Medical Image Analysis 63 (2020) 101695

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning

Xi Wang[a], Hao Chen[a,*], An-Ran Ran[b], Luyang Luo[a], Poemen P. Chan[b], Clement C. Tham[b], Robert T. Chang[c], Suria S. Mannil[c], Carol Y. Cheung[b], Pheng-Ann Heng[a]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[b] Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China
[c] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 31 October 2019
Revised 2 February 2020
Accepted 30 March 2020

Keywords:
Optical coherence tomography
Deep learning
Glaucoma screening
Semi-supervised multi-task learning

ABSTRACT

Glaucoma is the leading cause of irreversible blindness in the world. Structure and function assessments play an important role in diagnosing glaucoma. Nowadays, Optical Coherence Tomography (OCT) imaging gains increasing popularity in measuring the structural change of eyes. However, few automated methods have been developed based on OCT images to screen glaucoma. In this paper, we are the first to unify the structure analysis and function regression to distinguish glaucoma patients from normal controls effectively. Specifically, our method works in two steps: a semi-supervised learning strategy with *smoothness assumption* is first applied for the surrogate assignment of missing function regression labels. Subsequently, the proposed multi-task learning network is capable of exploring the structure and function relationship between the OCT image and visual field measurement simultaneously, which contributes to classification performance improvement. It is also worth noting that the proposed method is assessed by two large-scale *multi-center* datasets. In other words, we first build the largest glaucoma OCT image dataset (i.e., *HK* dataset) involving 975,400 B-scans from 4,877 volumes to develop and evaluate the proposed method, then the model without further fine-tuning is directly applied on another independent dataset (i.e., *Stanford* dataset) containing 246,200 B-scans from 1,231 volumes. Extensive experiments are conducted to assess the contribution of each component within our framework. The proposed method outperforms the baseline methods and two glaucoma experts by a large margin, achieving volume-level Area Under ROC Curve (AUC) of 0.977 on *HK* dataset and 0.933 on *Stanford* dataset, respectively. The experimental results indicate the great potential of the proposed approach for the automated diagnosis system.

1. Introduction

Glaucoma is the most frequent cause of irreversible blindness worldwide, which is a heterogeneous group of diseases that damages the optic nerve and can result in vision loss (Bourne et al., 2013). It is projected to affect 111.8 million people in 2040 (Tham et al., 2014). Glaucoma is featured by loss of retinal ganglion cells, thinning of retinal nerve fiber layer (RNFL), and cupping of the optic disc (Jonas et al., 2017). In reality, the most common form of glaucoma (i.e., open-angle glaucoma) is usually chronic and painless. The self-detection of visual field (VF) defect always occurs at the advanced stage of glaucoma. As a consequence, a substantial proportion of glaucoma patients remain undiagnosed. Early diagnosis via ophthalmological examination is essential as it could facilitate immediate treatment for early-stage glaucoma patients and delay the progression of the disease.

On the clinical background, the structure assessment and the function measurement are necessary and combined to diagnose and manage glaucoma. The structure assessment includes the evaluation of the optic disc and RNFL, which is performed through imaging techniques, like fundus photography and optical coherence tomography (OCT). In general, fundus photography provides the top view of the retina and optic nerve head (ONH). In the last decade, majority of works in literature are dedicated to the analysis on fundus images for screening glaucoma (Chen et al., 2015; Cerentinia et al., 2018; Masumoto et al., 2018; Shibata et al., 2018; Li et al., 2018c; Medeiros et al., 2018; Christo-

* Corresponding author.
E-mail address: hchen@cse.cuhk.edu.hk (H. Chen).

https://doi.org/10.1016/j.media.2020.101695
1361-8415/

pher et al., 2018a; Raghavendra et al., 2018). Although fundus imaging is the easiest test in clinic, the major drawback is that there is very limited information characterizing glaucoma in such modality. On the other hand, OCT is a non-contact and non-invasive imaging modality that generates high-resolution, cross-sectional images (i.e., B-scans) of the retina. It can provide an objective and quantitative assessment of various retinal structures. Owing to its rich and comprehensive information for glaucoma diagnosis, OCT has gained more and more popularity at present.

The most commonly used functional testing in the context of glaucoma evaluation is the standard automated perimetry, which is regarded as the clinically gold standard for the assessment of visual function. It provides a systematic way for clinicians to assess the regions of a patient's field of vision affected by glaucoma and the severity of vision loss (Lucy and Wollstein, 2016). In general, VF test can generate three important global indices, named *VF measurement*: visual field index (VFI), mean deviation (MD), and pattern standard deviation (PSD). Explicitly, VFI is a global metric that represents the entire VF as a single number in percentage, where 100% denotes a normal VF while 0% represents a perimetrically blind field (Iutaka et al., 2017). MD reflects the overall depression of the field (Yaqub, 2012), which is a weighted average decibel deviation from age normal database. The lower the MD value, the more damaged the visual function (Iutaka et al., 2017). Moreover, the PSD value is the standard deviation of the departure between the measured VF pattern and the normal hill of vision, which reflects the roughness of the VF. These VF indices have already been studied in the previous researches to screen glaucoma (Silva et al., 2013; Asaoka et al., 2016; Kim et al., 2017).

Automated glaucoma OCT image screening tool is quite appealing in clinical practice as the examination of OCT imaging usually requires highly trained ophthalmologists and is always partly subjective and time-consuming. Moreover, evaluation of the relationship between structural and functional damage can provide valuable insight into how visual function works according to the degree of the structural damage, which can help our understanding of glaucoma. This is quite significant in diagnosing, staging, and monitoring glaucoma patients. OCT image analysis has attracted a great number of researchers in recent years (Wang and Wang, 2019; Fang et al., 2019; Maetschke et al., 2019). Nevertheless, most of the previous works were developed based on machine learning methods and heavily relied on established features, such as the measurements on RNFL thickness and ganglion cell layer thickness (Huang and Chen, 2005; Medeiros et al., 2009; Kim et al., 2016; Christopher et al., 2018b). A pioneering work (Maetschke et al., 2019) recently proposed a 3D convolutional neural network (CNN) to directly classify the downsampled OCT volumes into glaucoma or normal, which considerably outperformed various feature-based machine learning algorithms. However, there are still three main challenges that have not been fully investigated yet. First, heretofore there is a dearth of studies that explores the structure and function relationship based on the raw OCT image and VF measurement for glaucoma screening. Second, owing to very limited images in current datasets, validation experiments of the previous methods are not comprehensive, which indeed constrains the development of robust and reliable approaches. Third, constructing a large medical dataset is always confronted with many difficulties and the missing label problem often occurs since it is inevitable to acquire incomplete clinical records due to some unexpected reasons, which is a common phenomenon in retrospective studies.

Here, we carried a thorough investigation on all of the challenges mentioned above. As we know, the OCT image is of high dimension and contains multiple B-scan images (e.g., $200 \times 200 \times 1024$), as illustrated in Fig. 1. Different from diabetic macular edema (DME) whose pathology or signs may only appear in a small fraction of B-scans (i.e., very sparse distribution), glaucoma can progressively cause structural changes in the whole retina. Thus a substantial number of B-scans within a volume could carry characteristics of glaucoma, and it may not require extra efforts on B-scan labels unlike some DME OCT image studies (Rasti et al., 2018b; 2018a). Notably, training a 3D CNN with large-scale volumes could be both memory- and computation-consuming. Besides, due to the scarcity of pre-trained 3D model available currently, training from scratch could easily lead to the local optima issue. Alternatively, it is much more feasible and reasonable to train a B-scan-wise deep network using B-scan images inheriting the volume-level label. There are two basic merits. First, the 2D CNN can enjoy transferred features learned from the large-scale natural images (e.g., ImageNet, Deng et al., 2009). Second, one OCT volume can provide multiple B-scan slices, which can substantially enrich the training set and thus effectively alleviate the over-fitting problem. As a result, we adopt a B-scan-wise network to tackle the glaucoma classification problem. Following Ben-Cohen et al. (2016) and Li et al. (2018b), we select every three successive B-scan images to form the 3-channel input so as to incorporate more information for every input. The volume-level label is propagated to every B-scan input.

In this paper, we proposed a semi-supervised multi-task learning network to address the glaucoma OCT image screening problem. A preliminary version of this work was presented in Wang et al. (2019b). The major differences are as follows: First, we elaborated a comprehensive literature review on glaucoma screening related tasks from the fundus image, VF, and OCT images. Second, we designed and conducted a number of experiments to verify the effectiveness of the B-scans sampling strategy. Besides, various networks, including VGG16 (Simonyan and Zisserman, 2014), ResNet18 (He et al., 2016), ResNet50 (He et al., 2016), DenseNet121 (Huang et al., 2017) and MobileNet (Howard et al, 2017), were compared and analyzed. We also investigated the impact of the hyper-parameter α in the weighted loss function on the joint training performance and used different similarity measures for surrogate assignment to improve the classification performance. Third, We reported the regression results on *HK* dataset, and further validated the robustness and generalization of our method on another dataset from Byers Eye Institute, Stanford University in the USA. Lastly, we thoroughly analyzed the advantages and limitations of our method and discussed about domain adaption to improve the performance on the cross-center datasets. Our main contributions are summarized as follows.

(1) To the best of our knowledge, we are the first to unify structure analysis and function regression for glaucoma screening based on OCT images. We develop a novel framework that explores the structure and function relationship between OCT image and VF measurement via a semi-supervised multi-task learning network.
(2) We build the largest glaucoma OCT dataset composed of 975,400 B-scans from 4,877 volumes of the optic disc in this study for algorithm development and evaluation.
(3) The proposed method achieves the best performance compared to other competitive methods. We evaluate the proposed method on two datasets, and the experimental results demonstrate its efficacy on recognition of glaucoma, with 92.7%, 94.1%% and 97.7% on the accuracy, F1 score and AUC at volume level on *HK* dataset, respectively. The mean absolute errors for regression tasks are 0.105, 3.508, and 2.034 on VFI, MD, and PSD, respectively. In particular, the model without further fine-tuning also achieves a promising AUC of 93.3% at volume level and 93.7% at eye-visit level on *Stanford* dataset.

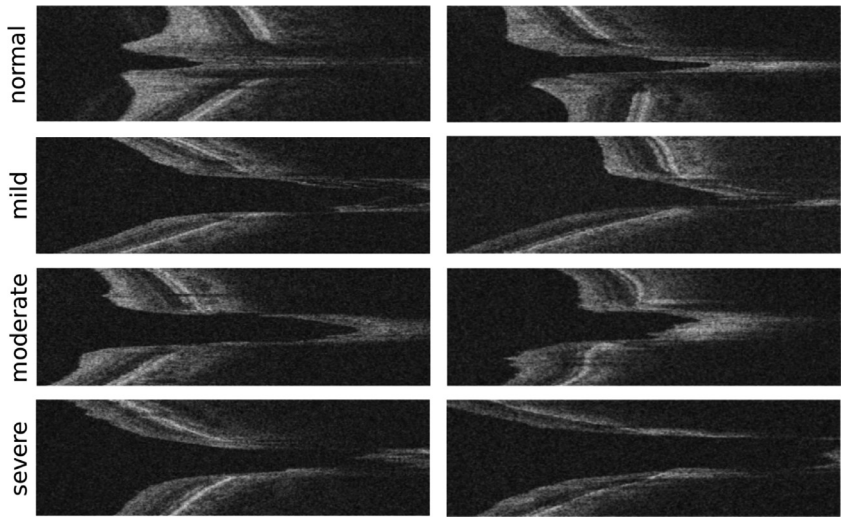

Fig. 1. Illustration of typical 2D B-scan images from volumetric OCT scans of the optic disc. The top panel shows the normal B-scans. The second to the bottom panel belong to mild glaucoma, moderate glaucoma, and severe glaucoma B-scans, respectively.

2. Related work

In this section, we will have a comprehensive literature review on automated glaucoma screening approaches. At first, we briefly recap the studies on fundus photos and VF test results, and then we revisit researches on OCT images, which are more relevant to our study in detail.

Analysis on color fundus images takes up the largest portion of the studies for glaucoma screening (Chen et al., 2015; Cerentinia et al., 2018; Masumoto et al., 2018; Shibata et al., 2018; Li et al., 2018c; Medeiros et al., 2018; Christopher et al., 2018a; Raghavendra et al., 2018). With the advance of deep learning techniques, CNNs continuously realized astonishing achievements in the field of computer vision (Krizhevsky et al., 2012; LeCun et al., 2015; Szegedy et al., 2015; He et al., 2016) and medical applications (Bejnordi et al., 2017; De Fauw et al., 2018; Chen et al., 2018; Wang et al., 2018; Song et al., 2015; 2016; Luo et al., 2019). Classic deep networks, e.g., ResNet, GoogleNet, VGG16, InceptionV3, have been widely leveraged by researchers to solve this classification problem, taking either the whole fundus image or the manually extracted region of interest as input (Cerentinia et al., 2018; Shibata et al., 2018; Li et al., 2018c; Medeiros et al., 2018; Christopher et al., 2018a). These methods have achieved promising results on different public datasets. Some researchers utilized VF test results to identify glaucoma (Asaoka et al., 2016; Kim et al., 2017; Li et al., 2018a; Kucur et al., 2018; Asaoka et al., 2019). For example, Asaoka et al. (2016) collected 52 features obtained by VF test as predictors and employed a deep feed-forward neural network to make predictions. Differently, a recent work took the pattern deviation plot from VF test as the input features of the deep network (Li et al., 2018a).

However, the usage of fundus photography is limited to the detection of signs visible in the retinal surface, and it cannot always identify the internal changes of the RNFL and ONH. By contrast, OCT imaging has an edge on generating cross-sectional images that embed important glaucomatous optic neuropathy related features from the entire retinal layered structure. Recently, an increasing number of researchers have shown interest in the analysis of glaucoma OCT images. Most of the existing approaches highly relied on the off-the-shelf features that could be readily obtained by OCT devices, and researchers were inclined to apply various machine learning methods like Bagging, Random Forest, and Multilayer Perceptron (Huang and Chen, 2005; Kim et al., 2016; Christopher et al., 2018b). The established features, such as the measurement of RNFL thickness, ONH thickness, and macular thickness, were extensively used in these works. For instance, Christopher et al. (2018b) proposed an unsupervised machine learning method to identify glaucoma on RNFL map features extracted from the OCT scan after careful dimension reduction. Besides, Kim et al. (2016) analyzed the topographic patterns of segmented thicknesses in open-angle glaucoma patients compared with healthy subjects. They found that the diagnostic ability of segmented mRNFL and GCL to discriminate between normal and glaucoma eyes was high and comparable to that of cpRNFL thickness (Kim et al., 2016). Some researchers also dedicated their efforts to detect and segment the optic disc and ONH using traditional approaches for OCT image analysis (Lee et al., 2009; Fu et al., 2014). Besides, cup-to-disc ratio (CDR) is another important indicator for glaucoma identification (Almazroa et al., 2015; Ramzan et al., 2018; Fu et al., 2018). Lots of previous studies tried to segment the optic disc and the optic cup from fundus photography and calculate the ratio for glaucoma screening (Almazroa et al., 2015), while Ramzan et al. (2018) proposed to extract inner limiting membrane (ILM) and retinal pigment epithelium (RPE) layers from OCT scans, then both the ILM layer and RPE breakpoints were used for quantifying the optic cup and the optic disc, respectively. Based on the calculated CDR, their method was able to classify the scans as normal or glaucomatous, with 87% of sensitivity, 72% of specificity, and 79% of accuracy. A recently published work (De Fauw et al., 2018) also developed a similar pipeline, where a segmentation network was first used to delineate 15 different retinal morphological features and OCT acquisition artifacts, and then the segmentation output was passed to the next classification network to make a referral triage decision from four categories and classify the presence of 10 different OCT pathologies. However, there was still a deficiency of approaches that could directly classify glaucoma OCT images until the 3D CNNs were employed to distinguish glaucoma from normal controls (Ran et al., 2019; Maetschke et al., 2019). In Maetschke et al. (2019), the OCT volumes were first downsampled into cubes with a much smaller size, and then they were fed into a 3D CNN to predict the class labels. More specifically, the 3D network was stacked by five successive 3D convolutional layers with zero-paddings, a global average pooling layer and a fully connected layer with the softmax activation. This 3D model significantly overwhelmed all other feature-based machine learning techniques with established features, achieving the best AUC of

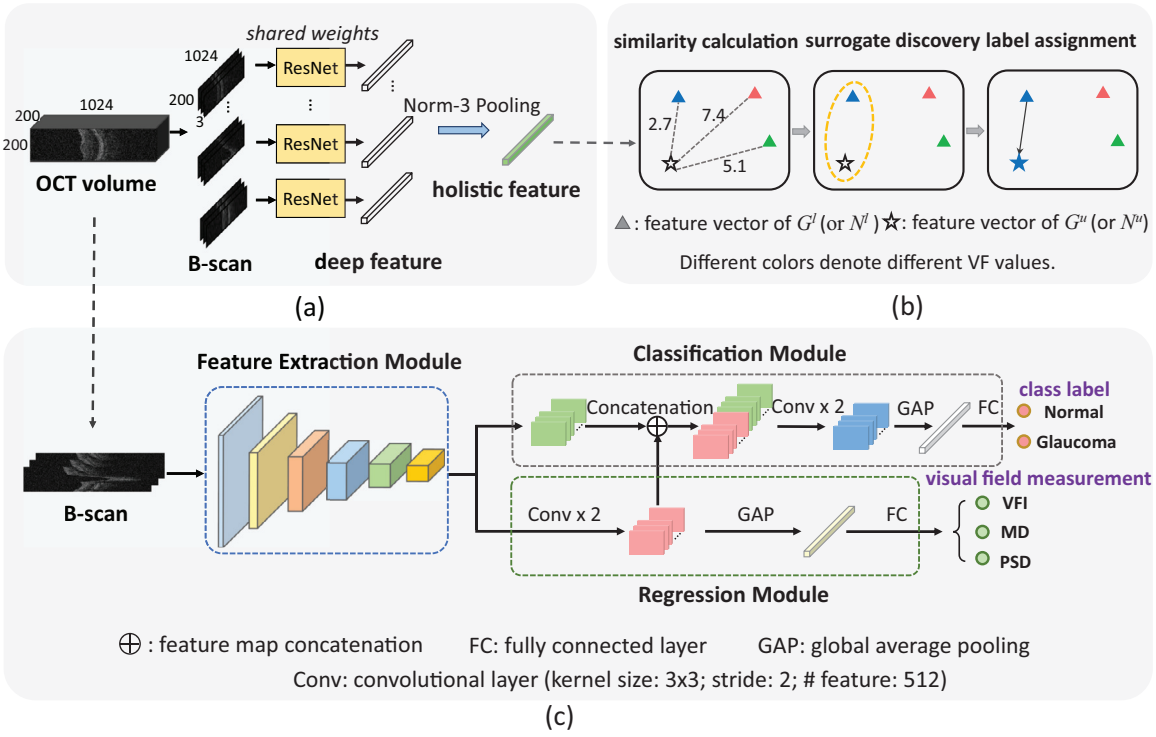

Fig. 2. An overview of the proposed method. (a) Initially, we train a CNN model with class labels (glaucoma/normal) to extract features of B-scans. Then the extracted features are aggregated to form a global representation of an OCT volume. (b) Next, we compute similarities between homogeneous groups to find the nearest neighbor of OCT images without VF measurement and enable its VF values for surrogate assignment. Here, G^l and G^u indicate glaucoma images with and without VF labels, while N^l and N^u denote normal images with and without VF labels. (c) Lastly, class labels and VF labels (ground truth and pseudo labels) are unified to train a multi-task learning network.

0.94. Nevertheless, the method was validated on a very small test dataset.

At present, there remains a dearth of studies that investigates the structure and function relationship from the OCT image and VF measurement for glaucoma analysis. To best of our knowledge, we are the first to make use of these clinical data to figure out whether such a relationship exists or not. Inspired by the previous works using multi-task learning networks for medical image analysis (Chen et al., 2016; 2017; Liu et al., 2018), e.g., the risk stratification and attribute score regression of lung nodules in CT images (Chen et al., 2017; Liu et al., 2018), we propose to incorporate structure analysis and function regression in a multi-task learning network. Specifically, it takes OCT images as input, and class labels and VF measurement as supervisions. Owing to incomplete VF measurement labels in the self-built dataset, a semi-supervised *smoothness assumption* based method is first applied to address the missing label problem ahead of glaucoma classification.

3. Methods

The main task of this study is to effectively classify glaucoma OCT images assisted by exploring the relationship between structure and function of glaucoma. Fig. 2 illustrates the overview of the proposed method. Specifically, it consists of two parts. The first part uses a semi-supervised learning technique to address the missing label problem of VF measurement for function regression. In particular, a CNN is trained under the full supervision of class labels, aiming to extract deep features from B-scan images. After B-scan feature aggregation, each OCT volume is represented by a holistic feature vector, which is used to calculate the similarity between any pair of homogeneous OCT images. Afterwards, pseudo labels are automatically propagated from the nearest neighbor. In the second part, a multi-task learning network is trained to unify structure analysis and surrogate-driven function regression for more accurate glaucoma screening.

3.1. Surrogate-driven labeling with semi-supervised learning

To solve the problem of missing regression labels for our multi-task learning, we borrow the spirit from semi-supervised learning and come up with an appropriate solution. In semi-supervised learning domain, the *smoothness assumption* points out that features close to each other are more likely to share the same label. This assumption is intimately linked to a definition of what it means for one feature to be near another feature, which can be embodied in a similarity function $S(\cdot, \cdot)$ on the input space (Chapelle et al., 2009).

To project OCT volumes into the input space for similarity calculation, there are a variety of representation approaches. Utilizing hand-crafted features (e.g., the Local Binary Patterns, Histogram of Oriented Gradient features, and RNFL thickness) is a very traditional method to depict OCT images (Lemaître et al., 2016; Alsaih et al., 2017). Nevertheless, great efforts are required to design and validate the effectiveness of these features. With the advance of CNN, deep neural networks, like auto-encoder, siamese networks, and generative adversarial networks (GANs), have demonstrated their superiority on feature representation (Kingma and Welling, 2013; Melekhov et al., 2016; Donahue et al., 2016). However, OCT volumes are always of high dimension, thus it is impractical to directly feed OCT volumes to train an auto-encoder or not cost-effective to train sophisticated networks (e.g., GANs) for feature embedding. Moreover, abstracting off-the-shelf features by available ImageNet pretrained models is also an alternative for feature embedding. However, these features are quite generic and always in lack of representation ability. To this end, we train a 2D CNN classifier as the feature representation model. Specifically,

this network first is pre-trained by ImageNet and then fine-tuned with B-scan images under the full supervision of class labels (glaucoma/normal). Next, the trained model is used to extract deep features of B-scan inputs from the Global Average Pooling (GAP) layer. Finally, following Wang et al. (2019a), we perform the *norm-3 pooling* below to aggregate B-scan features into a holistic representation $\mathcal{F}$ for each OCT volume:

$$\mathcal{F} = (\sum_{i=1}^{n} f_i^3)^{\frac{1}{3}} \tag{1}$$

where f is the feature representation of a B-scan input, and n is the number of B-scans from the same volume.

According to the type of class labels and the availability of the VF measurement, the training set is grouped into four clusters: (1) G^l: glaucoma images with VF measurement; (2) G^u: glaucoma images without VF measurement; (3) N^l: normal images with VF measurement and (4) N^u: normal images without VF measurement. Here, we call G^l and G^u (or N^l and N^u) homogeneous groups. Through our feature representation model and feature aggregation, each OCT volume is represented by a feature vector, i.e., $\mathcal{F} \in \mathbb{R}^m$ where m is the length of $\mathcal{F}$. In this study, several widely used similarity measures, including Euclidean distance, Cosine similarity, and Manhattan distance, are employed to quantifies the similarity between any pair of homogeneous OCT images (i.e., $\mathcal{F}^l \in G^l$ and $\mathcal{F}^u \in G^u$), as shown below:

$$S_{euclidean}(\mathcal{F}^l, \mathcal{F}^u) = [\sum_{i=1}^{m} (\mathcal{F}_i^l - \mathcal{F}_i^u)^2]^{\frac{1}{2}} \tag{2}$$

$$S_{cosine}(\mathcal{F}^l, \mathcal{F}^u) = \frac{\sum_{i=1}^{m} \mathcal{F}_i^l \mathcal{F}_i^u}{\sqrt{\sum_{i=1}^{m} (\mathcal{F}_i^l)^2} \sqrt{\sum_{i=1}^{m} (\mathcal{F}_i^u)^2}} \tag{3}$$

$$S_{manhattan}(\mathcal{F}^l, \mathcal{F}^u) = \sum_{i=1}^{m} \mid \mathcal{F}_i^l - \mathcal{F}_i^u \mid \tag{4}$$

The larger S_{cosine} is, the more similar two feature vectors are. But for Manhattan distance and Euclidean distance, the larger $S_{manhattan}$ or $S_{euclidean}$ is, the more dissimilar two features are.

According to the *semi-supervised smoothness assumption*, for each OCT volume without VF measurement $\mathcal{F}_j^u \in G^u$ (or $\mathcal{F}_j^u \in N^u$), we first find its nearest neighbor in its homogeneous group that has VF measurement $\mathcal{F}_{i^*}^l \in G^l$ (or $\mathcal{F}_{i^*}^l \in N^l$), where $i^* = \arg\min_i S(\mathcal{F}_i^l, \mathcal{F}_j^u)$, and then appoint the VF measurement of $\mathcal{F}_{i^*}^l$ to $\mathcal{F}_j^u$. Consequently, we manage to find suitable surrogates for all missing VF measurements in this semi-supervised learning fashion.

3.2. Multi-task learning for structure and function analysis

We build an end-to-end multi-task learning CNN with a primary task to classify B-scan images into glaucoma and normal, and an auxiliary task to investigate the relationship between structural and functional changes of glaucoma eyes. Particularly, this network is composed of three components: a shared feature extraction module, a classification module, and a regression module. As illustrated in Fig. 2(c) and Fig. 3, we employ ResNet18 as the backbone in the feature extraction module whose weights are shared by both classification and regression tasks. The rest of the network consists of two branches, one for glaucoma discrimination and the other for VF measurement regression. The two tasks are trained jointly.

3.2.1. Visual field measurement regression

For each attribute of VF measurement, i.e., VFI, MD, and PSD, we formulate it as an individual regression task. It is worth mentioning that we scale the real value of MD and PSD into the range of [0,1] ahead of the model training. Within the regression module, two convolutional layers with ReLU activation and batch-normalization are inserted before the GAP layer. Fully connected layers with Sigmoid activation are then individually used to regress these attributes. The regression tasks are driven by minimizing the mean square error:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^{N} \| y_i^r - \hat{y}_i^r(x_i; \theta_s, \theta_{reg}) \|_2^2 \tag{5}$$

where N denotes the number of training samples, x_i is the input of three adjacent B-scan images, y_i^r is the clinical measurement of VF attribute, and $\hat{y}_i^r$ is the corresponding prediction of the network. θ_s denotes shared weights in the feature extraction module, and θ_{reg} denotes features in the regression module, respectively. In this paper, we use superscripts r and c for discrimination between **r**egression and **c**lassification task.

3.2.2. Glaucoma classification

The classification module performs the primary task for glaucoma screening. In clinical routine practice, the VF measurement is an essential indicator of functional change for glaucoma diagnosis. Hence, it is reasonable to hypothesize that if the relationship between structure and function is appropriately discovered, the learned features in the regression module could exert a positive influence on the classification task. Based on this assumption, we concatenate the attribute regression feature maps with those originated from the feature extraction module, supposing that the features learned in the regression module could provide the classifier with helpful guidance. After our feature aggregation layer, two convolutional layers with ReLU activation and batch-normalization are appended, followed by a GAP layer. Lastly, a fully connected layer with softmax activation outputs class probabilities. Here, the typical binary cross-entropy loss is utilized for training this classifier:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} y_i^c \log \hat{y}_i^c(x_i; \theta_s, \theta_{cls}) \tag{6}$$

where y_i^c is the ground truth label while $\hat{y}_i^c$ is the likelihood predicted by the classifier. Similarly, θ_{cls} stands for the weights in the classification module.

3.2.3. Joint training of multi-task learning network

Finally, the multi-task learning network is trained by minimizing the weighted combination of the mean square error losses and the binary cross-entropy loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \sum_{j=1}^{3} \alpha_j \mathcal{L}_{reg}^j \tag{7}$$

where α_j is the hyper-parameter balancing $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}^j$.

At the testing stage, we apply the following volume-level inference strategy to obtain the final prediction: For any unseen volume, we feed a fixed number of B-scan inputs (i.e., 78 3-channel inputs sampled from the 60th to 139th B-scans) into our multi-task learning model and get the class likelihood and regression values from the classification branch and the regression branch, respectively. Afterwards, we take the average of B-scan-level predictions as the volume-level result.

4. Experiments and results

4.1. Materials and datasets

In this study, we constructed the largest scale glaucoma OCT image cohort with the collaboration of The Chinese University of

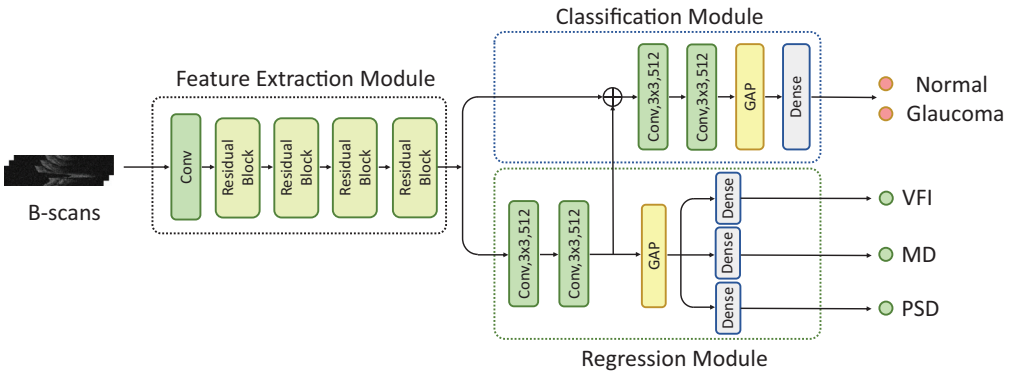

Fig. 3. The multi-task learning network for glaucoma classification and visual field measurement regression.

Table 1
Data distribution in *HK* and *Stanford* datasets.

	Volumes	B-scans	Patients	Eyes	Cases
HK	4,877	975,400	930	1,587	3,184
Stanford	1,231	246,200	305	583	1,148

Table 2
Number of OCT volumes and cases in training, validation and testing sets. *G* and *N* denote glaucoma and normal respectively.

		HK		Stanford	
		Volumes	Cases	Volumes	Cases
Training	*G*	1,752	1,153	-	-
	N	1,143	761	-	-
Validation	*G*	574	372	-	-
	N	441	265	-	-
Testing	*G*	600	394	806	750
	N	367	255	425	398
All	*G*	2,926	1,901	-	-
	N	1,951	1,281	-	-

Hong Kong (CUHK) Eye Center and the Hong Kong Eye Hospital, named *HK* dataset. It is the primary dataset that is used to develop and evaluate algorithms. This dataset consists of 4,877 volumetric OCT images of the optic disc imaged by Cirrus HD-OCT (Carl Zeiss Meditec, Inc., Dublin, CA, USA) from 930 subjects. It is worth noting that each investigated subject (one eye or both two eyes involved) might have several follow-ups. Also, during each follow-up, several OCT images may be taken, which eventually results in 3,182 eye visits in total. Specially, we denote the eye-visit result as the *case-level* result. Besides, all the included subjects performed the VF test that was determined by the Humphrey Field Analyzer II (Carl Zeiss Meditec, Inc., Dublin, CA, USA) and a combination of SITA-standard and SITA-fast. All the VF reports included were reviewed both quantitatively and qualitatively. Note that a part of VF measurements for some follow-ups are unavailable in this study, hence there are 974 OCT volume in lack of VF measurement labels.

Two glaucoma specialists, who completed a two-year Glaucoma Fellowship at CUHK and with more than 10-year clinical experience in managing glaucoma, worked individually to label all the OCT images into glaucoma and normal, taking VF test results and other clinical records as reference. A senior glaucoma expert was consulted in case of disagreement. Subsets of 2,895, 1,015 and 967 images are randomly selected for training, validation and testing, respectively. The random sampling is at patient level so as to prevent leakage and biased estimation of the testing performance. According to the accessibility of VF measurements in the training set, we re-configure the training set as follows: (i) *Part*: 1,979 images whose VF measurements exist. (ii) *All*: all 2,895 images.

Moreover, an external validation dataset provided by Byers Eye Institute, Stanford University in the USA, named *Stanford* dataset, is used to assess the generalization capability of the proposed method. The dataset is also imaged by Cirrus HD-OCT (Carl Zeiss Meditec, Inc., Dublin, CA, USA). All VF measurements in *Stanford* dataset are not provided. It is worth noting that no fine-tuning is performed on *Stanford* dataset, and the test results are directly generated by the models trained on *HK* dataset. The statistical details of the two datasets are reported in Table 1, and the numbers of OCT volumes and cases are shown in Table 2.

4.2. Evaluation metrics

The classification performance is measured via three criteria: accuracy, F1 score, and AUC. To obtain the *case-level* prediction, the averaging method is used to aggregate the results of all images during each eye visit to a single one. As for the regression tasks, we utilize mean absolute error (mAE) as the evaluation metric.

4.3. B-scan sampling strategy

Thinning of the retinal nerve fiber layer and cupping of the optic disc are two major structural changes caused by glaucoma. One of the most prominent advantages of the OCT image over the fundus image is its rich information covering the whole retina and the optic disc. However, the encoded information varies in the successive B-scan images. Specifically, in the head and the tail of the OCT volume along the B-scan axis, there is an obvious deficiency of content depicting the optic disc, as illustrated in Fig. 4. On the contrary, in the middle part of OCT volume, images with richer content are more capable of characterizing the structural changes. In order to investigate the importance of B-scans on the classification performance, we train a B-scan-wise network (i.e., *2D-ResNet* described in Section 4.6) using the training set in Table 2, and then perform inference on both the validation set and the testing set (called **E1**). With regard to the inference result on the validation set, we record the probability of each B-scan input within OCT volumes. We calculate the average of every ith (i =1, 2, ..., 200) B-scan's probabilities of all OCT images in the homogeneous group (being normal or glaucoma). In this way, we can obtain the average response of B-scans at different positions to the corresponding ground truth label. As the Fig. 5 shows, the red curve means the average confidence of normal B-scans predicted as normal, and the blue curve denotes the average confidence of glaucoma B-scans predicted as glaucoma. Informatively, the middle part of B-scans in normal and glaucoma OCT volumes have a relatively higher response. Such observation indicates that these B-scans are highly discriminative. Next, we perform another experiment where we only consider 60th to 139th B-scans for training and validation

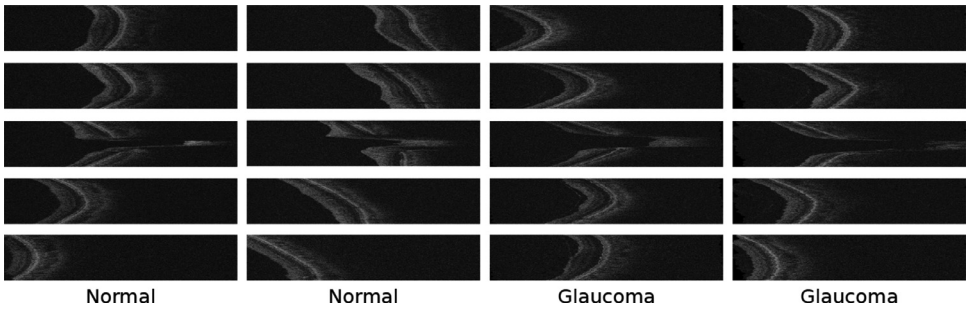

Fig. 4. B-scans at different position within the OCT volume. B-scans in the same column come from the same OCT volume and are the 1st, 50th, 100th, 150th and 200th B-scan respectively from the upper panel to the bottom panel.

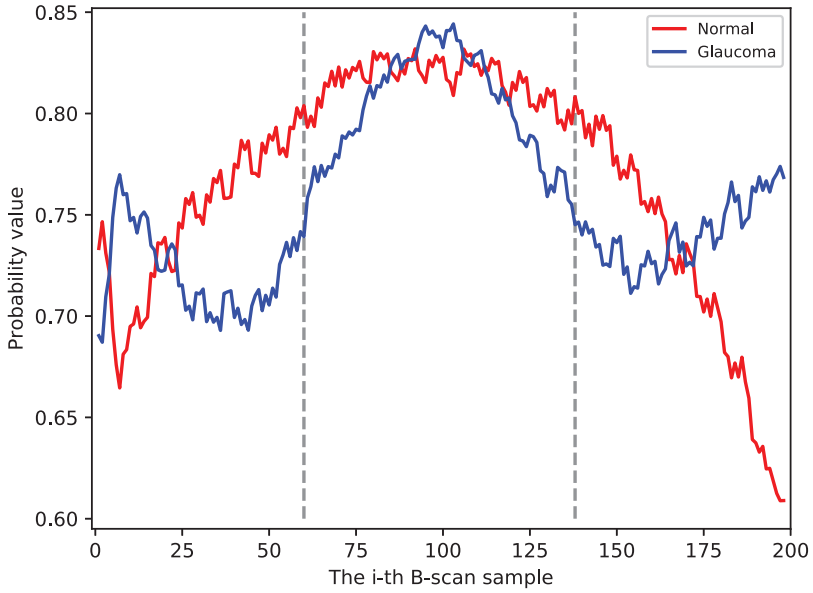

Fig. 5. The average response of the B-scan sample along the B-scan axis to the corresponding ground truth label of the validation set.

Table 3
Results of **E1** and **E2** on the validation and testing set.

		Accuracy		*F*1 score		AUC	
		Volume	Case	Volume	Case	Volume	Case
Val	E1	0.907	0.900	0.911	0.913	0.959	0.951
	E2	0.923	0.920	0.925	0.921	0.975	0.969
Test	E1	0.868	0.869	0.870	0.871	0.953	0.953
	E2	0.908	0.911	0.904	0.909	0.968	0.964

(named **E2**). The results on the validation and testing set of **E1** and **E2** are reported in Table 3. It is obvious that the model **E2**, trained with selected B-scans, can achieve a better result. Therefore, in the following experiments related to 2D CNNs, we only take the 60th to 139th B-scans into account.

4.4. Comparison of different backbone networks

To investigate the impact of different backbone networks on the classification performance, we replace the ResNet18 with several popular classification CNNs, including VGG16 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2016), DenseNet121 (Huang et al., 2017) and MobileNet (Howard et al., 2017). Not only the classification performance (i.e., volume-level AUC) is considered, but also the number of parameters of the network, computation memory and efficiency are included for a fair and comprehensive comparison. Especially, because the minimum input size of DenseNet required in *Keras* package is 221, we enlarge the B-scan images to 221×1132 accordingly, preserving the aspect ratio. For other networks, the input size remains unchanged, i.e., 200×1024. Table 4 summaries the statistics and classification performance of investigated networks, including the number of parameters, input size, random-access memory (RAM) per training image required, time to process an OCT volume, and volume-level AUC score on the validation and testing set. According to the AUC score, it is noticed that most of the networks realize comparable classification performance on *HK* testing set. Although ResNet18 is the second fastest network, a little slower than MobileNet, it has the minimum memory cost during the training phase and also achieves better performance than others. Considering the efficiency and classification precision, ResNet18 is more suitable for this classification task. Hence, it is selected as the backbone for algorithm development in this study.

4.5. Impact of α in weighted loss function

In this section, we conduct a number of experiments on the *HK* validation set to analyze the influence of the weight α in Eq. (7) on the joint training. Note that Euclidean distance is calculated to quantify similarity for the surrogate assignment ahead of the multi-task learning. It can also be replaced with Cosine similarity or Manhattan distance. A series of values of $\alpha \in \{0.25, 0.33, 0.50, 1.0\}$ are explored. The smaller α is, the more penalty is imposed on the binary cross-entropy loss for the classification task.

Table 4
Comparison of different backbone networks.

Networks	Input size	Million Parameter	Megabyte RAM	Million Mul-Add	Second Time cost	AUC Val	AUC Test
VGG16	200 × 1024	33.61	365.103	67.18	3.82	0.962	0.967
ResNet18	200 × 1024	11.19	**227.426**	22.37	2.46	**0.974**	**0.968**
ResNet50	200 × 1024	23.55	717.732	47.05	3.66	0.961	0.965
DenseNet121	221 × 1132	7.04	983.670	13.99	4.61	0.959	0.967
MobileNet	200 × 1024	**3.23**	307.686	**6.44**	**2.37**	0.971	0.963

Table 5
Performance of the proposed method under different weights of α on *HK* validation set.

α	Accuracy		*F*1 score		AUC		VFI	MD	PSD
	Volume	Case	Volume	Case	Volume	Case			
0.25	0.921	0.921	0.926	0.926	**0.977**	0.978	**0.110**	3.953	2.298
0.33	0.926	**0.929**	**0.941**	**0.941**	**0.977**	**0.979**	0.114	**3.941**	**2.047**
0.50	**0.927**	0.928	0.932	0.933	0.973	0.974	0.120	4.088	2.097
1.00	0.926	0.925	0.931	0.930	0.974	0.975	0.122	4.115	2.183

Table 5 reports the results of the proposed method under different values of α. We can observe that when α is set small, such as 0.25 and 0.33, the multi-task learning model is relatively robust, and could achieve a better result. By contrast, as α increases to a large value (i.e., 0.50 and 1.0), there is a considerable performance drop, especially on AUC score, which indicates that the proposed method is sensitive to large values of α. It is also worth noting that our method under $\alpha = 0.33$ has the most stable results and approaches the peak performance, which is better than $\alpha = 0.25$ on the whole. Hence, α is set 0.33 in the following experiments.

4.6. Quantitative evaluation and comparison

At present, quite few works have been proposed for glaucoma OCT image classification. Hence, we implement several baseline methods, including two existing methods as well as two variants of the proposed method, for comparison:

(i). 3D-CNN: the implementation of the approach proposed in Maetschke et al. (2019), which is trained with downsampled 3D volumes. It contains five convolutional layers, a GAP layer, a fully connected layer, and a softmax layer.
(ii). 3D-ResNet: the 3D implementation of ResNet that takes raw volumes as input (Ran et al., 2019). The number of filters in convolutional layers are halved to reduce the computational cost.
(iii). 2D-ResNet: ResNet18 trained with B-scan images sampled from OCT volumes.
(iv). 2D-ResNet-MT: 2D-ResNet with Multi-Task learning network for classification and regression, as shown in Fig. 3. There is no surrogate label assignment. Specifically, if the training sample is lack of VF measurement label, the regression loss is ignored.
(v). 2D-ResNet-SEMT: the proposed SEmi-supervised Multi-Task learning network with the surrogate label assignment.

4.6.1. Comparison of different similarity measures

In this study, three types of similarity measures, i.e., Euclidean distance, Cosine similarity, and Manhattan distance, are respectively applied for surrogate label assignment. The testing results of two datasets are listed in Tables 6 and 7. It is obvious that Euclidean distance and Manhattan distance perform better than Cosine similarity does. Besides, Manhattan distance achieves comparable or superior classification performance to Euclidean distance, and it has the best regression results on *HK* dataset, achieving the lowest mAE scores. Therefore, Manhattan distance is finally utilized in the proposed framework to compare with the state-of-the-art and baseline methods.

4.6.2. Experimental results on HK dataset

The experimental results of compared methods on *HK* dataset are listed in Tables 8 and 9. From the classification result, we can notice that when models were trained with the same set *All*, 2D-ResNet is superior to 3D-ResNet. There are two possible reasons. One is that training the 3D network with the high-dimensional data is extremely difficult. In fact, the validation loss oscillates wildly during the training phase, which thus makes it hard for model selection. The other is that there is a deficiency of 3D pre-trained models available, so training from scratch could readily result in the local optima. By exploring the structure and function relationship through 2D-ResNet-MT, the classification performance is improved, which verifies our hypothesis that the extra information from the regression module is helpful. Noticeably, an improvement is also observed when the training samples increase (i.e., from *Part* to *All*). Inspiringly, the proposed framework achieves the best results among aforementioned methods on all metrics. With surrogate label assignment for VF measurement, the classification module can receive more reliable information from the regression branch.

Compared to the pioneering work 3D-CNN (Maetschke et al., 2019), all of our semi-supervised multi-task learning methods outperform it by a large margin, with 4.3%, 6.0%, and 1.5% performance improvement on accuracy, F1 score and AUC at volume level. By and large, 3D-CNN has two main drawbacks. First, the input volumes are compressed intensively, which may lead to discriminative information loss. Second, its generalization ability is quite limited due to the shallow network structure.

To further evaluate the clinical usability of the proposed method for glaucoma OCT screening, we invited another two glaucoma experts to identify glaucoma based on the printout of the OCT images in the format that ophthalmologists usually read in clinic. They reviewed the printouts individually masked from any other clinical notes to make the decision, either glaucoma or normal, for each testing image. Apparently, the proposed method exceeds expert performance significantly, particularly on AUC.

With a closer observation of the predictions on the *HK* testing set, there are a total of 71 out of 967 OCT volumes misclassified by the proposed method. Following Masumoto et al. (2018), we categorize all these images into three groups according to the glaucoma VF damage measured by Humphrey Field Analyzer II. Explicitly, mean deviation of −6 dB or better is classified as *mild*, from

Table 6
Comparison with different similarity measures on *HK* dataset.

Similarity	Accuracy		*F*1 score		AUC		VFI	MD	PSD
	Volume	Case	Volume	Case	Volume	Case			
Cosine	0.914	0.917	0.923	0.926	0.977	0.979	0.119	3.793	2.261
Euclidean	**0.927**	0.934	**0.930**	**0.937**	**0.978**	**0.980**	0.108	3.740	2.124
Manhattan	**0.927**	**0.935**	0.921	0.929	0.977	0.979	**0.105**	**3.508**	**2.034**

Table 7
Comparison with different similarity measures on *Stanford* dataset.

Similarity	Accuracy		*F*1 score		AUC	
	Volume	Case	Volume	Case	Volume	Case
Cosine	0.839	0.846	0.871	0.876	0.925	0.930
Euclidean	0.858	0.864	**0.891**	**0.895**	0.930	0.932
Manhattan	**0.860**	**0.865**	0.889	0.894	**0.933**	**0.937**

Table 8
Comparison with different methods and expert performance for glaucoma classification on *HK* dataset.

Data	Methods	Accuracy		*F*1 score		AUC	
		Volume	Case	Volume	Case	Volume	Case
Part	2D-ResNet (He et al., 2016)	0.823	0.829	0.835	0.830	0.960	0.955
	2D-ResNet-MT	0.878	0.882	0.890	0.887	0.971	0.968
All	3D-CNN (Maetschke et al., 2019)	0.884	0.889	0.881	0.890	0.962	0.959
	3D-ResNet (Ran et al., 2019)	0.880	0.875	0.878	0.874	0.958	0.956
	2D-ResNet (He et al., 2016)	0.908	0.911	0.904	0.909	0.968	0.964
	2D-ResNet-MT	0.915	0.912	0.923	0.917	0.975	0.971
	2D-ResNet-SEMT	**0.927**	**0.935**	**0.941**	**0.948**	**0.977**	**0.979**
	Expert 1	0.912	0.912	0.917	0.917	0.918	0.918
	Expert 2	0.905	0.905	0.913	0.913	0.914	0.914

Table 9
Comparison with different methods and expert performance for regression tasks on *HK* dataset. The mAE score is calculated on the original unscaled data of VFI, MD and PSD.

Data	Methods	VFI	MD	PSD
Part	2D-ResNet-MT	0.167	4.519	2.566
All	2D-ResNet-MT	0.114	4.117	2.308
	2D-ResNet-SEMT	**0.105**	**3.508**	**2.034**

Table 10
Cohen's kappa coefficient between different methods/experts and ground truth.

Methods	*HK*		*Stanford*	
	Volume	Case	Volume	Case
3D-CNN (Maetschke et al., 2019)	0.756	0.768	0.615	0.628
3D-ResNet (Ran et al., 2019)	0.748	0.744	0.590	0.625
2D-ResNet (He et al., 2016)	0.805	0.818	0.630	0.639
2D-ResNet-MT	0.825	0.824	0.655	0.662
2D-ResNet-SEMT	**0.844**	**0.861**	**0.698**	**0.710**
Expert1	0.817	0.826	–	–
Expert2	0.804	0.807	–	–

−6 to −12 dB as *moderate*, and −12 dB or worse as *severe*. Among all wrong predictions, the majority are mild glaucoma images (39), followed by the normal (21) and moderate glaucoma ones (11). Importantly, no severe glaucoma images are mispredicted by our model. It is challenging to recognize early-stage glaucoma due to very slight structural change at this stage, which can also be easily misdiagnosed by the experienced experts.

In addition to classification results, we also report the mAE score of the regression tasks for VF measurements (i.e., VFI, MD, and PSD) in Table 9. It is noticeable that all of the regression results are significantly improved when the surrogate labels fill the vacancy of the missing VF measurement via semi-supervised learning, which further verifies the efficacy of this component in the proposed framework.

Furthermore, we calculated the Cohen's kappa coefficient between different methods/experts and ground truth labels, which is frequently utilized in healthcare studies to measure the rater reliability. As shown in Table 10, our 2D-ResNet18-SEMT achieved the highest score of 0.844 and 0.861 at volume level and case level on *HK* dataset, accordingly.

4.6.3. Experimental results on Stanford dataset

To further assess the performance of the proposed approach and baseline methods, we conducted identical tests on *Stanford* dataset by using the models trained on *HK* dataset. As VF measurements of *Stanford* dataset are not available, so only classification performance is reported in Table 11.

As expected, the results are consistent with those on *HK* dataset. Admittedly, there is a performance drop on *Stanford* dataset on the whole. This might be mainly attributed to the large ethnic difference existing between the two datasets. The subjects investigated within *HK* dataset primarily include Asian ethnic origin, while the *Stanford* dataset has a more ethnically diverse population, like white Americans, African Americans, etc. As reported in Tham et al. (2014), the risk and subtypes of glaucoma vary among races and countries, so *Stanford* dataset might have subtypes of glaucoma that might be absent in *HK* dataset, and vice versa. The different data distribution might explain why such a gap of testing performance between *HK* dataset and *Stanford* dataset exists.

By the large, our proposed method achieved the best performance among all approaches, with 0.860, 0.889 and 0.933 on accuracy, *F*1 score and AUC at volume level and with 0.865, 0.894 and 0.937 on accuracy, *F*1 score and AUC at case level, which demonstrates the robustness and efficacy of the proposed method.

Table 11
Comparison with different methods on *Stanford* dataset.

Data	Methods	Accuracy		*F*1 score		AUC	
		Volume	Case	Volume	Case	Volume	Case
Part	2D-ResNet (He et al., 2016)	0.764	0.773	0.792	0.780	0.907	0.909
	2D-ResNet-MT	0.802	0.809	0.833	0.839	0.912	0.918
All	3D-CNN Maetschke et al. (2019)	0.825	0.830	0.865	0.869	0.899	0.905
	3D-ResNet (Ran et al., 2019)	0.807	0.822	0.845	0.858	0.900	0.914
	2D-ResNet (He et al., 2016)	0.827	0.831	0.863	0.866	0.901	0.912
	2D-ResNet-MT	0.835	0.838	0.865	0.868	0.926	0.929
	2D-ResNet-SEMT	**0.860**	**0.865**	**0.889**	**0.894**	**0.933**	**0.937**

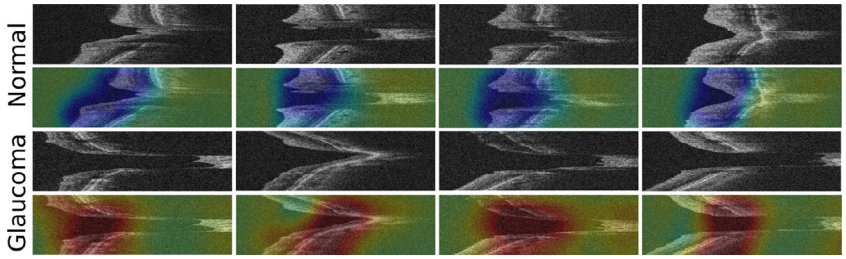

Fig. 6. Visualization of discriminative regions. The first two rows show the pairs of normal B-scan and corresponding CAM. The last two rows show the pairs of glaucomatous ones.

4.7. Qualitative evaluation

Class Activation Maps (CAMs) (Zhou et al., 2016) were computed to visualize the discriminative regions that played an essential role in class prediction. For each pair in Fig. 6, the upper one is the input, and the lower one is the corresponding CAM overlapped with the input. Specifically, the most warm-colored area demonstrates the most discriminative region to identify glaucomatous optic neuropathy and vice versa. Clearly, those regions found by CNN in glaucoma and normal images are quite different. In normal B-scan images, there is barely any response within retinal areas, while such regions contrarily have a high response in glaucoma images. Such observation is exactly corresponding with the clinical diagnosis of glaucoma.

4.8. Implementation details and computation cost

In this section, we present more details about the implementation environment and data augmentation strategies. The proposed method was implemented with Python and *Keras* package, on a workstation equipped with CPU of 3.5 GHz Intel Core i7-5930 and GPU of Nvidia GeForce GTX Titan X. During the whole training phase, online data augmentation techniques were applied, including horizontal flipping, vertical flipping, and 180° rotation. While at the testing stage of the multi-task learning network, only flipping was applied, and the ultimate prediction of B-scan input was the mean of the augmented results. Then we utilized the volume-level inference strategy introduced in Section 3.2.3 to get volume-level results. All 2D CNNs in this paper were initialized by the pre-trained model from ImageNet. And CNNs were then trained by Adam optimizer with the initial learning rate of $1e-05$. We decreased the learning rate by a factor of 0.9 if the training error stagnated. The batch size was set as 25, composed of random 5 B-scan samples from 5 different OCT volumes. During the inference stage, it takes about 2.51 s to process a OCT volume for the proposed method.

5. Discussion

Automatically recognizing structural change from optic disc OCT images plays a vital role in clinical glaucoma screening. Most of the previous approaches address this problem by firstly segmenting retinal layers, the optic disc cup, and rim, and subsequently extracting useful features, like retinal nerve fiber layer thickness and cup-to-disc ratio, to make the final prediction. Undeniably, the discrimination ability of the classification model primarily relies on the preceding segmentation tasks, which are quite complicated and laborious in practice. Hence their discrimination ability is still very limited. Although the recently proposed 3D CNN could directly classify OCT volumes into glaucoma and normal. The OCT images are downsampled extensively, which actually results in a classification performance drop. In the ophthalmology field, the relationship between glaucomatous structural and functional changes is very complicated. Heretofore there are few studies that explore the structure-function relationship based on the raw OCT images and the VF measurement for glaucoma screening.

To overcome the aforementioned shortcomings, in this paper, we propose a robust two-stage framework for glaucoma classification based on optic disc OCT images. To the best of our knowledge, it is the first study that explores the structure-function relationship based on OCT images and VF measurement in a multi-task learning network, to improve the classification performance. The first stage fills the vacancy of missing VF measurement via a semi-supervised learning fashion, by firstly OCT volume representation, then similarity calculation among *homogeneous* groups, and finally surrogate assignment. This setting could guarantee that surrogate labels would not introduce serious noise (e.g., if an OCT image belongs to normal, its surrogate of VF measurement assigned must be in the normal range). In the second stage, a multi-task learning network effectively classifies OCT B-scans into glaucoma and normal by taking advantage of the regression features. Through the fusion of features from the shared feature extraction module and the regression module, the classification performance is improved, which further demonstrates the benefit of the explored structure-function relationship. In addition to a series of preliminary experiments to investigate the effects of B-scan samples, different networks (e.g., VGG16, ResNet50, DenseNet121, and MobileNet), and the hyper-parameter α in the weighted loss function on the classification performance, we also conducted a number of ablation studies to see how the similarity measures (i.e., Cosine similarity, Euclidean distance, and Manhattan distance) influence the semi-supervised multi-task learning.

It is worth mentioning that we built the largest glaucoma OCT image dataset (*HK* dataset) to develop and validate our method. The trained model was additionally tested on an independent dataset (*Stanford* dataset) without any fine-tuning. The proposed approach outperforms all baseline methods considerably, even though there is a slight decrease in performance on *Stanford* dataset. The underlying reason is that there exists a large ethnic difference between the two datasets and thus induces instinct data distribution among them. Comparative analysis between the assessment of expert ophthalmologists and automated methods is clinically meaningful and interesting. From Tables 8 and 10, it is observed that almost all deep learning methods overwhelm specialists by a large margin. In particular, the proposed method achieves the best performance on all criteria, implying the great potential of automated glaucoma screen tool in real-world clinical practice.

Although our method realized fairly good performance in the two datasets, it still has some limitations. First, the framework of our method at the training stage is not end-to-end, where the hard assignment for VF measurement and the multi-tasking training work in a cascaded manner. The quality of the surrogates also exerts an important influence on the subsequent network training. In the future work, we shall incorporate the former into the multi-task network where the soft label is assigned via adaptive feature embedding and similarity calculation. Besides, domain adaptation is a promising tool to narrow the gap of image variance from different datasets. Hence, we will apply this technique to the cross-institution dataset and further improve the generalization ability of the proposed method.

6. Conclusions

In this study, we present a deep learning framework to screen glaucoma based on OCT images of the optic disc. We first use a semi-supervised learning method to address the miss VF measurement label problem in the training set, and then we build a multi-task learning network to explore the relationship between the functional and structural change of glaucoma and classify OCT images into glaucoma and normal in the meanwhile. The structure-function relationship explored has been verified to be beneficial to accuracy improvement. Extensive experiments on two large-scale cross-center datasets manifest the effectiveness of the proposed approach, which outperforms other competitive methods by a large margin. Besides, the comparison with glaucoma specialists provides strong evidence that our proposed framework has promising clinical potential for automated glaucoma screening in the near future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xi Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Hao Chen:** Conceptualization, Methodology, Supervision, Writing - review & editing. **An-Ran Ran:** Resources, Data curation, Visualization, Writing - review & editing. **Luyang Luo:** Software, Investigation, Visualization, Writing - review & editing. **Poemen P. Chan:** Resources, Data curation. **Clement C. Tham:** Funding acquisition. **Robert T. Chang:** Resources, Data curation. **Suria S. Mannil:** Resources, Data curation. **Carol Y. Cheung:** Resources, Data curation, Supervision. **Pheng-Ann Heng:** Writing - review & editing, Project administration, Funding acquisition.

Acknowledgments

The work described in this paper is supported in part by the Key-Area Research and Development Program of Guangdong Province, China under Grant 2020B010165004, grants from the National Natural Science Foundation of China with Project No. U1613219, Research Grants Council – General Research Fund, Hong Kong (Ref: 14102418) and Shenzhen Science and Technology Program under Project No. JCYJ20180507182410327.

References

Almazroa, A., Burman, R., Raahemifar, K., Lakshminarayanan, V., 2015. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. J. Ophthalmol. 2015.

Alsaih, K., Lemaitre, G., Rastgoo, M., Massich, J., Sidibé, D., Meriaudeau, F., 2017. Machine learning techniques for diabetic macular edema (DME) classification on SD-OCT images. Biomed. Eng. Online 16 (1), 68.

Asaoka, R., Murata, H., Hirasawa, K., Fujino, Y., Matsuura, M., Miki, A., Kanamoto, T., Ikeda, Y., Mori, K., Iwase, A., et al., 2019. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. Am. J. Ophthalmol. 198, 136–145.

Asaoka, R., Murata, H., Iwase, A., Araie, M., 2016. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. Ophthalmology 123 (9), 1974–1980.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Ben-Cohen, A., Diamant, I., Klang, E., Amitai, M., Greenspan, H., 2016. Fully convolutional network for liver segmentation and lesions detection. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 77–85.

Bourne, R.R.A., Stevens, G.A., White, R.A., Smith, J.L., Flaxman, S.R., Price, H., Jonas, J.B., Keeffe, J., Leasher, J., Naidoo, K., Pesudovs, K., Resnikoff, S., Taylor, H.R., 2013. Causes of vision loss worldwide, 1990–2010: a systematic analysis. Lancet Glob. Health 1 (6), e339–e349.

Cerentinia, A., Welfera, D., d'Ornellasa, M.C., Haygertb, C.J.P., Dottob, G.N., 2018. Automatic identification of glaucoma using deep learning methods. In: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics, MEDINFO 2017, 245. IOS Press, p. 318.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning (Chapelle, O. et al., Eds.; 2006) [book reviews]. IEEE Trans. Neural Netw. 20 (3), 542.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2018. Voxresnet: deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage 170, 446–455.

Chen, H., Qi, X., Yu, L., Heng, P.-A., 2016. DCAN: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496.

Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., Cheng, J.-Z., 2017. Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in ct images. IEEE Trans. Med. Imaging 36 (3), 802–814.

Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J., 2015. Glaucoma detection based on deep convolutional neural network. In: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 715–718.

Christopher, M., Belghith, A., Bowd, C., Proudfoot, J.A., Goldbaum, M.H., Weinreb, R.N., Girkin, C.A., Liebmann, J.M., Zangwill, L.M., 2018a. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. Sci. Rep. 8 (1), 16685.

Christopher, M., Belghith, A., Weinreb, R.N., Bowd, C., Goldbaum, M.H., Saunders, L.J., Medeiros, F.A., Zangwill, L.M., 2018b. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. Invest. Ophthalmol. Vis. Sci. 59 (7), 2748–2756.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial Feature Learning. arXiv:1605.09782.

Fang, L., Wang, C., Li, S., Rabbani, H., Chen, X., Liu, Z., 2019. Attention to lesion: lesion-aware convolutional neural network for retinal optical coherence tomography image classification. IEEE Trans. Med. Imaging 38 (8), 1959–1970.

Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging 37 (7), 1597–1605.

Fu, H., Xu, D., Lin, S., Wong, D.W.K., Liu, J., 2014. Automatic optic disc detection in OCT slices via low-rank reconstruction. IEEE Trans. Biomed. Eng. 62 (4), 1151–1158.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, M.-L., Chen, H.-Y., 2005. Development and comparison of automated classifiers for glaucoma diagnosis using stratus optical coherence tomography. Invest. Ophthalmol. Vis. Sci. 46 (11), 4121–4129.

Iutaka, N.A., Grochowski, R.A., Kasahara, N., 2017. Correlation between visual field index and other functional and structural measures in glaucoma patients and suspects. J. Ophthal. Vis. Res. 12 (1), 53.

Jonas, J.B., Aung, T., Bourne, R.R., Bron, A.M., Ritch, R., Panda-Jonas, S., 2017. Glaucoma. Lancet 390, 2183–2193.

Kim, H.J., Lee, S.-Y., Park, K.H., Kim, D.M., Jeoung, J.W., 2016. Glaucoma diagnostic ability of layer-by-layer segmented ganglion cell complex by spectral-domain optical coherence tomography. Invest. Ophthalmol. Vis. Sci. 57 (11), 4799–4805.

Kim, S.J., Cho, K.J., Oh, S., 2017. Development of machine learning models for diagnosis of glaucoma. PLoS One 12 (5), e0177726.

Kingma, D. P., Welling, M., 2013. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105.

Kucur, Ş.S., Holló, G., Sznitman, R., 2018. A deep learning approach to automatic detection of early glaucoma from visual fields. PLoS One 13 (11), e0206081.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lee, K., Niemeijer, M., Garvin, M.K., Kwon, Y.H., Sonka, M., Abramoff, M.D., 2009. Segmentation of the optic disc in 3-D OCT scans of the optic nerve head. IEEE Trans. Med. Imaging 29 (1), 159–168.

Lemaître, G., Rastgoo, M., Massich, J., Cheung, C.Y., Wong, T.Y., Lamoureux, E., Milea, D., Mériaudeau, F., Sidibé, D., 2016. Classification of SD-OCT volumes using local binary patterns: experimental validation for DME detection. J. Ophthalmol. 2016.

Li, F., Wang, Z., Qu, G., Song, D., Yuan, Y., Xu, Y., Gao, K., Luo, G., Xiao, Z., Lam, D.S., et al., 2018a. Automatic differentiation of glaucoma visual field from non-glaucoma visual filed using deep convolutional neural network. BMC Med. Imaging 18 (1), 35.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A., 2018b. H-DenseuNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging 37 (12), 2663–2674.

Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., He, M., 2018c. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 125 (8), 1199–1206.

Liu, L., Dou, Q., Chen, H., Olatunji, I.E., Qin, J., Heng, P.-A., 2018. MTMR-Net: multi-task deep learning with margin ranking loss for lung nodule analysis. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 74–82.

Lucy, K.A., Wollstein, G., 2016. Structural and functional evaluations for the early detection of glaucoma. Exp. Rev. Ophthalmol. 11 (5), 367–376.

Luo, L., Chen, H., Wang, X., Dou, Q., Lin, H., Zhou, J., Li, G., Heng, P.-A., 2019. Deep angular embedding and feature correlation attention for breast MRI cancer analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 504–512.

Maetschke, S., Antony, B., Ishikawa, H., Wollstein, G., Schuman, J., Garnavi, R., 2019. A feature agnostic approach for glaucoma detection in OCT volumes. PLoS One 14 (7), e0219126.

Masumoto, H., Tabuchi, H., Nakakura, S., Ishitobi, N., Miki, M., Enno, H., 2018. Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. J. Glaucoma 27 (7), 647–652.

Medeiros, F.A., Jammal, A.A., Thompson, A.C., 2018. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. Ophthalmology 126 (4), 513–521.

Medeiros, F.A., Zangwill, L.M., Alencar, L.M., Bowd, C., Sample, P.A., Susanna, R., Weinreb, R.N., 2009. Detection of glaucoma progression with stratus oct retinal nerve fiber layer, optic nerve head, and macular thickness measurements. Invest. Ophthalmol. Vis. Sci. 50 (12), 5741–5748.

Melekhov, I., Kannala, J., Rahtu, E., 2016. Siamese network features for image matching. In: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 378–383.

Raghavendra, U., Fujita, H., Bhandary, S.V., Gudigar, A., Tan, J.H., Acharya, U.R., 2018. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. Inf. Sci. 441, 41–49.

Ramzan, A., Akram, M.U., Shaukat, A., Khawaja, S.G., Yasin, U.U., Butt, W.H., 2018. Automated glaucoma detection using retinal layers segmentation and optic cup–to-disc ratio in optical coherence tomography images. IET Image Proc. 13 (3), 409–420.

Ran, A.R., Cheung, C.Y., Wang, X., Chen, H., Luo, L.-y., Chan, P.P., Wong, M.O., Chang, R.T., Mannil, S.S., Young, A.L., et al., 2019. Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. Lancet Digit. Health 1 (4), e172–e182.

Rasti, R., Mehridehnavi, A., Rabbani, H., Hajizadeh, F., 2018a. Automatic diagnosis of abnormal macula in retinal optical coherence tomography images using wavelet-based convolutional neural network features and random forests classifier. J. Biomed. Opt. 23 (3), 035005.

Rasti, R., Rabbani, H., Mehridehnavi, A., Hajizadeh, F., 2018b. Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Trans. Med. Imaging 37 (4), 1024–1034.

Shibata, N., Tanito, M., Mitsuhashi, K., Fujino, Y., Matsuura, M., Murata, H., Asaoka, R., 2018. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Sci. Rep. 8 (1), 14665.

Silva, F.R., Vidotti, V.G., Cremasco, F., Dias, M., Gomi, E.S., Costa, V.P., 2013. Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using spectral domain OCT and standard automated perimetry. Arq. Bras. Oftalmol. 76 (3), 170–174.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Song, Y., Tan, E.-L., Jiang, X., Cheng, J.-Z., Ni, D., Chen, S., Lei, B., Wang, T., 2016. Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans. Med. Imaging 36 (1), 288–300.

Song, Y., Zhang, L., Chen, S., Ni, D., Lei, B., Wang, T., 2015. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans. Biomed. Eng. 62 (10), 2421–2433.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tham, Y.-C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.-Y., 2014. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology 121 (11), 2081–2090.

Wang, D., Wang, L., 2019. On OCT image classification via deep learning. IEEE Photon. J. 11 (5), 1–14.

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Huang, Q., Cai, M., Heng, P.-A., 2018. Weakly supervised learning for whole slide lung cancer image classification. Med. Imaging Deep Learn..

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.-A., 2019a. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern..

Wang, X., Chen, H., Luo, L., Ran, A.-r., Chan, P.P., Tham, C.C., Cheung, C.Y., Heng, P.-A., 2019b. Unifying structure analysis and surrogate-driven function regression for glaucoma OCT image screening. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 39–47.

Yaqub, M., 2012. Visual fields interpretation in glaucoma: a focus on static automated perimetry. Commun. Eye Health 25 (79–80), 1.

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