Neurocomputing 359 (2019) 285–297

Contents lists available at ScienceDirect

Neurocomputing

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

A spatial-aware joint optic disc and cup segmentation method

Qing Liu[a], Xiaopeng Hong[b,c], Shuo Li[d], Zailiang Chen[a], Guoying Zhao[c], Beiji Zou[a,e,*]

[a] School of Computer Science and Engineering, Central South University, China
[b] Xian Jiaotong University, China
[c] Centre for Machine Vision and Signal Analysis, University of Oulu, Finland
[d] Department of Medical Imaging, Western University, Canada
[e] Hunan Province Engineering Technology Research Center of Computer Vision and Intelligent Medical Treatment, China

ARTICLE INFO

Article history:
Received 27 August 2018
Revised 10 May 2019
Accepted 13 May 2019
Available online 22 May 2019

Communicated by Pingkun Yan

Keywords:
Joint OD and OC segmentation
Conditional probability
Spatial-aware error function
Glaucoma screening

ABSTRACT

When dealing with the optic disc and cup in the optical nerve head images, their joint segmentation confronts two critical problems. One is that the spatial layout of the vessels in the optic nerve head images is variant. The other is that the landmarks for the optic cup boundaries are spatially sparse and at small spatial scale. To solve these two problems, we propose a spatial-aware joint segmentation method by explicitly considering the spatial locations of the pixels and learning the multi-scale spatially dense features. We formulate the joint segmentation task from a probabilistic perspective, and derive a spatial-aware maximum conditional probability framework and the corresponding error function. Accordingly, we provide an end-to-end solution by designing a spatial-aware neural network. It consists of three modules: the atrous CNN module to extract the spatially dense features, the pyramid filtering module to produce the spatial-aware multi-scale features, and the spatial-aware segmentation module to predict the labels of pixels. We validate the state-of-the-art performances of our spatial-aware segmentation method on two public datasets, i.e., ORIGA and DRISHTI. Based on the segmentation masks, we quantify the cup-to-disk values and apply them to the glaucoma screening. High correlation between the cup-to-disk values and the risks of the glaucoma is validated on the dataset ORIGA.

1. Introduction

Joint segmentation of the optic disc (OD) and optic cup (OC) in the optic nerve head (ONH) images is a fundamental task for the automated glaucoma screening and progression assessment. In an ONH image, the OD (see Fig. 1(a)) is the place that the optic nerve entering the retinal, and exhibits as an ellipse-like bright region. It is also the convergent point that the major vessels converge from both the superior and inferior directions. The OD is divided into two different parts. One is the centre portion called OC (see Fig. 1(b)) and the other is between the OD and OC, called the neural retinal rim (see Fig. 1(b)). The OC is a pit with no nerve fibres. Its boundaries are commonly determined based on the pallor and kinks of the vessels [1]. The pallor is defined as the area with maximum colour contrast inside the OD region. The kinks, also called r-bends, are defined as the bending of the vessels when they traverse the OC boundaries and dip into the optic pit [1] (see Fig. 1(b)). Clinically, the ratio of the vertical OC diameter to the vertical OD diameter, called the vertical cup-to-disc ratio (CDR), is manually estimated by ophthalmologists. It is used as a common measurement for the glaucoma diagnosis and progression assessment [2]. Such a manual assessment is labour intensive and impossible for the large scale population screening. To make the screening automated and assist the progression assessment, the joint segmentation of OD and OC is desired.

One way to automatically segment the OD and/or OC is to design hand-crafted features according to the clinical principles. For example, Salazar-Gonzalez et al. [3] incorporate the appearance and vessel structure priors into a graph-cut formulation to segment the OD. Joshi et al. [4,5], Damon et al. [6] and Wong et al. [7] segment the OC by detecting the vessel kinks. Cheng et al. [8,9] model for the appearance and learn classifiers for OD and OC separately to classify the superpixels. Riding the wave of deep learning technology, deep methods such as lightweight U-Net [10], MNet [11] and FC-DenseNet [12] are proposed and outperform most of those using the hand-crafted features. However, there still exist two challenges: (1) the variance of the spatial layout of the vessels, and (2) the ill-defined OC boundaries. Next we will detail these two challenges and our considerations.

The first challenge is that the superior and inferior parts of the OD and OC in the ONH image are covered by the major vessels while the nasal and temporal parts are often covered by the small

* Corresponding author.
E-mail addresses: hongxiaopeng@ieee.org (X. Hong), xxxyczl@csu.edu.cn (Z. Chen), guoying.zhao@oulu.fi (G. Zhao), bjzou@csu.edu.cn (B. Zou).

https://doi.org/10.1016/j.neucom.2019.05.039
0925-2312/

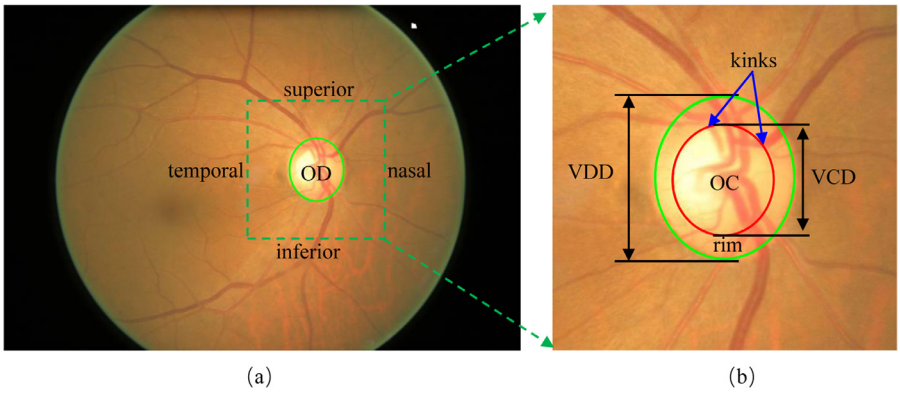


Fig. 1. The structure of the optic disc and optic cup. (a) is the ONH image and (b) is cropped from (a). The optic disc is enclosed by the green circle, and the optic cup is enclosed by the red circle. The region between them is the neural retinal rim. Two examples for the kinks of vessels are pointed by the blue arrows. The vertical cup-to-disc ratio is estimated by the ratio of the vertical diameter of the optic cup (VCD) to the vertical diameter of the optic disc (VDD).

vessels (see Fig. 1(a)). This results in the variance of the subspaces embedding the feature representations of the different image parts. Directly learning one classifier as the segmenter from those features encounters significant difficulty. In fact, the label prediction not only depends on the context features, but also depends on the spatial locations. We therefore explicitly consider the spatial locations of the pixels and propose a spatial-aware segmentation method. In detail, we jointly learn multiple spatial-aware classifiers. Each classifier predicts the labels of pixels belonging to the same image part while different classifiers predict the labels of pixels belonging to different image parts.

The second challenge is that the OC boundaries are subtle. The appearance difference between the OC and the rim is subtle. The most reliable landmarks for the OC boundary delineation are the spatially sparse kinks of vessels at small spatial scale. It requires that the feature extractor is able to not only preserve the local spatial structure details to encode the small scale kinks of the vessels, but also exploit large context to infer the subtle OC boundaries according to their sparse landmarks. In standard CNNs such as [13–16], context is captured by consecutive standard convolutional layers with down sampling. The down sampling operation enlarges the receptive field size. But it also reduces the spatial resolution of the feature maps and easily fails to encode the small scale structures.

To exploit large context and preserve the spatial structures at small scale, this paper proposes to use an atrous CNN [17,18]. On one hand, it enlarges the receptive field size and exploits large context by the atrous filters. On the other hand, it allows us to reduce the times of the down sampling operation compared to the standard CNNs and preserve the spatial structures at small scale.

Fig. 2(a) illustrates the overview of the proposed Spatial-aware neural Network (SAN). It consists of three modules: (1) the atrous CNN module to exploit the large context feature maps incorporating the spatial structure details; (2) the pyramid filtering module to produce multi-scale features for each part; (3) spatial-aware segmentation module to predict the labels of the pixels. Its state-of-the-art segmentation performances are validated on two public datasets, i.e., ORIGA [19] and DRISHTI [20]. Based on the segmentation masks, the CRD value is calculated and applied to the glaucoma perdition. Validation on the ORIGA dataset [19] demonstrates the effectiveness of our method for the CDR estimation and glaucoma prediction.

The contributions of this paper are summarised as follows:

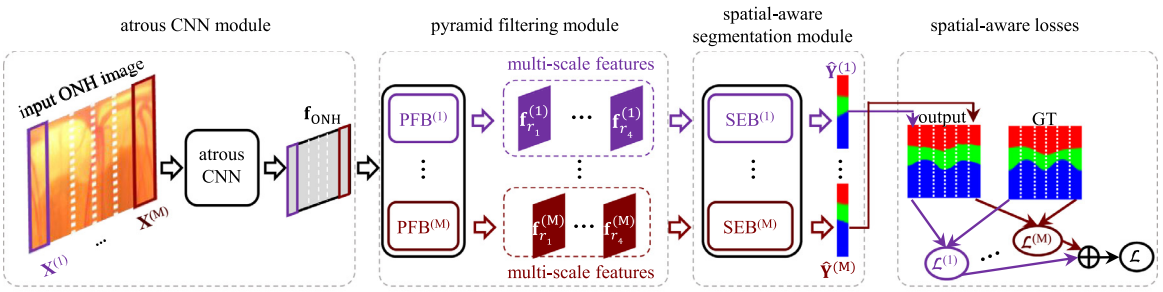
- We highlight the importance of explicitly considering the variance of the spatial layout of the vessels and the spatial sparsity of the kinks of the vessels at small scale for the joint OD and OC segmentation task.
- We propose a spatial-aware joint segmentation method by formulating it as a maximum conditional probability framework from a probabilistic perspective.
- We provide an end-to-end solution by designing a spatial-aware neural network.

The rest of this paper is organised as follows. Section 2 reviews the state-of-the-art methods. Section 3 provides a detailed derivation for the spatial-aware joint segmentation task and description of the proposed SAN. Section 4 reports the experimental results. In Section 5, we provide performances on the CDR estimation based on the segmentation masks and discuss the potential application on the glaucoma screening. The following section concludes this paper.

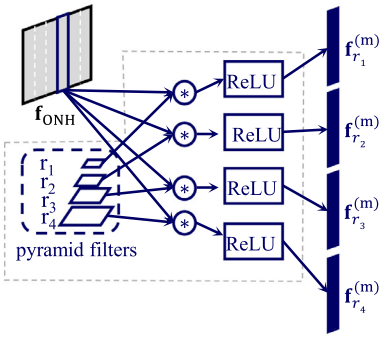
2. Related works

During the past years, tremendous effort has been made for the segmentations of the OD and/or the OC. According to the features that the previous methods exploit for the segmentation tasks, we divide them into two classes: hand-crafted feature based methods and deep feature based methods.

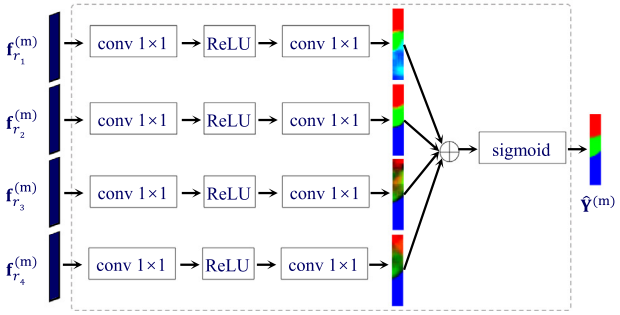
(1) *Hand-crafted feature based segmentation methods.* Conventional medical image analysis relies on hand-crafted features such as texture descriptors [21,22]. The OD and OC have high contrast to their surrounding regions and both of them are circular-like or ellipse-like. Inspired by these two priors, most previous methods focus on how to incorporate them into the segmentation model. For example, Aquino et al. [23] detect the candidate edges and perform Circular Hough Transform to obtain the approximate OD boundaries whereas deformable model is adopted to delineate the boundaries in [24–26]. Liu et al. [27,28] propose a saliency-based OD segmentation method. Xu et al. [29] use low-rank representation to distinguish the OD region and Salazar-Gonzalez et al. [3] use graph-cut to segment the OD. Zheng et al. [30] simultaneously segment the OD and OC with graph-cut. To better integrate the priors, supervised methods such as region-level classifier [31] and superpixel-level classifiers [8,32] are learnt to segment the OD. Similar method is also proposed to segment the OC in [9]. In [33], a shape-appearance model is learnt to segment the OD. [1], Joshi et al. [4], [5]], Damon et al. [6] and Wong et al. [7] focus on the kink detection by analysing the vessels. The key component of these methods is the design of the local hand-crafted features to encode the appearance prior and the local structure of the kinks. They are susceptible to the vessel occlusions, pathological regions, low



(a) The spatial-aware neural network (SAN)



(b) PFB: pyramid filtering block



(c) SEB: spatial-aware segmentation block

Fig. 2. An overview of the proposed spatial-aware neural network (SAN) for the joint OD and OC segmentation. (a) The spatial-aware neural network (SAN). It includes an atrous CNN module, a pyramid filtering module consisting of M parallel pyramid filtering blocks (PFB), and a spatial-aware segmentation module consisting of M parallel spatial-aware segmentation blocks (SEB). The learning of the network is supervised by M spatial-aware error functions. (b) Components of the pyramid filtering block (PFB) associated with part m. (c) Components of the segmentation block (SEB) associated with part m. The detailed configurations for the atrous CNN module, PFB and SEB block are provided in Appendix.

contrast, and low image quality due to the limited representation ability of the local hand-crafted features.

(2) *Deep feature based segmentation methods.* The huge success of the CNNs achieved in many visual tasks directs the efforts into designing deep neural network architectures for the segmentations of the OD and OC. In [34], a deep retinal image understanding network is proposed to jointly segment the OD and vessels. In [35], a semi-supervised method based on the variational auto-encoder [36] is proposed to segment the OC. Cheng et al. [37] first improve the contrast of the retinal image, then train a deep model to segment the OC. To segment both the OD and OC, Sevastopolsky [10] separately train two lightweight U-Net models. Al-Bander et al. [12] formulate the joint segmentation task as a multi-class classification task and propose a U-shaped architecture with dense block. In [11], the task is formulated as a multi-label task and solved by the proposed MNet. In [38], an optic disc segmentation method based on atrous convolution and spatial pyramid pooling is proposed. In [39], a context encoder with dense atrous convolution and residual multi-kernel pooling are proposed to encoder the features for OD segmentation. However, they ignore the spatial structure variance in the ONH image and the spatial sparsity of the small scale kinks of vessels, which limits their segmentation performances.

Our SAN has significant differences with existing deep learning implementations for the joint OD and OC segmentation [10–12]. First, our SAN is "Backbone Style". It extracts spatially dense feature maps with large context as well as local structure information by an atrous CNN module, which allows to perform dense prediction directly. Instead, [10,12] and [11] are "Encoder-Decoder Style". They first design an encoder to extract spatially coarse feature maps incorporating with large context, then design a decoder to recover the spatial information. Second, our SAN considers the variance of the spatial layout of the vessels in the ONH image and processes different parts by different segmenters while [10,12] and [11] treat all the pixels equally and perform segmentation on the whole ONH image with one segmenter.

3. Methodology

In this section, we first detail our formulation for the task of the spatial-aware joint segmentation of OD and OC from a probabilistic perspective, then we introduce our design of SAN according to the formulation.

3.1. Formulation from a probabilistic perspective

We formulate the segmentation as a multi-class dense classification task. The goal is to assign a label to each pixel in the input image. Since the ONH image consists of OC, rim and background, among which the OC and rim make up of the OD, we let the label space be {*OC*, *rim*, *background*}. Formally, given the training data set $(\mathcal{X}, \mathcal{Y}) = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{N}$, where $\mathbf{X}_i = \{x_{i,j}\}_{j=1}^{|\mathbf{X}_i|}$ indicates the input image represented by the polar coordinate system with $|\mathbf{X}_i|$ pixels, $\mathbf{Y}_i = \{y_{i,j}\}_{j=1}^{|\mathbf{X}_i|}$ denotes the corresponding ground truth map for image $\mathbf{X}_i$ and $y_{i,j} \in \{oc, rim, background\}$, assuming that the pixels are independent and identity distributed, the joint segmentation is formulated to optimize the following likelihood function:

$$\mathbf{\Theta}^{\star} = \arg\max_{\mathbf{\Theta}} P(\mathcal{Y}|\mathcal{X}, \mathbf{\Theta}) \,, \tag{1}$$

where $\mathbf{\Theta}$ represents the model parameters and $P(\mathcal{Y}|\mathcal{X}, \mathbf{\Theta})$ is the likelihood function:

$$P(\mathcal{Y}|\mathcal{X}, \mathbf{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{|\mathbf{X}_i|} P(y_{i,j}|x_{i,j}, \mathbf{\Theta}) \ . \tag{2}$$

In the ONH image, the differences between OC and rim are subtle. Large context is necessary to infer the label of pixels according to the sparse landmarks of the OC boundaries. Thus, to introduce context information, we let the $y_{i,j}$ not only depends on $x_{i,j}$, but also on its neighbours. Let $\mathcal{N}(x_{i,j})$ be the set of pixels including $x_{i,j}$ and its neighbours and assume that each $\mathcal{N}(x_{i,j})$ is sampled from the ONH images independently. The conditional probability over the data set can be expressed as:

$$P(\mathcal{Y}|\mathcal{X}, \mathbf{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{|\mathbf{X}_i|} P(y_{i,j}|\mathcal{N}(x_{i,j}), \mathbf{\Theta}) \ . \tag{3}$$

To handle the variance of the spatial layout of the vessels in the ONH image, we assume that the label of the pixel not only depends on its context information, but also on its spatial location. Following this assumption, we introduce a matrix variable $\mathbf{L}$ representing the spatial locations of the pixels in images to the conditional probability function:

$$P(\mathcal{Y}|\mathcal{X}, \mathbf{L}, \mathbf{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{|\mathbf{X}_i|} P(y_{i,j}|\mathcal{N}(x_{i,j}), l_{i,j}, \mathbf{\Theta}) \ , \tag{4}$$

where $l_{i,j}$ is the spatial location of the jth pixel in $\mathbf{X}_i$.

Intuitively, we partition the ONH image into M parts uniformly from left to right such that $\mathbf{X}_i = \{\mathbf{X}_i^{(1)} \cup \ldots \cup \mathbf{X}_i^{(M)}\}$ and let $\mathbf{\Theta}$ consist of three parts $\mathbf{\Theta} = \{\theta_{fea}, \theta_{ms}, \theta_{cls}\}$. θ_{fea} is parameters shared by all the parts in the ONH images and associates to the common feature extraction. $\theta_{ms} = \{\theta_{ms}^{(m)}\}_{m=1}^{M}$ and $\theta_{cls} = \{\theta_{cls}^{(m)}\}_{m=1}^{M}$ are spatial-aware parameters. $\theta_{ms}^{(m)}$ is specific to the mth part in the ONH images and associates to the spatial-aware feature extraction. $\theta_{cls}^{(m)}$ associates to the classification of the pixels in the mth part in the ONH images. Thus the conditional probability function can be written as:

$$P(\mathcal{Y}|\mathcal{X}, \mathbf{L}, \mathbf{\Theta}) = \sum_{m=1}^{M} \prod_{i=1}^{N} \prod_{j=1}^{|\mathbf{X}_i|} P(y_{i,j}|\mathcal{N}(x_{i,j}), \theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)}) \cdot \mathbf{1}\{l_{i,j} == m\}, \tag{5}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function.

In the joint OD and OC segmentation, the classes {*oc*, *rim*, *background*} are mutually exclusive. We assume that the class labels are independent given the pixels and their neighbours. Then it can be proved that the optimisation of maximising Eq. (5) is equivalent to minimising the following spatial-aware error function:

$$\mathcal{L}(\mathbf{\Theta}) = \sum_{m=1}^{M} \mathcal{L}^{(m)}(\theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)}) \ , \tag{6}$$

where $\mathcal{L}^{(m)}(\theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)})$ is the error function for the mth part in the ONH images:

$$\begin{aligned} &\mathcal{L}^{(m)}(\theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)}) \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{|\mathbf{X}_i^{(m)}|} \sum_{t} \mathbf{1}\{y_{i,j} == t\} \\ &\quad \cdot \log P(y_{i,j} = t|\mathcal{N}(x_{i,j}), x_{i,j} \in \mathbf{X}_i^{(m)}, \theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)}), \end{aligned} \tag{7}$$

where t belongs to $\{oc, rim, background\}$, $\mathbf{1}\{\cdot\}$ is the indicator function, and $P(y_{i,j} = t|\mathcal{N}(x_{i,j}), x_{i,j} \in \mathbf{X}_i^{(m)}, \theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)})$ is the conditional probability that measures how likely the pixel $x_{i,j}$ belongs to class t.

Overall, the objective of our spatial-aware segmentation method is to find one common feature extractor $\Phi(\cdot, \theta_{fea})$ for the whole ONH image, and M spatial-aware feature generators $\{\Upsilon^{(m)}(\cdot, \theta_{ms}^{(m)})\}_{m=1}^{M}$ and M spatial-aware classifiers $\{\Psi^{(m)}(\cdot, \theta_{cls}^{(m)})\}_{m=1}^{M}$ such that the loss function Eq. (6) is minimised. To this end, we design a spatial-aware neural network and we next will detail it.

3.2. Network architectures

To minimise the loss function Eq. (6) in an end-to-end way, we design a spatial-aware neural network (SAN). Its architecture is illustrated in Fig. 2. It consists of three modules: (1) the atrous CNN module, used as the common feature extractor $\Phi(\cdot, \theta_{fea})$ for the whole ONH image; (2) the pyramid filtering module containing M Pyramid Filtering Blocks (PFBs), used as the spatial-aware feature generators $\{\Upsilon^{(m)}(\cdot, \theta_{ms}^{(m)})\}_{m=1}^{M}$ to extract multi-scale features; (3) the spatial-aware segmentation module including M spatial-aware SEgmentation Blocks (SEBs), used as the spatial-aware classifiers $\{\Psi^{(m)}(\cdot, \theta_{cls}^{(m)})\}_{m=1}^{M}$. Since polar transformation is able to balance the class distributions [11] and increase the diversity of the training data [40], we feed the SAN with the ONH images represented by polar coordinate. Next we will detail each module in the proposed SAN.

3.2.1. Atrous CNN module

To simultaneously exploit the large context and preserve the local spatial structures, we use the atrous CNN module [17] for feature extraction. The atrous CNN module [17] is modified from VGG16 [14]. To reduce the loss of the spatial resolution of the feature maps and preserve the details of the spatial structures at small scale, the atrous CNN module removes the down sampling operators in the last two stages in VGG16 [14] and replaces the standard convolutional layers in the last stage in VGG16 [14] by atrous convolutional layers. The parameters in the atrous CNN module to be optimised constitute θ_{fea} in Eq. (6). They are shared by the whole ONH image. The detailed configuration for the atrous CNN module is provided in Fig. 9 in Appendix.

Compared to existing deep OD and OC segmentation models MNet [11] and lightweight U-Net [10] which are also modified from the VGG16 [14], using the atrous CNN module [17] as the feature extractor has following two advantages. First, it produces spatially denser feature maps. The atrous CNN module only performs down sampling operation three times and down samples the input by a factor of eight while MNet [11] and lightweight U-Net [10] perform down sampling operation four times and down samples the input by a factor of 16. This implies that the atrous CNN module preserves more spatial structure details than MNet [11] and the lightweight U-Net [10]. Second, the atrous CNN module has a larger receptive field size than the MNet [11] and the lightweight U-Net [10]. This indicates that the atrous CNN module exploits larger context than the MNet [11] and the lightweight U-Net [10].

3.2.2. Pyramid filtering module

In the ONH image, the scale of the rim and OC varies across individuals. To further enhance the scale invariant of the features, we design the pyramid filtering module as the spatial-aware feature generators $\{\Upsilon^{(m)}(\cdot, \theta_{ms}^{(m)})\}_{m=1}^{M}$ to produce multi-scale features.

The pyramid filtering module consists of M parallel pyramid filtering blocks (PFBs). Each PFB corresponds to one spatial-aware generator $\Upsilon^{(m)}(\cdot, \theta_{ms}^{(m)})$. Fig. 2(b) illustrates its components. Formally, let $\mathbf{f}_{ONH} \in \mathcal{R}^{H \times W \times C}$ be the feature maps that the atrous CNN module outputs, where $H \times W$ is the spatial resolution of the feature maps and C is the channel number, the pyramid filtering

module produces multi-scale feature maps by learning M pyramids of filters $\theta_{ms} = \{\theta_{ms}^{(m)}\}_{m=1}^{M}$ in parallel via the M PFBs. Particularly, suppose that the mth PFB consists of a spatial pyramid of filters at R scales $\theta_{ms}^{(m)} = \{\mathbf{w}_1^{(m)}, \mathbf{w}_2^{(m)}, ..., \mathbf{w}_R^{(m)}\}$, the outputs of the PFB are the features at R scales $\{\mathbf{f}_1^{(m)}, \ldots, \mathbf{f}_R^{(m)}\}$, where $\mathbf{f}_r^{(m)} \in \mathcal{R}^{H\times W_m \times C_{out}}$ are the features with C_{out} channels at scale r:

$$\mathbf{f}_r^{(m)}[i,j] = ReLU\left((\mathbf{f}_{ONH}^{(m)} *_{(r\cdot\alpha_0)} \mathbf{w}_r^{(m)})[i,j]\right) . \tag{8}$$

In Eq. (8), $\mathbf{f}_{ONH}^{(m)} \in \mathcal{R}^{H\times W_m \times C}$ is the mth part in $\mathbf{f}_{ONH}$. It satisfies $\mathbf{f}_{ONH} = \mathbf{f}_{ONH}^{(1)} \cup \ldots \cup \mathbf{f}_{ONH}^{(M)}$ and $\mathbf{f}_{ONH}^{(m_1)} \cap \mathbf{f}_{ONH}^{(m_2)} = \emptyset$ if $m_1 \neq m_2$. α_0 is the smallest scale range, and $*_{(r\cdot\alpha_0)}$ means the atrous convolution operator with atrous rate $r\cdot\alpha_0$. *ReLU* is the rectified linear unit [41] to increase the nonlinearity and sparsity of the network. The detailed configuration for one PFB is provided in Fig. 10 in Appendix.

3.2.3. Spatial-aware segmentation module and spatial-aware supervision

After the spatial-aware feature maps are obtained, we design the spatial-aware segmentation module associated with parameters $\theta_{cls} = \{\theta_{cls}^{(m)}\}_{m=1}^{M}$ as the spatial-aware classifiers $\{\Psi^{(m)}(\cdot, \theta_{cls}^{(m)})\}_{m=1}^{M}$ to predict the labels of the pixels. It consists of M spatial-aware segmentation blocks (SEBs). Fig. 2(c) shows the components of one SEB block.

Formally, suppose the multi-scale feature maps from the mth PFB be $\{\mathbf{f}_r^{(m)}\}_{r=1}^{R}$ where $\mathbf{f}_r^{(m)} \in R^{H\times W_m \times C_{out}}$ denotes the feature maps of the mth part at scale r, for each part $\mathbf{X}^{(m)}$, the corresponding SEB outputs the probability maps by:

$$P(\mathbf{Y}^{(m)} | \{\mathbf{f}_r^{(m)}\}_{r=1}^{R}, \theta_{cls}^{(m)}) = g\left(\sum_{r=1}^{R} g_2(ReLU(g_1(\mathbf{f}_r^{(m)}, \varpi_r^{(m)})), \omega_r^{(m)}\right), \tag{9}$$

where g_1 and g_2 are linear functions with parameters $\varpi_r^{(m)}$ and $\omega_r^{(m)}$ respectively, $\theta_{cls}^{(m)} = \{\varpi_r^{(m)}, \omega_r^{(m)}\}_{r=1}^{R}$, and g is the softmax function. g_1 aggregates the input features first. Then ReLU is used to regularise the aggregated features. g_2 maps the features of each pixel to the label activations. Thereafter g maps the label activations of each pixel to a probability vector. The detailed configuration of the SEB is provided in Fig. 10 in Appendix.

3.2.4. Learning

Given the training dataset for the segmentation of the OD and OC, we adopt stochastic gradient descent to minimise the error function Eq. (6) and obtain the optimal parameters:

$$(\theta_{fea}, \theta_{ms}, \theta_{cls})^* = \arg\min \mathcal{L}(\theta_{fea}, \theta_{ms}, \theta_{cls}) , \tag{10}$$

where $\theta_{ms} = \{\theta_{ms}^{(m)}\}_{m=1}^{M}$ and $\theta_{cls} = \{\theta_{cls}^{(m)}\}_{m=1}^{M}$.

In the testing phase, given an ONH image $\mathbf{X}$ represented by the polar coordinate, the label prediction map $\hat{\mathbf{Y}}^{(m)}$ for a specific part $\mathbf{X}^{(m)}$ is given by:

$$\hat{\mathbf{Y}}^{(m)} = \max_t P(y_j = t | \mathcal{N}(x_j), x_j \in \mathbf{X}^{(m)}, (\theta_{fea}, \theta_{ms}^{(m)}, \theta_{cls}^{(m)})^*) . \tag{11}$$

By concatenating all the spatial-aware label prediction maps, the whole label prediction map for the input $\mathbf{X}$ is obtained:

$$\hat{\mathbf{Y}} = \hat{\mathbf{Y}}^{(1)} \cup \ldots \cup \hat{\mathbf{Y}}^{(M)}. \tag{12}$$

Thereafter, we perform the inverse radial transformation on $\hat{\mathbf{Y}}$, and obtain the segmentation represented by the Cartesian coordinate. As the same as that in the previous methods [9,11,42], an ellipse fitting is performed on both the OC mask and the OD mask to yield the final segmentation masks.

4. Experimental results

The proposed spatial-aware joint segmentation method is first evaluated and compared with the state-of-the-art segmentation methods on the public dataset ORIGA [19] from two aspects: the overlapping error at region level and the average boundary localisation error at boundary level. It then is evaluated and compared on the dataset DRISHT [20] in terms of the dice coefficient at region level and the average boundary localisation error at boundary level.

4.1. Datasets and experimental setup

Datasets. We validate our SAN on two public datasets: ORIGA [19] and DRISHTI [20]. ORIGA [19] contains 650 fundus images in which 168 images are from glaucomatous eyes and 482 images are from normal eyes. ORIGA [19] provides the boundaries manually delineated by experts, CDR value and the label indicating glaucoma or not for each fundus image. It is divided into 325 training images including 73 glaucoma cases and 325 testing images including 95 glaucoma cases. DRISHTI [20] contains 101 colour retinal fundus images consisting 50 training images and 51 testing images. Each image provides four manual segmentation masks for the OD and OC respectively, which are delineated by four experts with 3, 5, 9, 20 years of clinical experience. The ground truth masks are generated if the regions get the support from at least three experts.

Experimental setup. Our SAN is built on the top of the implementation of Deeplab V2 [17] within the Caffe framework [43]. We initialise the weights θ_{fea} in the atrous CNN module by the Deeplab V2 [17], and initialise the weights θ_{ms} in the pyramid filtering module and the weights θ_{cls} in the spatial-aware segmentation module with Gaussian distribution with zero mean and standard deviation 0.01. We adopt the stochastic gradient descent to optimise the network parameters. The learning rate is set to 0.0001 for θ_{fea} and 0.002 for θ_{ms} and θ_{cls}. The maximum number of training iterations is 18,000. The number of parts M is set to six. The smallest feature scale α_0 is set to six and the number of scales R is set to four. Our work focuses on the segmentation task, thus we simply suppose that the OD centroids are given. The input of proposed SAN is a 401×401 ONH image represented by the polar coordinate which is obtained by performing a radial transformation on an 803×803 ONH image centred on the OD centroid. In the training stage, we augment the training dataset by performing up-sampling ($1.05\times$ original image size), down-sampling ($0.95\times$ original image size) and horizontal flipping on the colour retinal fundus images in the Cartesian coordinate system. Then for each image, nine ONH images size of 803×803 are randomly cropped whose centroids are near the OD centroid, and transformed into the polar coordinate system with the size of 401×401. With this augmentation, the size of training set increases by a factor of 54.

4.2. Evaluation metrics

Region level metrics. The overlapping error $E(\hat{Y}, Y)$ is a widely used metric in previous methods such as [11,44]. It measures the ratio of the number of the wrongly detected pixels and missing detection pixels to the number of pixels in the union set of the segmentation mask $\hat{Y}$ and the ground truth Y. It is defined as:

$$E(\hat{Y}, Y) = 1 - \frac{Area(\hat{Y} \cap Y)}{Area(\hat{Y} \cup Y)} , \tag{13}$$

where $Area(\cdot)$ accounts for the number of the non-zero elements. Additionally, as same as previous method [45], the Dice similarity coefficient $Dice(\hat{Y}, Y)$ is also used. It measures the extent of overlapping between the segmentation mask $\hat{Y}$ and the ground truth Y,

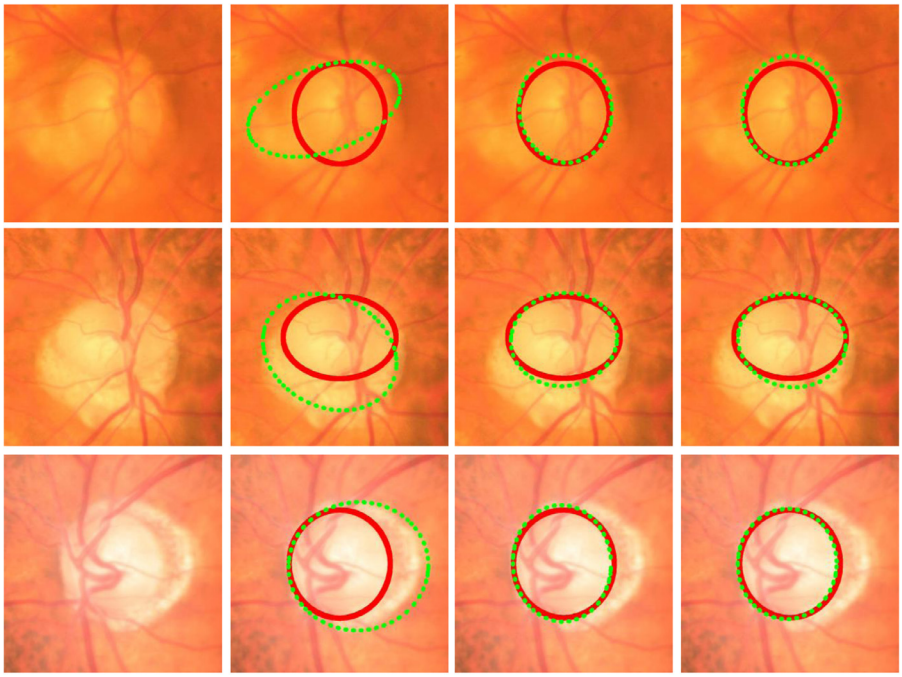


Fig. 3. Examples from ORIGA [19] for OD segmentation by (b) MNet [11], (c) DeepLab V2 [17] and (d) our SAN. The solid elliptic contours in red are the OD boundaries manually delineated by experts and the dashed elliptic contours in green are the OD boundaries of the segmented OD masks by segmentation methods.

and is defined as:

$$Dice(\hat{Y}, Y) = \frac{2 \cdot Area(\hat{Y} \cap Y)}{Area(\hat{Y}) + Area(Y)}, \tag{14}$$

Boundary level metric. Similar to [45], the average boundary localisation error (BLE) is used to evaluate the segmentation performance at a boundary level. It measures the average absolute distance between the boundary $\hat{B}$ of the segmentation mask and the boundary B of the ground truth segmentation. It is defined as:

$$BLE(\hat{B}, B) = \frac{1}{n} \sum_{\theta \in \{\theta_1, \ldots, \theta_n\}} |d(\hat{B}, \theta) - d(B, \theta)|, \tag{15}$$

where $d(\hat{B}, \theta)$ computes the Euclidean distance of the boundary point on $\hat{B}$ in the direction θ to the centroid of the ONH image, and $\{\theta_1, \ldots, \theta_n\}$ is the set of the uniformly sampled directions. As same as the setting used in [45], n is set to 24.

4.3. Results on ORIGA

Segmentation evaluation. For quantitative comparison, we compare our SAN with 11 methods, i.e., R-bend [5], ASM [46], Superpixel [9], LRR [29], lightweight U-Net [10], MNet [11], FC-DenseNet [12], DeepLab V2 [17], JointRCNN [47], DeepLab V3 [48] and PSPNet [49]. The first four methods use hand-crafted features while the last seven are deep learning methods. Among deep learning methods, lightweight U-Net [10], MNet [11], FC-DenseNet [12] and JointRCNN [47], are designed for the OD and OC segmentation. DeepLab V2 [17], DeepLab V3 [48] and PSPNet [49] are originally designed for natural scene image segmentation and fine-tuned to make them be more sophisticated on the joint OD and OC segmentation task. Lightweight U-Net [10], MNet [11], DeepLab V2 [17], JointRCNN [47] and ours adopt VGG16 [14] or lightweight VGG16 [10] as backbone. FC-DenseNet [12] adopts DenseNet [50] as backbone while DeepLab V3 [48] and PSPNet [49] adopt ResNet101 [15] as backbone. The results of R-bend [5], ASM [46], Superpixel [9], LRR [29] and lightweight U-Net [10] are from [11]. The results of MNet [11] are provided by the authors. The results of JointRCNN [47] are directly from the original paper. As for DeepLab V2 [17], we use the original implementation by the authors. As for PSPNet [49] and DeepLab V3 [48], we use the implementation provided by [51]. Compared to the original implementations of PSPNet [49] and DeepLab V3 [48], the implementation by Huang et al. [51] integrates the synchronous Batch Normalization (synBN) and achieves better performances than the original implementations. The hyper-parameters for PSPNet [49] and DeepLab V3 [48] are set followed by PSPNet [49] except for the batch size. In detail, SGD is adopted as the optimizer. The base learning rate is set to 0.01 and the poly learning rate policy is adopted. The max iteration is set to 40000. Momentum and weight decay are set to 0.9 and 0.0001 respectively. Different from PSPNet [49] in which the batch size is set to 16, the batch size in our implementation is set to six due to limitation of GPUs. Performances of those methods are reported in Table 1.

Obviously, as shown in Table 1, the deep learning methods except for lightweight U-Net [10] achieve better performances than those using hand-crafted features. Comparing with lightweight U-Net [10], MNet [11], DeepLab V2 [17], JointRCNN [47] and FC-DenseNet [12] taking VGG16 [14], lightweight VGG16 [10] or DenseNet [50] as backbone, our SAN achieves lowest overlapping errors on disc, cup and rim segmentation. Our SAN also achieves lower *BLE*, which means that the detected boundaries of OD and OC are closer to the manually delineated boundaries. Comparing with deep methods taking the ResNet101 [15] as backbone, our SAN achieves comparable performances on disc, cup and rim segmentation to DeepLab V3 [48], but higher overlapping errors than PSPNet [49] on the disc segmentation by 0.2%, cup segmentation by 0.4% and rim segmentation by 1%. In terms of *BLE*, PSPNet [49], DeepLab V3 [48] and ours achieve comparable performances. However, as we claimed before, the PSPNet [49] and DeepLab V3 [48] adopt the ResNet101 [15] as backbone and both of them use synchronous Batch Normalization (synBN) while our proposed SAN

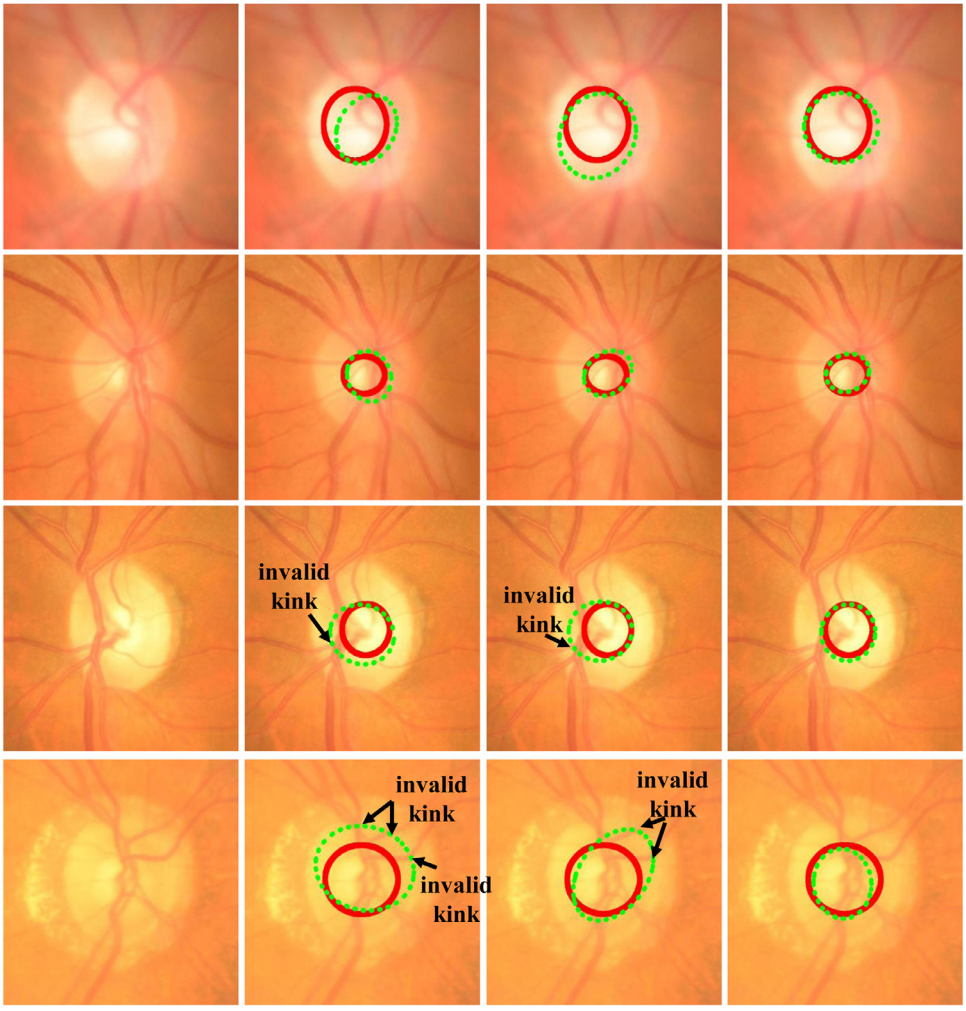


Fig. 4. Examples from ORIGA [19] for OC segmentation by (b) MNet [11], (c) DeepLab V2 [17] and (d) our SAN. The solid elliptic contours in red are the OC boundaries manually delineated by experts and the dashed elliptic contours in green are the OC boundaries of the segmented OC masks by segmentation methods.

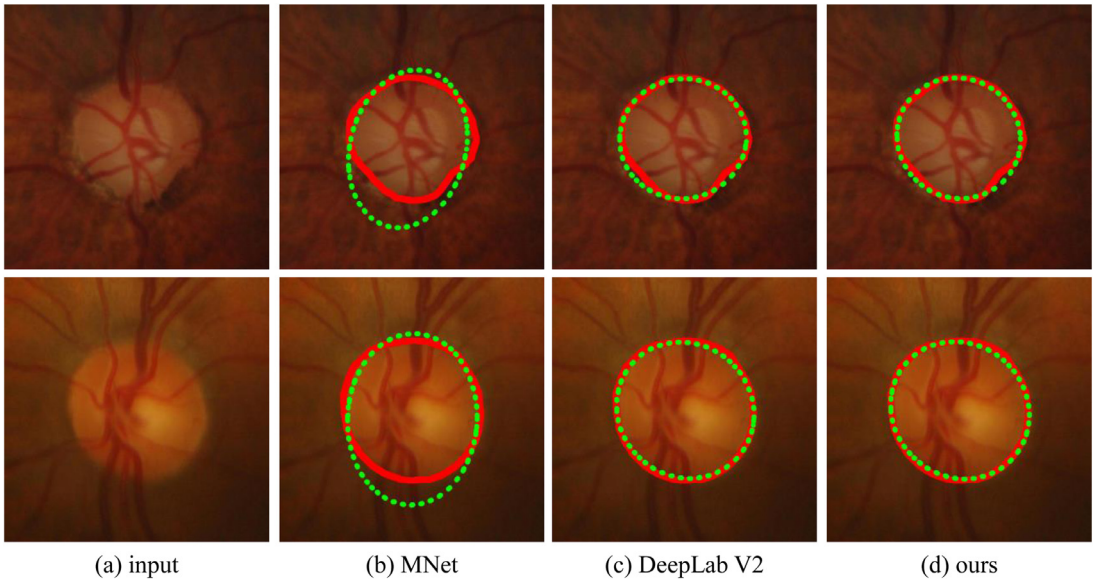


Fig. 5. Examples from DRISHTI [20] for OD segmentation by MNet [11], DeepLab V2 [17] and our SAN. The solid elliptic contours in red are OD boundaries manually delineated by experts and the dashed elliptic contours in green are OD boundaries of the segmented OD masks by segmentation methods.

Table 1
Performance comparisons of the different methods on ORIGA [19].

Methods		E_{disc}	BLE_{disc}/std	E_{cup}	BLE_{cup}/std	E_{rim}
Hand-crafted	R-bend [5]	0.129	–	0.395	–	–
	ASM [46]	0.148	–	0.313	–	–
	Superpixel [9]	0.102	–	0.264	–	0.299
	LRR [29]	–	–	0.244	–	–
Deep learning	lightweight U-Net [10]	0.115	–	0.287	–	0.303
	MNet [11]	0.071	6.88/6.98	0.230	14.64/10.37	0.233
	DeepLab V2 [17]	0.065	6.11/3.02	0.221	14.26/8.93	0.235
	JointRCNN [47]	0.063	–	0.209	–	–
	FC-DenseNet [12]	0.067	–	0.231	–	–
	DeepLab V3 [48]	0.058	5.38/3.44	0.208	13.27/9.01	0.209
	PSPNet [49]	0.057	5.25/3.15	0.204	12.92/8.80	0.205
	Ours	**0.059**	**5.59/2.91**	**0.208**	**13.40/8.93**	**0.215**

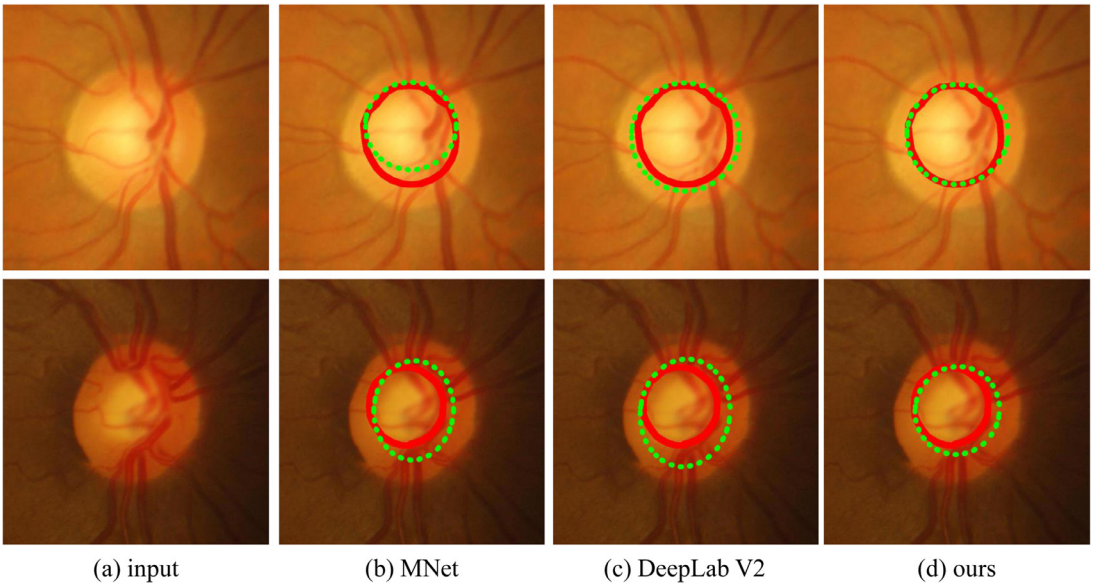


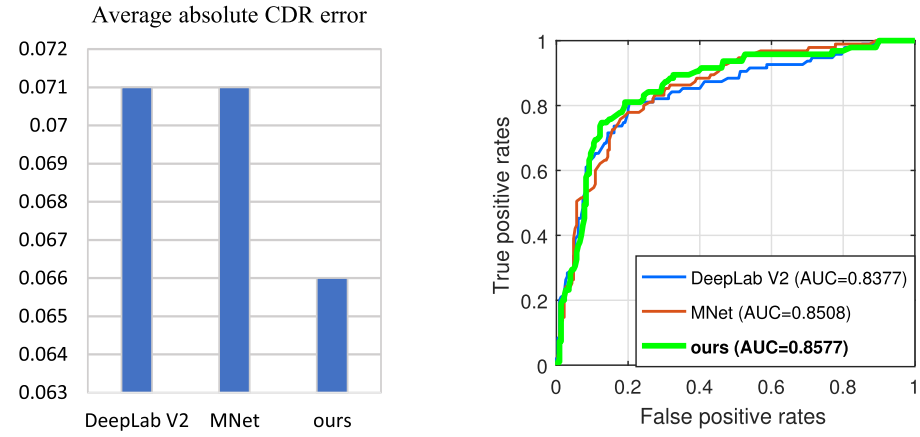
Fig. 6. Examples from DRISHTI [20] for OC segmentation by MNet [11], DeepLab V2 [17] and our SAN. The solid elliptic contours in red are the OC boundaries manually delineated by experts and the dashed elliptic contours in green are the OC boundaries of the segmented OC masks by segmentation methods.

adopts VGG16 [14] as backbone and does not use batch normalization. Both the strong backbone and synBN contribute to better performances than that using the VGG16 without batch normalization. However, strong backbone and batch normalization/synBN need to estimate more parameters in these methods. As a result, the models sizes of them are much larger than ours.

For qualitative comparison, we compare our method with two deep learning methods, i.e. MNet [11] and DeepLab V2 [17]. The former takes the lightweight VGG16 [10] as backbone and the later takes the VGG16 [14] as backbone. Three examples of the OD segmentation results are shown in Fig. 3, in which the solid elliptic contours in red are the OD boundaries manually delineated by experts and the dashed elliptic contours in green are the boundaries of the OD segmentation mask by the segmentation methods. The OD boundaries of these three example images are seriously blurred due to the surrounding confounding bright structures. Those bright structures seriously confuse the MNet [11] and are wrongly classified as OD pixels. On the contrary, both our SAN and DeepLab V2 [17] are robust to those confounding bright structures, and able to distinguish between the OD pixels and the confounding bright pixels correctly. The possible reason is that both DeepLab V2 [17] and our SAN exploit larger context information than MNet [11], which helps to correctly infer the labels of those confounding pixels.

Fig. 4 shows four examples of the OC segmentation results by MNet [11], DeepLab V2 [17] and the proposed SAN. In these ONH images, the boundaries of the OC regions are very ambiguous and difficult to be manually delineated by humans without professional experiences. Nevertheless, the proposed SAN still works well. As is shown in the first two examples in Fig. 4, no matter the ONH image is heavily blurring or the scale of the OC is small, our SAN works much better than the MNet [11] and DeepLab V2 [17]. Furthermore, our SAN performs better when the ONH images contains confusing invalid kinks than the MNet [11] and DeepLab V2 [17], as is shown in Fig. 4(c) and (d). Those invalid kinks exhibit similar appearances and structures to the valid kinks at different spatial locations. To distinguish between those invalid kinks and valid kinks at different locations by one segmenter like MNet [11] and DeepLab V2 [17] is difficult. Instead, our SAN considers the spatial layout of the vessels and learns specific segmenters for each specific part in the ONH images. Thus it is able to distinguish the valid kinks from the invalid kinks.

Ablation study of the SAN. We look into the effect of enabling and disabling different module of our proposed SAN on ORIGA [19]. In Table 2, the performances of different settings are reported. In Table 2, the baseline method consists of the atrous CNN module and standard dense classification. We can see that polar transformation on the input images reduces the overlapping errors of disc, cup and rim segmentation from 8.0%, 24.4%, 27.3% to 6.7%, 22.2%, 23.2% respectively. Simply integrating the pyramid feature module (PFM) with the baseline model changes little on the



(a) Average CDR errors (b) ROC curves

Fig. 7. Performance comparisons on the CDR estimation and glaucoma screening. (a) shows the mean absolute CDR estimation errors by DeepLab V2 [17], MNet [11] and ours. (b) shows the ROC curves and the AUC scores for glaucoma screening on ORIGA dataset [19] by the CDR values estimated according to the segmentation masks by the proposed SAN, DeepLab V2 [17] and MNet [11].

Table 2
Influences of modules in SAN on ORIGA [19]. Baseline: the atrous CNN module + standard dense classification. PFM: pyramid feature module. SASM: spatial-aware segmentation module.

Method	Input Image	E_{disc}	E_{cup}	E_{rim}
Baseline	Cartesian	0.080	0.244	0.273
Baseline	Polar	0.067	0.222	0.232
Baseline + PFM	Polar	0.068	0.220	0.233
Baseline + SASM	Polar	0.063	0.218	0.222
Baseline + PFM + SASM	Polar	**0.059**	**0.208**	**0.215**

Table 3
Performance comparisons of the different settings for the number of parts M on ORIGA [19].

M	E_{disc}	BLE_{disc}/std	E_{cup}	BLE_{cup}/std	E_{rim}
1	0.067	6.26/3.19	0.218	14.11/9.12	0.232
2	0.063	6.02/2.99	0.213	13.78/8.98	0.223
4	0.062	5.95/2.93	0.210	13.53/8.86	0.219
6	**0.059**	**5.59/2.91**	**0.208**	**13.40/8.93**	**0.215**
8	0.059	5.50/2.72	0.210	13.47/8.88	0.214

performances. Integrating with the spatial-aware segmentation module (SASM) further reduces the overlapping errors. When we integrate the baseline with PFM and SASM, the overlapping errors are further reduced.

Validation of the spatial-aware segmentation module. To investigate how the explicit process for the part with similar spatial properties contributes to the joint segmentation, we train five models respectively by setting the number of the part M to be one, two, four, six and eight. The performances are reported in Table 3. We can see that the overlapping errors E_{disc}, E_{cup}, E_{rim} as well as the boundary localisation errors BLE_{disc} and BLE_{cup} decrease as the number of part M increases from one to six. This implies that explicit process for the specific part in the ONH image is capable to handle the variance of the spatial layout of the vessels in the ONH images and improve the segmentation performances. When M increases to eight, the performances change slightly.

Validation of the pyramid filtering module. To investigate how the multi-scale features contribute to the joint segmentation, we range the number of scales from one to five and train five models. The performances on ORIGA [19] are reported in Table 4. Obviously, models using multi-scale features (i.e. $R \geq 2$) achieve better performances than the model extracting single scale features

Table 4
Performance comparisons of the different setting for the number of scales R on ORIGA [19].

R	E_{disc}	BLE_{disc}/std	E_{cup}	BLE_{cup}/std	E_{rim}
1	0.063	5.96/3.44	0.218	13.92/8.90	0.222
2	0.061	5.80/2.92	0.215	13.68/8.89	0.219
3	0.062	5.85/2.82	0.213	13.60/9.35	0.219
4	**0.059**	**5.59/2.91**	**0.208**	**13.40/8.93**	**0.215**
5	0.060	5.72/2.96	0.210	13.45/8.96	0.217

Table 5
Performance comparisons of the different methods on DRISHTI [20].

Methods		OD		OC	
		Dice	BLE/std	Dice	BLE/std
Hand-crafted	R-bend [5]	0.96	8.93/2.96	0.77	30.51/24.80
	Multiview [52]	0.96	8.93/2.96	0.79	25.28/18.00
	Superpixel [9]	0.95	9.38/5.75	0.80	22.04/12.57
	Graph cut prior [30]	0.94	14.74/15.66	0.77	26.70/16.67
	BB-CRF [45]	0.97	6.61/3.55	0.83	18.61/13.02
Deep learning	MNet [11]	0.95	8.93/21.15	0.87	17.00/10.19
	DeepLab V2 [17]	0.98	4.56/1.78	0.87	16.20/11.98
	Ours	**0.98**	**4.22/1.72**	**0.89**	**13.55/9.07**

(i.e. $R = 1$). When the number of the scales increases from one to four, the segmentation performances are gradually improved. This implies that the multi-scale features contribute to the joint segmentation. As the number of the scales increases from four to five, the performances degrade slightly. This is because the atrous rate of the spatial filters used to produce the features at the fifth scale becomes 30. Spatial filters with such a large atrous rate tends to degrade to simple 1×1 filters [48] and no more context information is exploited when M increases from four to five.

4.4. Results on DRISHTI

Segmentation evaluation. On DRISHT [20], we compare the proposed SAN with five hand-crafted feature based segmentation methods, i.e., R-bend [5], Multiview [52], Superpixel [9], Graph cut prior [30] and BB-CRF [45], and two deep learning based methods MNet [11] and DeepLab V2 [17]. The performances are reported in Table 5. It is obvious that: (1) the deep learning based methods achieve better performances than most of the hand-crafted methods, especially on OC segmentation task; (2) the dice coefficient

of the proposed SAN and DeepLab V2 [17] reach 98%, which indicates that the OD segmentation ability of both methods on DRISHT [20] almost reach the expert level; (3) compared with the state of the arts, the proposed SAN gets better performances on OC segmentation and comparable performances on OD segmentation. The visual comparisons on OD and OC segmentation to MNet [11] and DeepLab V2 [17] are illustrated in Fig. 5 and Fig. 6 respectively. As we can see from Fig. 5, MNet [11] fails on the OD segmentation tasks while our SAN and DeepLab V2 [17] work well. As shown in Fig. 6, the boundaries of the OC masks by our SAN are closer to the manual OC boundaries than those by MNet [11] and DeepLab V2 [17].

5. Glaucoma screening

Glaucoma is the foremost cause of the irreversible vision impairment and irreversible blindness [53]. An important measure for the glaucoma diagnosis in clinical is the CDR value. Generally, the larger CDR value is, the higher risk of glaucoma is. Unlike direct glaucoma screening which maps the image space to image-level label space such as [54], we first apply our spatial-aware segmentation method to the CDR estimation, and then screen glaucoma according to the CDR.

We compare the performances of our method both on the CDR estimation and glaucoma screening with two deep models, i.e., the MNet [11] and DeepLab V2 [17]. We only evaluate the proposed method on ORIGA [19] since it provides the clinical diagnoses.

To evaluate the CDR estimation, we follow [11] and use the absolute CDR error δ_E, which is defined as:

$$\delta_E(\hat{CDR}, CDR_{GT}) = |\hat{CDR} - CDR_{GT}|, \tag{16}$$

where $\hat{CDR}$ is estimated based on the predicted OC and OD segmentation masks, and CDR_{GT} is manually estimated by the experienced clinician. Fig. 7(a) shows the average errors. DeepLab V2 [17] and MNet [11] get the average absolute CDR error 0.071 while ours gets much lower average absolute CDR error 0.066.

To show the correlation between the estimated CDR value and glaucoma, we report the Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC), as introduced in [11]. To plot the ROC curve, the CDR values are first thresholded and the subjects whose CDR values are larger than the threshold value are regarded as glaucomatous suspect. Then we calculate the false positive rate and true positive rate. By changing the threshold value, the ROC curve is obtained and AUC is computed. Fig. 7(b) shows the ROC curves and AUC values. The AUCs of DeepLab V2 [17], MNet [11] and our SAN are 83.77%, 85.08% and 85.56% respectively. Ours is 1.79% higher than the DeepLab V2 [17] and 0.48% higher than the MNet [11]. This reflects that the CDR values computed based on the segmentation masks by our SAN has highest correlation to the glaucoma screening.

One thing worth to mention here is that the performances of different segmentation methods on the joint segmentation task, CDR calculation task and glaucoma screening task are not always consistent. (1) The performances of DeepLab V2 [17] on segmentation tasks are better than MNet [11] while their average CDR errors are same. This is reasonable because the pixel domain used to measure the performances for the segmentation task and the CDR estimation task are different. The former is measured with regard to all the pixels in the ONH images while the latter is only with regard to four points defining the vertical OC diameter and OD diameter. (2) The average CDR errors of the DeepLab V2 [17] and the MNet [11] are same while the AUC for glaucoma screening of the DeepLab V2 [17] is worse than that of the MNet [11]. This is reasonable because both overestimation for the CDR values of the glaucoma cases and underestimation for the CDR values of the normal cases not only result high AUC when performing glaucoma prediction, but also result high average CDR errors. Nevertheless, our spatial-aware method consistently achieves better performances on the joint segmentation task, the CDR calculation task and the glaucoma screening task than DeepLab V2 [17] and MNet [11].

6. Conclusion and future work

This paper proposes a spatial-aware joint OD and OC segmentation method and designs a spatial-aware neural network to implement it. Experimental results on two public datasets demonstrate its effectiveness on the OD and OC segmentation. Based on the segmentation masks, the CDR is estimated and shows high correlation to the risk of glaucoma, which indicates that the proposed method can play a significant role in the glaucoma screening system and assist the ophthalmologist to assess the progression of glaucoma. Our method has the following advantages:

- It is able to distinguish the OD region from the confounding bright structures surrounding the OD by extracting spatially dense context features with large size of the receptive field of view, and achieves the state-of-the-art performances on the OD segmentation on the datasets ORIGA [19] and DRISHTI [20].
- It is able to distinguish the valid kinks from the invalid kinks by explicitly considering the spatial locations, and performs better on the OC segmentation on the datasets ORIGA [19] and DRISHTI [20] than the previous methods.
- It obtains better performances when applying it to the CRD estimation and the glaucoma screening on ORGIA dataset [19].

We note that our segmentation method also can help the prediction of other optic nerve head related eye diseases such as the onarteritic anterior ischaemic optic neuropathy [55] and optic neuritis [56], etc. This will be our future work. Additionally, the idea of explicitly processing the regions with similar spatial layout in a different way can be used to the segmentation tasks on other biomedical images that also have the similar characteristics as the ONH images, for example the cardiac ventricle MR images.

Conflict of interest

None.

Acknowledgments

The authors would like to thank the Cixi Institute of BioMedical Engineering for providing us the ORIGA dataset and the Medical Image Processing (MIP) group, IIIT Hyderabad for providing us the DRISHT dataset.

This work of Q. Liu is supported by the China Postdoctoral Science Foundation under Grant no. 2017M620356 and the Postdoctoral Science Foundation of Central South University; B. Zou and G. Zhao are partially supported by the International (Regional) Joint Research Program of Hunan Province under Grant no. 2017WK2074; B. Zou and Z. Chen are partially supported by the National Natural Science Foundation of China under Grant nos. 61573380 and 61672542.

Appendix A. The detailed configuration for SAN

In this section, we provide the detailed configuration for the proposed SAN. As shown in Fig. 8, it consists of an atrous CNN module (see Fig. 9), and a pyramid filtering module including six parallel PFBs (see Fig. 10), and a segmentation module including

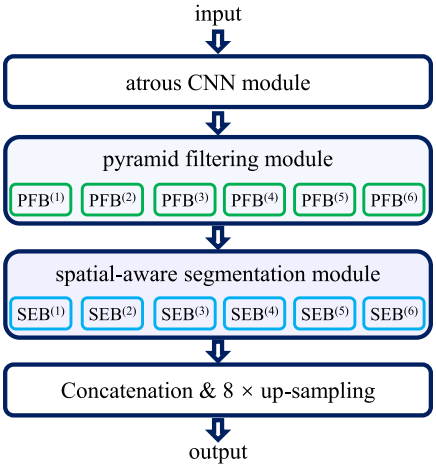


Fig. 8. The architecture of the proposed SAN. It consists of the atrous CNN module, and the pyramid filtering module including six PFBs, and the spatial-aware segmentation module including six SEBs, and the concatenation and 8× up-sampling operators.

six parallel SEBs, (see Fig. 10), the concatenation and 8× up-sampling operators. In these figures, the configuration for each type of layers are illustrated as follows:

- convolutional layer
  Conv < receptive field size> − <number of channels >
- atrous convolutional layer
  AtrousConv < receptive field size> − <number of channels> − <atrous rate >
- max pooling layer
  MaxPooling < receptive field size> − <stride >
- dropout
  Dropout- < dropout rate >

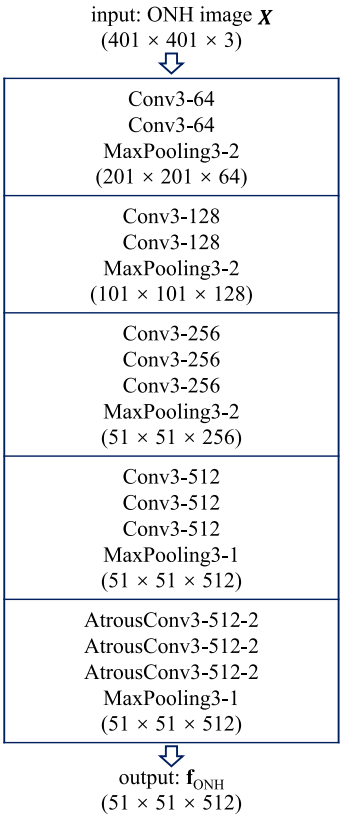


Fig. 9. The detailed configuration for the atrous CNN. It consists of five stages. The first four stages are equipped with two or three convolutional layers (Conv), and one max pooling layer (MaxPooling). The last stage consists of three atrous convolutional layers (AtrousConv) and one max pooling layer (MaxPooling). We note that each Conv/AtrousConv layer is equipped with the ReLU activation layer [41], which is not shown for brevity.

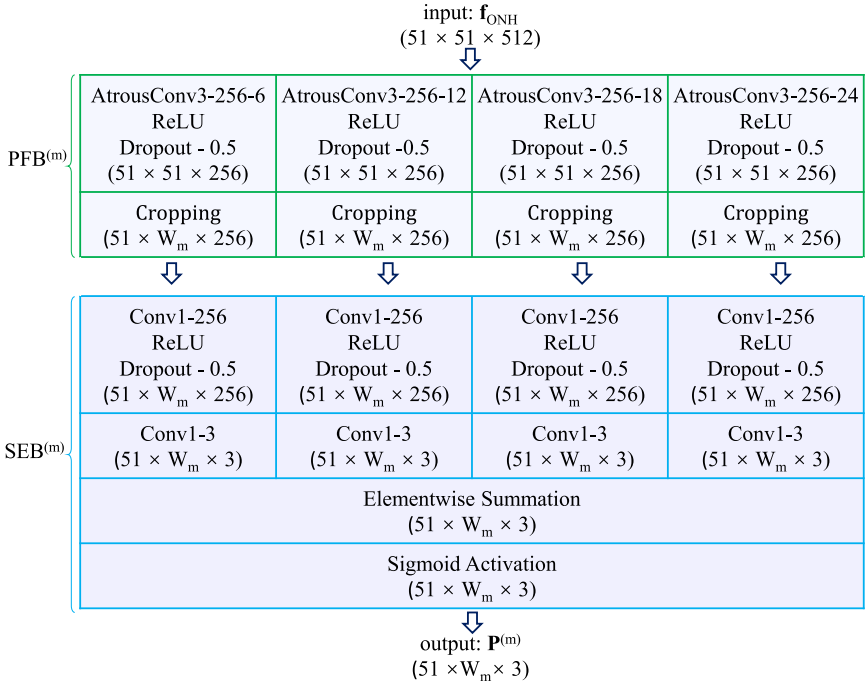


Fig. 10. The detailed configuration for the pyramid filtering block (i.e. PFB$^{(m)}$ in Fig. 8) and spatial-aware segmentation block (i.e. SEB$^{(m)}$ in Fig. 8). The input of PFB$^{(m)}$ is the output of the atrous CNN module, i.e., $\mathbf{f}_{ONH}$, and its output is directly fed into SEB$^{(m)}$. SEB$^{(m)}$, yielding the probability matrix $\mathbf{P}^{(m)} \in \mathcal{R}^{51\times W_m\times 3}$ for part m. Each element $\mathbf{P}^{(m)}_{i,j,t}$ is the probability that the point $[i, j]$ belongs to class t, where $t \in \{oc, rim, background\}$. W_m is the width of the mth part of $\mathbf{f}_{ONH}$. It equals to eight if $m \in \{1, 3, 4, 6\}$, or equals to ten if $m = 2$, or equals to nine if $m = 5$.

The shape of each stage's output is given in the bracket and denoted as (height × width × number of channels).

References

[1] B. Schwartz, Cupping and pallor of the optic disc, Arch. Ophthalmol. 89 (4) (1973) 272–277.
[2] J.B. Jonas, A. Bergua, P. Schmitz Valckenberg, K.I. Papastathopoulos, W.M. Budde, Ranking of optic disc variables for detection of glaucomatous optic nerve damage, Investig. Ophthalmol. Vis. Sci. 41 (7) (2000) 1764.
[3] A. Salazar-Gonzalez, D. Kaba, Y. Li, X. Liu, Segmentation of the blood vessels and optic disk in retinal images, IEEE J. Biomed. Health Inform. 18 (6) (2014) 1874–1886.
[4] G.D. Joshi, J. Sivaswamy, K. Karan, P. R., S.R. Krishnadas, Vessel bend-based cup segmentation in retinal images, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 2536–2539.
[5] G.D. Joshi, J. Sivaswamy, S.R. Krishnadas, Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment, IEEE Trans. Med. Imaging 30 (6) (2011) 1192–1205.
[6] W.W.K. Damon, J. Liu, T.N. Meng, Y. Fengshou, W.T. Yin, Automatic detection of the optic cup using vessel kinking in digital retinal fundus images, in: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), 2012, pp. 1647–1650.
[7] D.W.K. Wong, J. Liu, N.M. Tan, F. Yin, B.H. Lee, Y.C. Tham, C. Cheung, T.Y. Wong, Detecting the optic cup excavation in retinal fundus images by automatic detection of vessel kinking, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 73–76.
[8] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, N.-M. Tan, C.-Y. Cheng, Y.C. Tham, T.Y. Wong, Superpixel classification based optic disc segmentation, in: K.M. Lee, Y. Matsushita, J.M. Rehg, Z. Hu (Eds.), Computer Vision – ACCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 293–304.
[9] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, N.M. Tan, D. Tao, C.Y. Cheng, T. Aung, T.Y. Wong, Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, IEEE Trans. Med. Imaging 32 (6) (2013) 1019–1032.
[10] A. Sevastopolsky, Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network, Pattern Recognit. Image Anal. 27 (3) (2017) 618–624.
[11] H. Fu, J. Cheng, Y. Xu, D.W.K. Wong, J. Liu, X. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, IEEE Trans. Med. Imaging PP (99) (2018). 1–1
[12] B. Al-Bander, B.M. Williams, W. Al-Nuaimy, M.A. Al-Taee, H. Pratt, Y. Zheng, Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis, Symmetry 10 (2018).
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[17] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[18] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 636–644.
[19] Z. Zhang, F.S. Yin, J. Liu, W.K. Wong, N.M. Tan, B.H. Lee, J. Cheng, T.Y. Wong, Origa-light: an online retinal fundus image database for glaucoma analysis and research, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 3065–3068.
[20] J. Sivaswamy, S.R. Krishnadas, G.D. Joshi, M. Jain, A.U.S. Tabish, Drishti-GS: retinal image dataset for optic nerve head(ONH) segmentation, in: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), 2014, pp. 53–56.
[21] H. Xiaopeng, Z. Guoying, P. Matti, C. Xilin, Combining LBP difference and feature correlation for texture description, IEEE Trans. Image Process. 23 (6) (2014) 2557–2568.
[22] H. Xiaopeng, Z. Guoying, Z. Stefanos, P. Maja, P. Matti, Capturing correlations of local features for image representation, Neurocomputing 184 (2016) 99–106.
[23] A. Aquino, M.E. Gegundez-Arias, D. Marin, Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques, IEEE Trans. Med. Imaging 29 (11) (2010) 1860–1869.
[24] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, L. Kennedy, Optic nerve head segmentation, IEEE Trans. Med. Imaging 23 (2) (2004) 256–264.
[25] H. Yu, E.S. Barriga, C. Agurto, S. Echegaray, M.S. Pattichis, W. Bauman, P. Soliz, Fast localization and segmentation of optic disk in retinal images using directional matched filtering and level sets, IEEE Trans. Inf. Technol. Biomed. 16 (4) (2012) 644–657.
[26] P.S. Mittapalli, G.B. Kande, Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma, Biomed. Signal Process. Control 24 (2016) 34–46.
[27] Q. Liu, B. Zou, J. Chen, W. Ke, K. Yue, Z. Chen, G. Zhao, A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images, Comput. Med. Imaging Graph. 55 (2017) 78–86. Special Issue on Ophthalmic Medical Image Analysis.
[28] B. Zou, Q. Liu, K. Yue, Z. Chen, J. Chen, G. Zhao, Saliency-based segmentation of optic disc in retinal images, Chin. J. Electron. 28 (1) (2019) 71–75.
[29] Y. Xu, L. Duan, S. Lin, X. Chen, D.W.K. Wong, T.Y. Wong, J. Liu, Optic cup segmentation for glaucoma detection using low-rank superpixel representation, in: P. Golland, N. Hata, C. Barillot, J. Hornegger, R. Howe (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, Springer International Publishing, Cham, 2014, pp. 788–795.
[30] Y. Zheng, D. Stambolian, J. O'Brien, J.C. Gee, Optic disc and cup segmentation from color fundus photograph using graph cut with priors, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 75–82.
[31] S. Roychowdhury, D.D. Koozekanani, S.N. Kuchinka, K.K. Parhi, Optic disc boundary and vessel origin segmentation of fundus images, IEEE J. Biomed. Health Inform. 20 (6) (2016) 1562–1574.
[32] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, B.H. Lee, C. Cheung, T. Aung, T.Y. Wong, Superpixel classification for initialization in model based optic disc segmentation, in: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 1450–1453.
[33] A. Li, Z. Niu, J. Cheng, F. Yin, D.W.K. Wong, S. Yan, J. Liu, Learning supervised descent directions for optic disc segmentation, Neurocomputing 275 (2018) 350–357.
[34] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, L. Van Gool, Deep retinal image understanding, in: S. Ourselin, L. Joskowicz, M.R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, 2016, pp. 140–148.
[35] S. Sedai, D. Mahapatra, S. Hewavitharanage, S. Maetschke, R. Garnavi, Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder, in: M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D.L. Collins, S. Duchesne (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, Springer International Publishing, Cham, 2017, pp. 75–82.
[36] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2013.
[37] J. Cheng, Z. Li, Z. Gu, H. Fu, D.W.K. Wong, J. Liu, in: Structure-preserving guided retinal image filtering and its application for optic disc analysis, 2018. arXiv:1805.06625.
[38] Z. Gu, P. Liu, K. Zhou, Y. Jiang, H. Mao, J. Cheng, J. Liu, Deepdisc: optic disc segmentation based on atrous convolution and spatial pyramid pooling, in: Computational Pathology and Ophthalmic Medical Image Analysis, Springer, 2018, pp. 253–260.
[39] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imaging (2019) 1–1. doi:10.1109/TMI.2019.2903562.
[40] H. Salehinejad, S. Valaee, T. Dowdell, J. Barfett, Image augmentation using radial transform for training deep neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.
[41] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
[42] J. Cheng, D. Tao, D.W.K. Wong, J. Liu, Quadratic divergence regularized SVM for optic disc segmentation, Biomed. Opt. Express 8 (5) (2017) 2687–2696.
[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22Nd ACM International Conference on Multimedia, in: MM '14, 2014, pp. 675–678.
[44] J. Cheng, J. Liu, D.W.K. Wong, F. Yin, C. Cheung, M. Baskaran, T. Aung, T.Y. Wong, Automatic optic disc segmentation with peripapillary atrophy elimination, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 6224–6227.
[45] A. Chakravarty, J. Sivaswamy, Joint optic disc and cup boundary extraction from monocular fundus images, Comput. Methods Programs Biomed. 147 (2017) 51–61.
[46] F. Yin, J. Liu, S.H. Ong, Y. Sun, D.W.K. Wong, N.M. Tan, C. Cheung, M. Baskaran, T. Aung, T.Y. Wong, Model-based optic nerve head segmentation on retinal fundus images, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 2626–2629.
[47] Y. Jiang, L. Duan, J. Cheng, Z. Gu, H. Xia, H. Fu, C. Li, J. Liu, JointRCNN: a region-based convolutional neural network for optic disc and cup segmentation, IEEE Trans. Biomed. Eng. (2019) 1–1. doi:10.1109/TBME.2019.2913211.
[48] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, in: Rethinking atrous convolution for semantic image segmentation, 2017. arXiv:1706.05587.
[49] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, in: Pyramid scene parsing network, 2017, pp. 2881–2890.
[50] G. Huang, Z. Liu, L. v. d. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.
[51] Z. Huang, Y. Wei, X. Wang, W. Liu, A pytorch semantic segmentation toolbox, 2018. (https://github.com/speedinghzl/pytorch-segmentation-toolbox).

[52] G.D. Joshi, J. Sivaswamy, S.R. Krishnadas, Depth discontinuity-based cup segmentation from multiview color retinal images, IEEE Trans. Biomed. Eng. 59 (6) (2012) 1523–1531.
[53] Y.-C. Tham, X. Li, T.Y. Wong, H.A. Quigley, T. Aung, C.-Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis, Ophthalmology 121 (11) (2014) 2081–2090.
[54] H. Fu, J. Cheng, Y. Xu, C. Zhang, D.W.K. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Trans. Med. Imaging 37 (11) (2018) 2493–2501.
[55] N.R. Miller, A.C. Arnold, Current concepts in the diagnosis, pathogenesis and management of nonarteritic anterior ischaemic optic neuropathy, Eye 29 (1) (2015) 65–79.
[56] V. Biousse, N.J. Newman, Diagnosis and clinical features of common optic neuropathies, Lancet Neurol. 15 (13) (2016) 1355–1367.

Qing Liu received the bachelor degree and Ph.D. degree in computer science and technology from the Central South University, Changsha, China, in 2011 and 2017 respectively. She is currently a postdoc researcher in Central South University. Her research interests include salient object detection and medical image analysis. She has authored or co-authored more than 10 papers in journals and conferences.

Xiaopeng Hong received his Ph.D. degree in computer application and technology from Harbin Institute of Technology, P. R. China, in 2010. He is now an associate Professor at Xian Jiaotong University. He was a Docent with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where he had been a scientist researcher from 2011 to 2018. Dr. Hong has published over 30 articles in mainstream journals and conferences such as IEEE T-PAMI, T-IP, CVPR and ACM UbiComp. His current research interests include multi-modal learning, affective computing, intelligent medical examination, and human-computer interaction, etc. His research has been reported by global media including MIT Technology Review and Daily Mail.

Shuo Li received his Ph.D. degree in computer science from Concordia University in 2006, Montreal, QC, Canada. He is an adjunct research professor at the Western University and adjunct scientist at the Lawson Health Research Institute. He is currently leading the Digital Imaging Group of London as the Scientific Director. His current research interests include automated medial image analysis and visualization.

Zailiang Chen received the Ph.D. degree in computer science from Central South University in 2012. He is currently an Associate Professor. He has authored or co-authored more than 40 papers in journals and conferences. His recent research interests include computer vision and medical image analysis.

Guoying Zhao is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland and a professor with the School of Information and Technology, Northwest University, China. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She has authored or co-authored more than 190 papers in journals and conferences. Her papers have currently over 9700 citations in Google Scholar (h-index 43). Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, dynamic texture recognition, human motion analysis, and person identification. Dr. Zhao was a Co-Chair of many International Workshops at ECCV, ICCV, CVPR, ACCV and BMVC. She was co-publicity chair for FG2018, has served as Area Chairs for several conferences. Currently, she is Associate Editor for Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and Image and Vision Computing Journals.

Beiji Zou received the B.S., M.S., and Ph.D. degrees from Zhejiang University in 1982, Qinghua University in 1984 and Hunan University in 2001 respectively. He is currently a Professor at the School of Computer Science and Engineering at Central South University. His research interests include computer graphics and image processing.