoma diagnosis.
- An innovative weakly-supervised multi-task learning approach is proposed to endow the model with the ability to discover the evidence regions, obtain the optic disc mask and complete diagnosis learning from weak-label data (given only the binary diagnostic labels (normal/glaucoma)).
- A newly-designed CNN with skip connections and densely connected layer is developed to capture multi-scale task-aware features to release the limitation from high variability and extreme inhomogeneity of fundus structure.

Related Work

Weakly-supervised learning (WSL) has attracted great interests nowadays because the amount of data with weak-label annotations is much bigger and is growing much faster than that with complete annotations, especially in medical field. The major problem of weakly-supervised visual analysis is how to accurately assign weak labels such as image-level annotations to corresponding pixels of training images. Recently, many methods are emerging to establish the desired pixel-label correspondence in training for weakly-supervised learning (Tang et al. 2018; Huang et al. 2018; Kwak et al. 2017). Pinheiro et al. (2015) proposed to utilize multiple instance learning (MIL) which puts more weigh on pixels important for classifying the image during training to obtain the pixel labels from image level supervision. Papandreou et al. (2015) adopted an alternating training procedure based on the Expectation-Maximization algorithm to predict the attribution of pixels. Wei et al. (2016) proposed a simple to complex learning method to gradually enhance the segmentation network. Huang et al. (2018) proposed a semantic segmentation network starting from the discriminative regions and progressively increase the pixel-level supervision using by seeded region growing.

Recently, many weakly-supervised learning methods have tried to provide an effective solution for medical prediction. Zhang et al. (2017) proposed a weakly-supervised learning method to identify evidence regions of lumbar spinal stenosis for localising and classifying vertebrae in MRI images. Quellec et al. (2017) proposed a solution to create heatmaps showing the suspicious lesions where the pixels play a role in the image-level predictions. Gondal et al. (2017) proposed a CNN-based method to detect lesion areas for diabetic retinopathy with image-level label. Weakly-supervised localization, segmentation and identification are attracting more and more attentions since plenty of training data with

weak labels are readily available and efficiently obtained from diagnostic reports.

Multi-task learning (MTL) is a learning paradigm in machine learning and its aim is to leverage shared information contained in multiple related tasks to help improve the performance of all the tasks (Zhang and Yang 2017; Chen et al. 2017). Designing deep architectures for joint tasks is popular and effective, which has been used for several vision tasks (Pinheiro, Collobert, and Dollár 2015; Girshick 2015). Those networks fall into two categories. The one is trained separately for different tasks, and the configurations of different tasks are similar, but the learned parameters are totally different. The other method solves a common problem with two objective terms, such as generating object segmentation proposals. He et al. (2017) proposed a general framework to detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance by adding a branch for bounding box recognition. This architecture has led to an effective multi-task neural network methods, which use different network branch performing the segmentation and detection tasks by sharing features. In medical analysis, Feng et al. (2017) exploited CNN for automated detection and segmentation of pulmonary nodules on lung computed tomography (CT) scans. Wang et al. (2017) proposed a convolution neural network based algorithm for simultaneously diagnosing diabetic retinopathy and detecting suspicious regions. Different from existing methods, we propose a distinguished method to simultaneously deliver multiple indispensable tasks given only the binary diagnosis labels. We call this kind of learning method as *weakly-supervised multi-task learning* (WSMTL).

The Proposed Method

As shown in Fig.1(b), our weakly-supervised multi-task learning framework consists of four distinguishing parts: (1) a newly-designed CNN with skip connections and dense block to automatically capture the multi-scale feature representation, (2) a pyramid integration structure with multi-layer global pooling and activation pyramid mapping learning only from the diagnostic labels to generate high-resolution evidence map for evidence identification and segmentation, (3) a deep neural network, named as Constrained Clustering Branch (CCB), implements for optic disc segmentation, and (4) a fully-connected discriminator for automated glaucoma diagnosis. The proposed framework forms a tree structured network architecture, which use three different branches performing the evidence identification, glaucoma diagnosis and optic disc segmentation tasks, and shares the feature representation constructed by CNN backbone.

CNN with Skip Connections for Multi-scale Representation

In our WSMTL framework, CNN with skip connections and dense block are employed to capture multi-scale feature representation to deal with the challenges introduced by high variability and extreme inhomogeneity of fundus structure. Skip connections give the backbone CNN access to different convolutional layers directly to capture coarse high-level semantic features and low-level high-resolution features. Dense block allows the network to reuse and bypass existing features from prior layers and ensures high accuracies in later layers.

The WSMTL implements 4 dense blocks and inserts the transition layer between adjacent dense blocks to adjust the resolution of feature maps. The transition layers used in our experiments consist of a batch normalization layer and an 1×1 convolutional layer followed by a 2×2 average pooling layer. Each dense block having an equal number of layers is defined following the design in DenseNets (Lin et al. 2016), and we set the number of output channels of the three scales to 6, 6, 12 and 24, respectively. Each output of the dense block is directly connected to corresponding feature map to enhance the semantics for low-level feature and spatial information for high-level features. Features from different convolutional dense block are directly concatenated while bypassing intermediate layers to classify fundus image as normal or glaucomatous during training. This architecture helps the deep convolutional neural network generate multi-scale feature maps with accurate spatial and semantic information closely related to glaucoma diagnosis. The multi-scale feature representation helps the model discover more fine-grained details of the meaningful evidence regions which closely related to the fundus structure.

For a given fundus image, let $f_k^i(x, y)$ represents the feature map of channel k at scale i and spatial location (x, y), and P_k^i indicates the feature vector pooled from corresponding feature map. During training, the classification score $\boldsymbol{S}$ can be obtained by a weighted sum as

$$\boldsymbol{S} = \sum_i \sum_k w_k^i \boldsymbol{P}_k^i \tag{1}$$

where w_k^i is the weight corresponding to glaucoma for feature channel k and spatial scale i. Essentially, w_k^i is updated using the gradients backpropagation during the network is training with the diagnosis label, thence, it indicates the importance of $\boldsymbol{P}_k^i$ for glaucoma diagnosis. Finally, the output of the softmax for glaucoma is given by $\frac{exp(\boldsymbol{S})}{\sum exp(\boldsymbol{S})}$. Given the binary diagnosis label, the network learns a feature hierarchy consisting of multi-scale feature maps with a scaling step of 2, where the feature hierarchy possesses more high-level semantics to represent the glaucomatous changes of fundus structure.

Pyramid Integration Structure for Evidence Identification

As shown in Fig.2, a pyramid integration structure is implemented for evidence identification by constructing a high-resolution evidence map. The pyramid integration structure is constituted by two blocks, multi-layer global pooling for generation of pyramid activations which highlight the discriminative pixels, and activation pyramid mapping for construction of evidence map by integrating the pyramid activations. By means of the two blocks, the pyramid integration

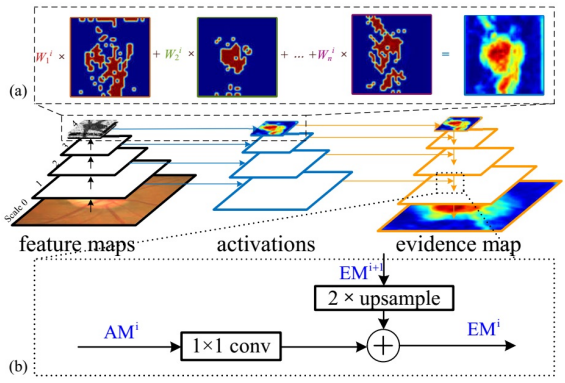

Figure 2: The pyramid integration structure for evidence identification. (a) multi-layer global pooling to produce glaucoma activation hierarchy highlighting discriminative pixels of feature maps to support glaucoma diagnosis. (b) activation pyramid mapping to generate evidence map by integrating activation hierarchy via upsampling and element-wise addition. The pyramid integration structure not only refines the spatial precision of the output evidence map, but also highlights more meaningful local pixels for evidence of glaucoma by merging the high-level glaucoma activations.

structure models the correspondence from binary diagnostic labels to spital pixels.

Multi-layer Global Pooling (MGP). MGP extends the global average pooling (GAP) to multiple convolutional layers to activate the discriminative pixels of fundus image at multiple scales. To identify the importance of feature map for glaucoma diagnosis, the weights w_k^i learned in discriminate layer are projected to feature maps. Let $\boldsymbol{AM}^i(x,y)$ represents the glaucoma activation at the i^{th} scale, then it is described more formally as follows

$$\boldsymbol{AM}^i(x,y) = \sum_k w_k^i f_k^i(x,y) \tag{2}$$

Pixels of $\boldsymbol{AM}$ with higher weights are activated with higher confidence to support glaucoma diagnosis. Accordingly, glaucoma activations highlight the discriminative pixels contributing to glaucoma diagnosis. Therefore, glaucoma activations bridges the gap between precise location and global semantics, and the correspondence from the diagnostic labels to the spatial pixels is effectively modeled.

Activation pyramid mapping (APM). The goal of APM is to construct a high-resolution evidence map by integrating the activation hierarchy generated by MGP via upsampling and element-wise addition. At each convolutional stage, the glaucoma activation $\boldsymbol{AM}^i(x,y)$ is merged with the evidence map $\boldsymbol{EM}^{i+1}$ from the top-down pathway. The pixel of evidence map with higher value indicates the higher confidence to evidence the glaucoma assessment. In this way, pyramid integration structure not only discovers more meaningful pixels or local regions for evidence identification, but also refines the spatial precision of evidence map for pixel-wise optic disc segmentation.

The basic block of APM is shown in Fig.2(b). The coarser-resolution evidence map $\boldsymbol{EM}^{i+1}$ from the top-down pathway is firstly upsampled by a factor of 2 to obtain the same spatial resolution as the corresponding glaucoma activation $\boldsymbol{AM}^i$. The upsampled evidence map $\widehat{\boldsymbol{EM}}^{i+1}$ is then merged with the corresponding glaucoma activations $\boldsymbol{AM}^i$ by element-wise addition as

$$\begin{aligned}\boldsymbol{EM}^i(x,y) &= \widehat{\boldsymbol{EM}}^{i+1}(x,y) + \boldsymbol{AM}^i(x,y) \\ &= \widehat{\boldsymbol{EM}}^{i+1}(x,y) + \sum_k w_k^i \boldsymbol{AM}_k^i(x,y)\end{aligned} \tag{3}$$

Constrained Clustering Branch for Optic Disc Segmentation

Our proposed CCB is a network head and used to reconstruct the segmentation mask of optic disc by implementing a constrained clustering algorithm, which clustering pixels with relational constraints. The method learns a similarity metric and clustering objective via deep neural network to achieve image segmentation incorporating with clustering algorithm.

In our WSMTL framework, CCB consists of a cluster assignment block, a pair generation block and a clustering objective. The cluster assignment block contains two fully-connected layers to generate assignment ("0" for background and "1" for foreground) for every pixel of input image. The output of the block represents a probabilistic assignment of a pixel to the cluster. The assignment between pixels and clusters are formed stochastically during optimization and is guided by the pairwise similarity. If there is a similar pair, their distribution should be similar, and they should be labeled with the same label and merged into one region. The pair generation block enumerates all pair of image pixels based on the 8-connected rule, which ensures that the clustering pixels are connected in the same region. The outputs of cluster assignment is enumerated in pairs before sending to the clustering objective. The proposed objective function is easily combined with the backbone network and optimized by stochastic gradient descent end-to-end.

The key to our method is the design of a clustering objective that can use pairwise information. In most approaches, the pairwise information is also called as constraints or similar pairs (Hsu, Lv, and Kira 2018). We use the pairwise KL-divergence to evaluate the distance between the assignment distributions, and use pre-learned similarity function to construct the contrastive loss as the clustering objective. Given a pair of pixels x_p, x_q, based on the cluster assignment block, their output are defined as $\mathcal{P}$ and $\mathcal{Q}$. If pixels x_p, x_q are dissimilar, the loss is given as

$$\mathcal{L}(x_p,x_q)^+ = \mathcal{D}_{KL}(\mathcal{P}\|\mathcal{Q}) + \mathcal{D}_{KL}(\mathcal{Q}\|\mathcal{P}) \tag{4}$$

$$\mathcal{D}_{KL}(\mathcal{P}\|\mathcal{Q}) = \sum_m^k (p_c log(\frac{p_c}{q_c})) \tag{5}$$

If pixels x_p, x_q come from a pair and are dissimilar, the hinge loss can be employed as

$$\mathcal{L}(x_p,x_q)^- = \mathcal{L}_t(\mathcal{D}_{KL}(\mathcal{P}\|\mathcal{Q}),\sigma) + \mathcal{L}_t(\mathcal{D}_{KL}(\mathcal{Q}\|\mathcal{P}),\sigma) \tag{6}$$

where

$$\mathcal{L}_t(x, \sigma) = max(x, \sigma - e) \tag{7}$$

Therefore, the total loss can be defined as clustering objective

$$\mathcal{L}(x_p, x_q) = \mathcal{F}(x_p, x_q)\mathcal{L}(x_p, x_q)^+ + (1 - \mathcal{F}(x_p, x_q))\mathcal{L}(x_p, x_q)^- \tag{8}$$

where $\mathcal{F}(x_p, x_q)$ is the similar function between pixels x_p and x_q.

In our experiments, the deep neural network proposed in (Zagoruyko and Komodakis 2015) is chosen to construct the similar function $\mathcal{F}(x_p, x_q)$. The network is used to predict image pixel similarity, while we use it to predict the pixels similarity based on their surroundings pixels. We use the cross-entropy loss and train it end-to-end. After training, the inference of similarity is achieved among all the pixel pairs in the input image. The output then be binaried as discrete similarity predictions and used by CCB as a similarity function.

Algorithm I summarizes the detail procedure of our optic disc segmentation algorithm. The algorithm contains four parts: Generate the initial pixel patches, similarity function pre-learning, training of CCB, inference of optic disc mask. For a excellent behavior, our segmentation algorithm first perform a preparation step to group pixels of evidence map into relative homogeneous patches. Compared with clustering directly to pixels, this strategy makes the segmentation algorithm more robust and less computation. Secondly, the similarity function $\mathcal{F}$ is learned by a pre-trained network proposed in (Zagoruyko and Komodakis 2015). Here, x_p, x_q denote the pixel patches. In this step, the pair generation block is also used to enumerate possible pairs of pixel patches. Then we refer to equation 8 as clustering loss $\mathcal{L}(x_p, x_q)$ to optimize the CCB. After the cluster assignment of CCB, the segmentation mask of the optic disc is obtained by retrieving the corresponding pixels in fundus image followed by an ellipse fitting algorithm.

Algorithm 1:

Input: Evidence map $\boldsymbol{EM}$
Output: Optic disc mask
STEP 1: Generate initial pixel patches
Generating the pixels patches $\boldsymbol{X} = x_1, x_2, ..., x_n$ by applying k-means.
STEP 2: Similarity function learning
Learning the similarity function $\mathcal{F}(x_p, x_q)$ for all the patch pairs using the network proposed in (Zagoruyko and Komodakis 2015);
STEP 3: CCB optimization
Optimizing the CCB using the clustering loss $\mathcal{L}(x_p, x_q)$ as equation 8.
STEP 4: Reference of Optic disc mask
Input a test evidence map , forward propagate the data through the CCB with trained weights, and get outputs for cluster assignment. Using ellipse fitting to obtain the optic disc segmentation mask.

Fully-Connected Discriminator for Glaucoma Diagnosis

Glaucoma diagnosis discriminates the input fundus image as the glaucomatous or normal case by the fully-connected discriminator, which comprehensively assesses features of raw fundus image, evidence map and optic disc segmentation. Firstly, we perform a simple convolutional layer with a kernel of 1×1 on the feature maps at each scale to obtain the relevant features. Secondly, we conduct a simple Gaussian filter on the segmented regions to capture the discriminative features for representation of optic disc. Finally, the multi-scale feature of fundus image and discriminative features of evidence map and optic disc are concatenation to establish a whole feature vector and input the soft-max function to determine whether the fundus belongs to glaucoma or normal.

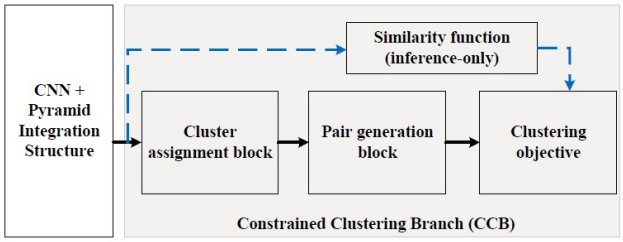

Figure 3: The Constrained Clustering Branch (CCB) for optic disc segmentation. The input is evidence map. The cluster assignment block is a two fully-connected layers and generates assignment ("0" for background and "1" for foreground) for every pixel of input image. Pair generation block enumerates all pair of image pixels based on the 8-connected rule. The parameters are trained by optimizing the cluster objective. When inference, it uses the forward propagation with only the cluster assignment to obtain the segmentation mask followed by an ellipse fitting.

Experiments

The effectiveness of the proposed WSMTL is validated in the three different tasks. Experimental results show that WSMTL successfully identifies clinically important yet easily missed evidence, and achieves 89.6% TP Dice of optic disc segmentation under 92.4% AUC for glaucoma diagnosis.

Dataset and Configurations

Dataset. Our WSMTL is validated with the challenging dataset ORIGA650 (Cheng et al. 2017) with 168 glaucomatous and 482 normal eyes. The 650 images with manual labeled optic disc mask are randomly divided into 325 training images (*Trainset*, including 73 glaucoma cases) and 325 testing images (*Testset*, including 95 glaucoma). We trained our method on the *Trainset* with only the diagnosis labels and tested the trained model on the *Testset* for evidence identification, optic disc segmentation and glaucoma diagnosis.

Data Augmentation. We augmented our data in order to reach our WSMTL the invariance properties. The types of augmentation used in this work include: (1) Randomly adjusting the optic disc coordinates based on Gaussian distribution in order to enhance the generation of our WSMTL, (2) Adding Gaussian noise directly to our image in order

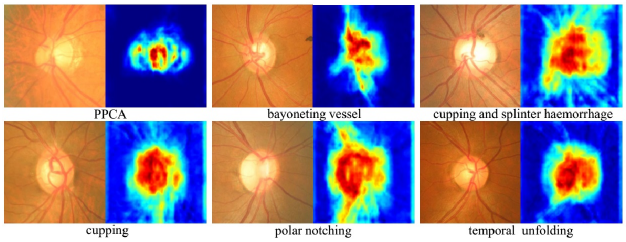

Figure 4: Examples of automatically discovered evidence regions by visualizing the top responses locations of the evidence map at different scales. Some easily missed abnormalities are retrieved based on high confidence evidence regions, and those abnormalities are clinically significant for glaucoma assessment, such as PPCA, bayoneting vessel, splinter haemorrhage, cupping, polar notching and temporal unfolding.

to simulate inherent noise, and (3) Horizontal reflecting the fundus images.

Training and Inference. The DenseNet is adopted as the backbone network and the last pooling layer *pool5*, and the FC layer, softmax are removed. The weights of remaining layers are initialized based on the pre-training on ImageNet. In our experiments, a multi-stage training strategy is developed so that all the evidence identification, optic disc segmentation and glaucoma diagnosis are achieved. First, we train the backbone CNN and pyramid integration structure together using the stochastic gradient descent with the only binary diagnosis labels. Because the diagnosis labels are high-level semantic information, we use it as supervision to train the deep neural network to discover the meaningful pixel region which is useful for evidence identification and segmentation. Then, based on the trained backbone CNN and pyramid integration structure, we train the CCB head via optimizing the clustering objective as equation 8. The trained WSMTL model is implemented to achieve the evidence identification, optic disc segmentation and glaucoma diagnosis simultaneously. As shown in the experiments, the weakly-supervised multi-task training and inference deliver excellent performance for all the three tasks.

Results and Analysis

Evidence Identification. As shown in Fig.3, The experimental results show that WSMTL obtains accurate visual evidence to highlight the abnormalities caused by glaucoma. Some clinically important abnormalities, such as PPCA, bayoneting vessel, splinter haemorrhage, cupping, polar notching and temporal unfolding, have been accurately identified by our WSMTL. In Fig.3, evidence regions are also visualized by showing the top responses location of evidence map with high confidence. Abnormalities, easily missed without the assistance of the evidence, are accurately located by retrieving the corresponding image regions. The clinically important evidence provides intuitive illustrations and interpretation for physicians and patients of how the diagnosis is made.

Optic Disc Segmentation and Glaucoma Diagnosis. As shown in Table.1, experimental results show that the proposed WSMTL achieves accurate optic disc segmentation (89.6% TP Dice) and glaucoma diagnosis (92.4% AUC). Here, TP Dice is employed to show the effectiveness of optic disc segmentation of glaucomatous fundus images since the precise segmentation of optic disc and accurate glaucoma diagnosis are mutually promoted. Compared with state-of-the-art methods(Maninis et al. 2016; Ronneberger, Fischer, and Brox 2015; Sedai et al. 2017), the proposed method increases the TP Dice by 4.7%, which indicates that our WSMTL identifies more precise contours of the abnormal optic disc. It is helpful for clinical glaucoma assessment to quantitative evaluation of optic disc. Fundamentally, the increasing of segmentation accuracy owes to multi-scale feature representation captured by skip connected CNN and high-resolution evidence map generated by pyramid integration structure. The results confirm that each component of the proposed framework is beneficial for accurate optic disc segmentation and glaucoma diagnosis.

Table.1 shows that the proposed WSMTL achieves the higher accuracy for glaucoma diagnosis since the model determines whether a unknown fundus is glaucoma or normal based on three sources: original features representation, extracted evidence map and segmentation mask. Our method obtains the highest AUC value of 0.924, which increases the AUC by 8.60% compared with (Fu et al. 2018), by 11.79% compared with (Cheng et al. 2013), by 20.2% compared with the conventional approach (Zhao et al. 2017). The accurate evidence identification and segmentation enhance diagnosis confidence.

Overall, the proposed WSMTL framework possesses the remarkable capableness and advantages on the two challenging tasks, which provides accurate quantitative evaluation for glaucoma assessment and reduces the rate of misdiagnosis in clinic. Therefore, the unified framework provides great help for clinical simultaneous optic disc segmentation and glaucoma diagnosis.

Conclusions

We proposed a novel Weakly-Supervised model to simultaneously achieve three clinical tasks: evidence identification, optic disc segmentation and glaucoma diagnosis. The model can be trained only with the binary diagnosis labels (normal/glaucoma), while obtains pixel-level evidence map, segmentation mask and diagnosis prediction simultaneously. This model is named as Weakly-Supervised Multi-Task Learning (WSMTL) in this paper. Specially, we develop a WSMTL framework which implements the newly-designed CNN for multi-scale representation of fundus image, the pyramid integration structure for evidence identification, the Constrained Clustering Branch (CCB) for optic disc segmentation, and the fully-connected discriminator for glaucoma diagnosis. By taking advantages of the newly-designed WSMTL framework, the proposed weakly-supervised model is capable of simultaneously delver the three indispensable parts of clinical practice in a unified model given only the binary diagnostic labels, which possesses a great potential for clinical assessment of fundus image.

Table 1: Performance of WSMTL used for optic disc segmentation and glaucoma diagnosis. Three criteria are evaluated to compare with state-of-the-art methods. High TP Dice indicates good performance on simultaneous segmentation and diagnosis.

Method	Dice	TP Dice[1]	AUC
Fully-supervised			
U-Net(Ronneberger, Fischer, and Brox 2015)	**0.87±0.09**	0.85±0.10	-
DRIU(Maninis et al. 2016)	0.82±0.09	0.81±0.11	-
Semi-supervised			
VAE(Sedai et al. 2017)	0.87±0.06	0.84±0.09	-
Our weakly-supervised			
1-layer MGP w/ APM	0.82±0.08	0.83±0.07	0.89
2-layers MGP w/ APM	0.85±0.07	0.86±0.05	0.91
3-layers MGP w/ APM	**0.87±0.06**	**0.89±0.04**	**0.92**
4-layers MGP w/ APM	0.86±0.07	0.88±0.06	0.92
3-layers MGP w/o APM	0.82±0.12	0.83±0.09	0.90

[1] TP Dice = Dice coefficient over truly detected glaucomatous images.

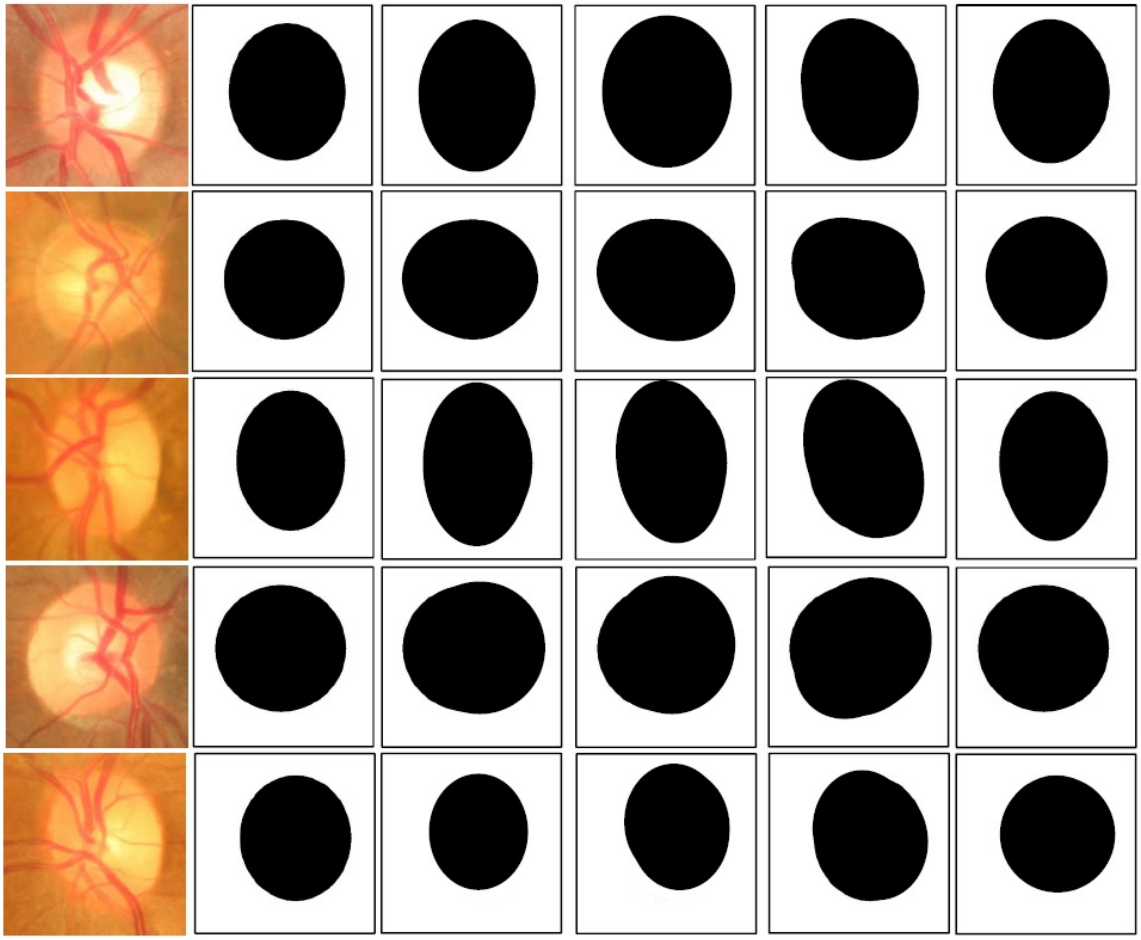

Figure 5: The visual examples of optic disc segmentation mask, where the white denotes the disc segmentations, while black denotes background. From left to right: fundus image, ground truth (GT), Unet, DRIU, VAE and our proposed method.

References

Almazroa, A.; Burman, R.; Raahemifar, K.; and Lakshminarayanan, V. 2015. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. *Journal of ophthalmology* 2015.

Chen, X.; Xu, Y.; Yan, S.; Wong, D. W. K.; Wong, T. Y.; and Liu, J. 2015. Automatic feature learning for glaucoma detection based on deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 669–677. Springer.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. A multi-task deep network for person re-identification. In *Pro-*

ceedings of the 31st AAAI Conference on Artificial Intelligence, 3988–3994.

Cheng, J.; Liu, J.; Xu, Y.; Yin, F.; Wong, D. W. K.; Tan, N.-M.; Tao, D.; Cheng, C.-Y.; Aung, T.; and Wong, T. Y. 2013. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Transactions on Medical Imaging* 32(6):1019–1032.

Cheng, J.; Zhang, Z.; Tao, D.; Wong, D. W. K.; Liu, J.; Baskaran, M.; Aung, T.; and Wong, T. Y. 2017. Similarity regularized sparse group lasso for cup to disc ratio computation. *Biomedical optics express* 8(8):3763–3777.

Feng, X.; Yang, J.; F.Laine, A.; and D. Angelini, E. 2017. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 568–576. Springer.

Fu, H.; Cheng, J.; Xu, Y.; Zhang, C.; Wong, D. W. K.; Liu, J.; and Cao, X. 2018. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Transactions on Medical Imaging* 37(11):2493–2501.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Gondal, W. M.; Köhler, J. M.; Grzeszick, R.; Fink, G. A.; and Hirsch, M. 2017. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In *Image Processing (ICIP), 2017 IEEE International Conference on*, 2069–2073. IEEE.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988. IEEE.

Hsu, Y. C.; Lv, Z.; and Kira, Z. 2018. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*.

Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7014–7023.

Kwak, S.; Hong, S.; Han, B.; et al. 2017. Weakly supervised semantic segmentation using superpixel pooling network. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4111–4117.

Lin, K.; Lu, J.; Chen, C. S.; and Zhou, J. 2016. Learning compact binary descriptors with unsupervised deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 1183–1192.

Maninis, K.-K.; Pont-Tuset, J.; Arbeláez, P.; and Van Gool, L. 2016. Deep retinal image understanding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 140–148. Springer.

Papandreou, G.; Chen, L.-C.; Murphy, K.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*.

Pinheiro, P. O., and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1713–1721.

Pinheiro, P. O.; Collobert, R.; and Dollár, P. 2015. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, 1990–1998.

Quellec, G.; Charrière, K.; Boudi, Y.; Cochener, B.; and Lamard, M. 2017. Deep image mining for diabetic retinopathy screening. *Medical image analysis* 39:178–193.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.

Sedai, S.; Mahapatra, D.; Hewavitharanage, S.; Maetschke, S.; and Garnavi, R. 2017. Semi-supervised segmentation of optic cup in retinal fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 75–82. Springer.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *arXiv preprint arXiv:1807.03342*.

Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; and et.al. 2017. Zoom-in-net: deep mining lesions for diabetic retinopathy detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 267–275. Springer.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M. M.; Feng, J.; Zhao, Y.; and Yan, S. 2016. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(11):2314–2320.

Zagoruyko, S., and Komodakis, N. 2015. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4353–4361.

Zhang, Y., and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhang, Q.; Bhalerao, A.; and Hutchinson, C. 2017. Weakly-supervised evidence pinpointing and description. In *International Conference on Information Processing in Medical Imaging*, 210–222. Springer.

Zhao, R.; Chen, Z.; Duan, X.; Chen, Q.; Liu, K.; and Zhu, C. 2017. Automated glaucoma detection based on multi-channel features from color fundus images. *Journal of Computer-Aided Design and Computer Graphics* 998–1006.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.