Artificial Intelligence In Medicine 126 (2022) 102261

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Weak label based Bayesian U-Net for optic disc segmentation in fundus images

Hao Xiong [*], Sidong Liu, Roneel V. Sharan, Enrico Coiera, Shlomo Berkovsky

Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

ARTICLE INFO

Keywords:
Optic disc segmentation
Bayesian U-Net
Expectation-maximization
Weak labels
Fundus image

ABSTRACT

Fundus images have been widely used in routine examinations of ophthalmic diseases. For some diseases, the pathological changes mainly occur around the optic disc area; therefore, detection and segmentation of the optic disc are critical pre-processing steps in fundus image analysis. Current machine learning based optic disc segmentation methods typically require manual segmentation of the optic disc for the supervised training. However, it is time consuming to annotate pixel-level optic disc masks and inevitably induces inter-subject variance. To address these limitations, we propose a weak label based Bayesian U-Net exploiting Hough transform based annotations to segment optic discs in fundus images. To achieve this, we build a probabilistic graphical model and explore a Bayesian approach with the state-of-the-art U-Net framework. To optimize the model, the expectation-maximization algorithm is used to estimate the optic disc mask and update the weights of the Bayesian U-Net, alternately. Our evaluation demonstrates strong performance of the proposed method compared to both fully- and weakly-supervised baselines.

1. Introduction

Retinal optic disc analysis in fundus imaging plays a pivotal role in identifying ophthalmic diseases [1–8]. The size, shape, depth and pathological changes of the optic disc are regarded as important index to judge retinopathy, such as glaucoma. As such, efficient optic disc segmentation is an important preprocessing step that can be incorporated into many automatic systems for eye related disease screening. Rather than manually identifying the optic disc, efficient and automatic optic disc segmentation has drawn widespread research attention.

Recent advances in deep learning substantially boost the performance of optic disc segmentation. Albeit deep learning approaches demonstrated strong performance, they are highly dependent on the ground truth optic disc annotations needed for the fully-supervised learning. In real-life scenarios, fully-supervised training of optic disc segmentation is a complex task, due to its inherent reliance on a large-scale optic disc segmentation training data containing fundus images with pixel-wise annotations. However, manually producing pixel-wise annotation is time consuming and expensive, and may also induce inter-annotator variance.

In addition to fully-supervised methods, Saha et al. proposed a weakly labeled multi-task learning method to segment vessels, lesions and optic disc concurrently [9]. Two datasets, one missing the vessel annotations and the other missing the lesion and optic disc annotations, were used for training, practically yielding a semi-supervised method using only partial ground truth annotation. Such partial annotations were created manually by domain experts, thus, rendering non-scalable. The limitations of fully-supervised and weakly labeled methods are evident; yet, a few weakly-supervised deep learning methods for optic disc segmentation, not relying on manually annotated pixel-level ground truth, have been proposed. Lu et al. proposed a weakly-supervised optic disc segmentation method using pseudo ground truth annotations [10]. They deployed a constrained CNN with image-level labels to produce a rough disc segmentation map and then refined the map by setting the area outside bounding box as a background. Pseudo ground truth maps were obtained by fusing these two disc segmentation maps and fed into a modified U-Net for the supervised learning.

However, the pseudo ground truth masks derived from the image-level and bounding box labels may be unreliable. This is mainly due to the image and bounding box being weak labels, not necessarily allowing to properly recognize optic disc area from fundus images during pseudo mask generation. Hence, the derived pseudo ground truth masks may turn out inaccurate (Fig. 1(d)) and limit the disc segmentation performance. To address the shortcomings of the image-level and

[*] Corresponding author.
E-mail address: hao.xiong@mq.edu.au (H. Xiong).

https://doi.org/10.1016/j.artmed.2022.102261
Received 3 June 2021; Received in revised form 18 January 2022; Accepted 20 February 2022
Available online 26 February 2022

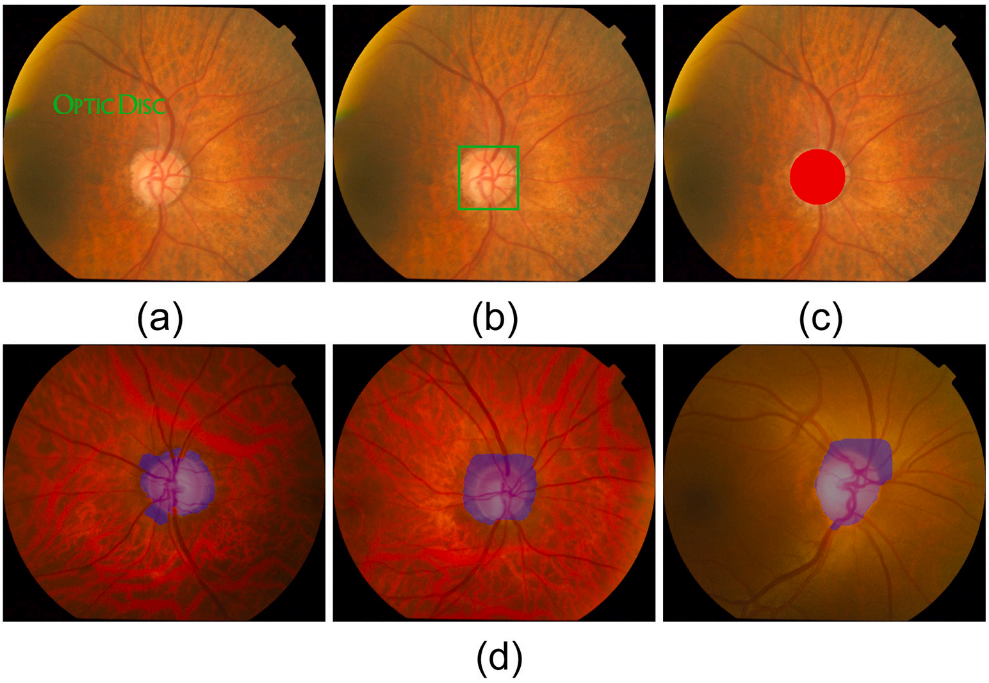

Fig. 1. Different weak labels for learning optic disc segmentation: (a) image-level label (semantic label) (b) bounding box label (c) Hough transform based label (d) pseudo masks (purple) generated by image-level and bounding box labels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

bounding box labels, we exploit the Hough transform based label as the main pseudo ground truth mask for the disc segmentation task (visual comparison is shown in Fig. 1). The Hough transform is able to detect a circular mask that largely covers the optic disc area [11]. Although more accurate than the image-level and bounding box labels, the Hough transform based labels have two limitations: (i) they are less accurate than the manual expert-annotated labels, and (ii) some fundus images cannot be successfully labeled by the Hough transform.

To alleviate these, in this work we propose a weak label based Bayesian U-Net built upon a novel probabilistic graphical model. To derive this model, the U-Net is represented as a Bayesian neural network by assigning priors to the network weights. Then, the derived Bayesian U-Net considers uncertainty during training. This benefits the subsequent inference stage, where a number of probability maps can be sampled from the trained Bayesian U-Net. Finally, the predictions are generated by averaging the probability maps. Compared to traditional networks having only one prediction, our approach generates multiple predictions and only highly probable pixels are included in the eventual mask. As a result, our Bayesian approach is demonstrated to be more robust and accurate than the non-Bayesian variant using the Hough transform.

For some fundus images, the optic disc masks cannot be labeled, since they cannot be detected by the Hough transform. In this case, a latent variable and a new variable are introduced into the graphical model to represent the masks of the failed fundus images and bounding box label, respectively. The proposed graphical model contains both latent variables and network weights for estimation during training. The Expectation-Maximization (E-M) algorithm is applied to estimate the latent variable and optimize the network parameters alternately, until convergence. In each iteration, the network is learned at the M-step using the estimated optic disc mask from the E-step. In the next iteration, the bounding box label and the optimized network from the previous iteration ensure more accurate estimation of the optic disc mask for subsequent network learning. This way, all the images have the associated pseudo masks for training – including those that the disc mask labeling failed, where the pseudo masks are estimated using the E-M algorithm. Experimental evaluation shows that the alternate learning results in a more robust training than the existing weakly-supervised optic disc segmentation methods.

In summary, the motivations for designing such a Bayesian model are two-fold. First, our Bayesian model considers uncertainty of the network weights, which is harnessed to compute the predictive distribution of the optic disc segmentation task. It has been shown that predictions of Bayesian networks considering such an uncertainty are more accurate than those determinant predictions of ordinary non Bayesian networks [12–15]. In our case, it helps mitigate the inaccurate predictions by training weak labels which is obviously less accurate than manually annotated labels. Second, instead of applying variational methods to approximate Bayesian inference for neural networks [16,17], the dropout as a Bayesian approximation [12] is exploited here as a tractable approximation of our model. Dropout is exploited extensively in deep learning to avoid overfitting [18]. This is particularly useful in our scenario, since the training dataset can be small, which is likely to lead to overfitting. Moreover, the main contribution of our work refers to the proposed weak label based optic disc segmentation method that outperforms the baseline weakly-supervised methods. The method does not rely on pixel-level annotations for training and can eliminate the need for manual annotations, thereby, having a high clinical applicability and paving the way to future applications in other medical image segmentation tasks.

2. Related work

This section overviews related work on fully-supervised, weakly-supervised, and unsupervised methods for optic disc segmentation. Since there are very few works on weakly-supervised disc segmentation, we broaden the scope to weakly-supervised methods for medical images.

2.1. Fully-supervised methods

The existing fully-supervised methods can be broadly classified into two groups: traditional machine learning and deep learning based methods. Traditional approaches generally exploit brightness and morphological features to address the optic disc segmentation task. Youssif et al. proposed to identify the position of the optic disc based on the information of the blood vessels in fundus images [19]. Rather than identifying the position of the optic disc, Zou et al. utilized the morphology and eclipse fitting to extract the optic disc area [20]. Besides these, several works aimed to detect the boundary of the optic disc. Considering the active contour, [21–23] applied level-set techniques to fit the contour to the optic disc boundaries. However, level-set based methods utilized deformable contours and, given that the optic disc is circular, [24–26] exploited circular based transformation techniques to detect the optic disc boundaries. In [27–29], image-level features were extracted to train classifiers and identify the optic disc area. Later on, [30,31] proposed superpixel based classification methods that extracted various hand-crafted, high-dimensional features for disc segmentation purposes. In general, the above works relied on a single fundus image to segment the optic disc. Alternatively, Abramoff et al. exploited stereo images surfacing disparity information, to distinguish the optic disc from image background [32]. However, the accuracy of such traditional machine learning methods was not sufficiently high, such that high segmentation accuracy could not be guaranteed, limiting their clinical applicability in practice.

The accuracy of the optic disc segmentation has been upgraded using more recent deep learning methods. Several deep learning based segmentation methods using U-Net [33–38] have been proposed. Here, [33,34] applied U-Net without much modifications to segment both the optic disc and optic cup for glaucoma detection. Likewise, [35] aimed to segment the optic disc and optic cup with a context extraction module extracting additional global information. Sevastopolsky et al. developed a cascaded network to segment the optic disc based on the U-Net network [37]. Fu et al. designed a U-shape like network with multi-scale input layer to achieve multiple-level receptive field sizes, which enhanced the accuracy of the segmentation [36]. The work in [38] integrated U-Net with an attention mechanism to consider channel dependencies between different levels of features for optic disc segmentation. In addition to U-Net based methods, Tan et al. trained a CNN to perform optic disc segmentation [39] and Zilly et al. proposed an ensemble learning based method to extract the optic disc using a CNN architecture [40]. Without exception, all these deep learning based methods were fully-supervised and relied on pixel-level annotations of the optic disc for training, which rendered non-scalable.

2.2. Unsupervised and weakly-supervised methods

The appearance of the fundus image may vary significantly across datasets. As a result, the optic disc and cup segmentation method trained on one dataset may not generalize and perform poorly on images from another dataset. Hence, *unsupervised* domain adaptation is considered as an effective way to mitigate this issue [41–44]. Wang et al. proposed patch-based output space adversarial learning framework (pOSAL) to address the domain shift challenge across datasets [41]. They designed a novel morphology-aware segmentation loss and patch based training scheme to capture the segmentation details and ensure accuracy. In [42], they also proposed the boundary and entropy-driven adversarial learning aiming to improve the optic disc and cup segmentation, particularly in the ambiguous boundary regions. To further improve the generalization of CNNs in target domains, a follow-up work of Wang et al. developed a pool containing diversified domain knowledge from multiple datasets [43]. Such a pool enriched the image features that could be exploited for an accurate segmentation of images from new domains. Besides, generative adversarial networks (GAN) were exploited by Bian et al. to make the testing images look similar to the training images [44]. Following this, a network trained on training images was utilized to perform the optic disc and cup segmentation on transformed testing images. In addition to these, Norouzifard et al. proposed an improved chaotic ICA based on a traditional imperialist competitive algorithm (ICA), an unsupervised clustering algorithm for the optic disc and cup segmentation [45].

A small number of *weakly-supervised* methods have been deployed for medical image segmentation. Rajchl et al. proposed a deep cut for brain and lung segmentation, which exploited an iterative optimization based on Grab cut to obtain fake labels from bounding box labels for CNN training [46]. Likewise, Yang et al. exploited the bounding box labels for renal tumor segmentation [47]. Here, the fake labels were extracted by convolutional conditional random fields from the bounding boxes and were then used by CNNs for training. More recently, Kervadec et al. proposed to add a size constraint on the loss function so that the size of the generated fake labels could be constrained [48]. Their evaluation on cardiac image segmentation demonstrated that the performance of their weakly-supervised method was close to that of fully-supervised methods. In another weakly-supervised scenario, Rajchl et al. trained a fully convolutional network using super-pixel annotations to segment fetal MR images [49]. The super-pixel annotation refers to a set of pixels sharing a similar texture, color and brightness. Girum et al. utilized the pseudo-contour landmarks as weak labels for prostate and cardiac image segmentation [50]. Specifically, a deep generative neural network first modeled the prior-knowledge predictions using the pseudo-contour landmarks and then the predictions were refined by a fully convolutional neural network. However, only a few approaches exploited weakly-supervised learning of the optic disc segmentation without manual, pixel-level annotation. To the best of our knowledge, the only work focusing on weakly-supervised learning of the optic disc segmentation was [10], where, Lu et al. exploited image level and bounding box labels to generate pseudo ground truth masks for training. However, such weak labels were found to be unreliable and generated inaccurate pseudo masks for segmentation learning. Our work sets out to fill this gap and proposes a novel weak label based algorithm for the segmentation of the optic disc.

3. Methods

3.1. Notation

Variables $X \in R^{N \times M}$ and $Y \in R^{N \times M}$ denote the training fundus images and optic disc masks, respectively. Here, we have N images and corresponding masks of M pixels each. We use x_n and y_n to denote individual image and mask in X and Y, respectively. $y_{nm} \in \{0, 1\}$ denotes the pixel label at position m in mask y_n, where **1** represents optic disc and **0** – background. The testing images and their predicted mask are represented by x^* and y^*, respectively, and ω denotes the network weights.

3.2. Preprocessing: Hough transform based labeling

At the preprocessing stage, the optic disc mask y of an image x is labeled by the Hough transform [51]. Then, the labeled mask y is regarded as a pseudo ground truth mask for the optic disc segmentation learning.

Given an image, image processing techniques, such as the Gaussian blur and image opening, are initially applied to obtain the region of interest (ROI) containing the brightest pixels in the image. Following this, the Canny edge detector is utilized to detect the edges within the ROI [52]. Evidently, the edges normally appear on the boundary of the optic disc. The detected edges are further broadened by image dilation, to make it visible and detectable with the Hough transform. As a consequence, the Hough transform can find the circular shape on the dilated edges.

3.3. Bayesian U-Net

We use U-Net [53] as the backbone of our method and derive its Bayesian variant, which takes the weight uncertainty of the U-Net into consideration. The network architecture of the Bayesian U-Net is shown

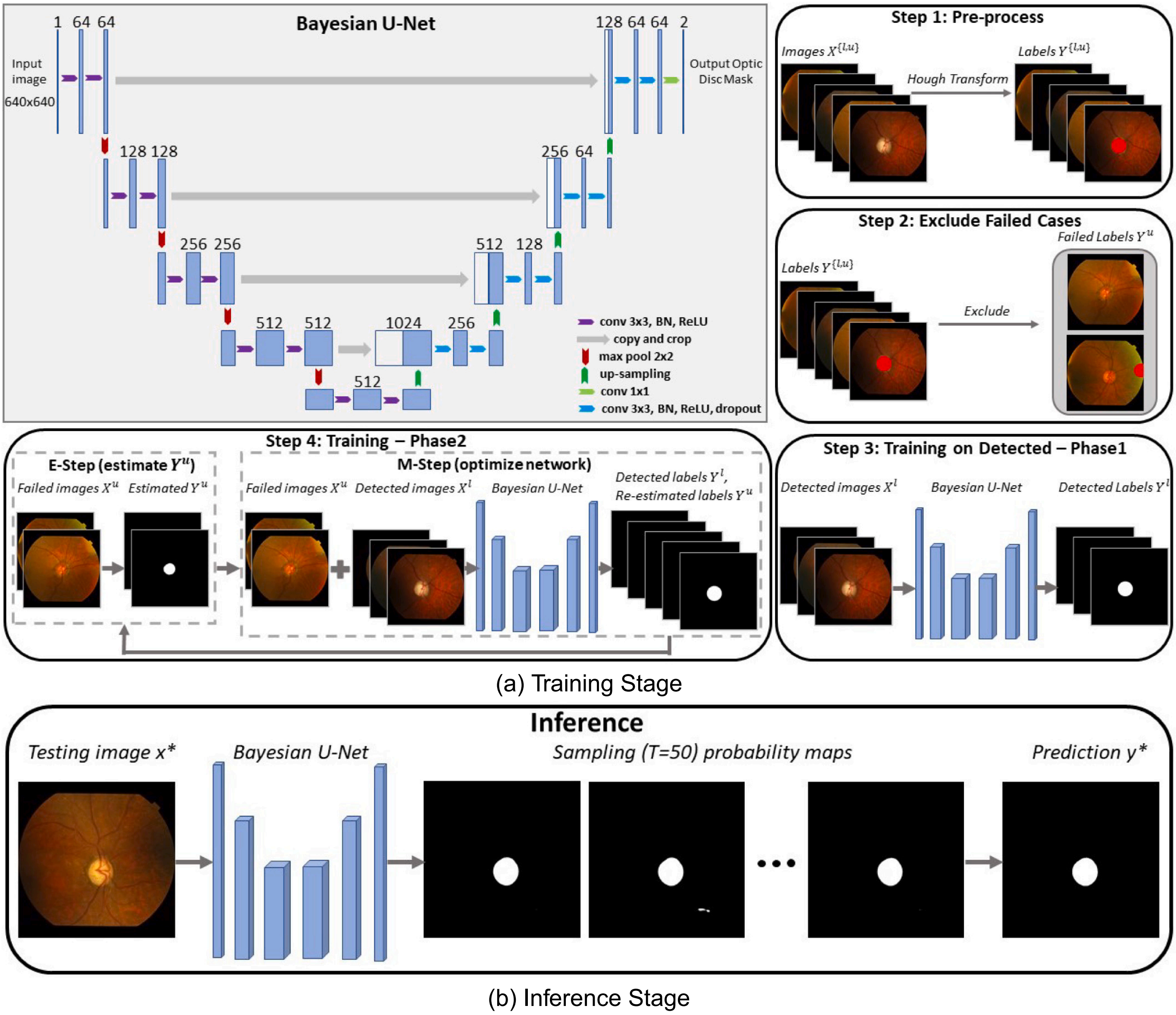

Fig. 2. (a) Training: Detailed architecture of the Bayesian U-Net (top left). Step 1: Generate pseudo ground truth masks of training images with the Hough transform. Step 2: Exclude images that cannot be labeled by the Hough transform. Images satisfying the following conditions were excluded: i) image cannot be detected by Hough transform at all, and ii) the detected Hough transform based labels substantially drift from the optic disc area. Step 3: Train the Bayesian U-Net with the detected images only. Step 4: Re-estimate the masks of the failed images at the E-Step and optimize the weights of the Bayesian U-Net using the E-M algorithm. (b) Inference: Given a testing image, 50 outputs (probability maps) are sampled directly from the Bayesian U-Net. The prediction is obtained by averaging these 50 probability maps. The final prediction is converted to a binary mask by setting the probability threshold at 0.5.

in Fig. 2. In the Bayesian U-Net, standard Gaussian prior is placed over the network weights ω, such that $\omega \sim p(\omega)$. Then, the Bayesian U-Net is defined as:

$$\begin{aligned} p(Y|X) &= \int p(Y|X,\omega)p(\omega)d\omega \\ &= \int \prod_{n=1}^{N} p(y_n|x_n,\omega)p(\omega)d\omega \\ &= \int \prod_{n=1}^{N}\prod_{m=1}^{M} p(y_{nm}|x_n,\omega)p(\omega)d\omega, \end{aligned} \tag{1}$$

where $p(y_{nm}|x_n,\omega) \propto \exp\,(f_m(y_{nm}|x_n,\omega))$. Here, $f_m(y_{nm}|x_n,\omega)$ is the output of the Bayesian U-Net at pixel m.

After training the Bayesian U-Net, the optic disc mask y^* of a testing image x^* can be predicted by the trained Bayesian U-Net as follows:

$$p(y^*|x^*,X,Y) \;\; = \int p(y^*|x^*,\omega)p(\omega|X,Y)d\omega. \tag{2}$$

Here, $p(\omega|X,Y)$ is the posterior of network weights ω, derived by:

$$p(\omega|X,Y) = \frac{p(Y|X,\omega)p(\omega)}{p(Y|X)} = \frac{p(Y|X,\omega)p(\omega)}{\int p(Y|X,\omega)p(\omega)d\omega}. \tag{3}$$

However, the integration $\int p(Y|X,\omega)p(\omega)d\omega$ in Eq. (3) is intractable. Hence, the prediction of Eq. (2) that relies on the posterior $p(\omega|X,Y)$, cannot be computed, and a variational distribution $q(\omega)$ is introduced to approximate the posterior $p(\omega|X,Y)$ and make Eq. (2) tractable. To make the distributions $q(\omega)$ and $p(\omega|X,Y)$ closer, the Kullback–Leibler (KL) divergence between $q(\omega)$ and $p(\omega|X,Y)$ is minimized as follows:

$$KL(q(\omega)\|p(\omega|X,Y)) \quad = -\underbrace{\int q(\omega)log\frac{p(\omega|X,Y)}{q(\omega)}d\omega}_{\mathrm{ELBO}\mathscr{L}(\mathrm{q}(\omega))}. \tag{4}$$

In Eq. (4), $\mathscr{L}(q(\omega))$ is also referred to as evidence lower bound (ELBO). Then, the network parameters ω can be learned by optimizing the ELBO $\mathscr{L}(q(\omega))$:

$$\mathscr{L}(q(\omega)) \quad = \int q(\omega)log\frac{p(Y|X,\omega)p(\omega)}{q(\omega)}d\omega + const. \tag{5}$$

3.3.1. Dropout as a Bayesian approximation

The ELBO $\mathscr{L}(q(\omega))$ can be further factorized as:

$$\mathscr{L}(q(\omega)) = \int q(\omega)logp(Y|X,\omega)d\omega + \int q(\omega)log\frac{p(\omega)}{q(\omega)}d\omega + const. \tag{6}$$

Since the first term in Eq. (6) is i.i.d, it can be re-written as:

$$\mathscr{L}(q(\omega)) = \sum_{n=1}^{N}\int q(\omega)logp(y_n|x_n,\omega)d\omega - KL(q(\omega)\|p(\omega)) + const. \tag{7}$$

As per [12], $\mathscr{L}(q(\omega))$ (ELBO) can be approximated as a dropout U-Net with a penalty term.

$$\mathscr{L}(q(\omega)) = \underbrace{\sum_{n=1}^{N} logp(y_n|x_n,\widehat{\omega})}_{\mathrm{U-Net\ with\ dropout}} + \underbrace{\|\omega\|^2}_{\mathrm{penalty}} + const. \tag{8}$$

Note that ω refers to the whole network weights and $\widehat{\omega}$ – to the remaining network weights after applying the dropout.

3.3.2. Introducing latent variable for unknown masks

Recall that some images may not be successfully labeled by the Hough transform. In these cases, variable Y denoting the optic disc mask becomes $Y^{\{l,u\}}$. Specifically, $Y^l \in R^{O\times M}$ denotes the masks of training images that were successfully labeled by the Hough transform, while $Y^u \in R^{P\times M}$ refers to the masks not detected by the Hough transform. O and P denote the number of detected and failed images ($N = O + P$), respectively. Hence, Y^u is the latent variable that needs to be estimated in the training. Likewise, $X^{\{l,u\}}$ are the training images of Y^l and Y^u, respectively. With these, Eq. (1) becomes:

$$p(Y|X) = \int p(Y^l|X^l,\omega)p(Y^u|X^u,\omega)p(\omega)d\omega. \tag{9}$$

To estimate the unknown masks Y^u, the bounding box labels Z^u of the failed images X^u are introduced and the newly defined graphical model becomes:

$$p(Y,Z^u|X) = \int p(Y^l|X^l,\omega)p(Y^u|X^u,\omega)p(Z^u|Y^u)p(\omega)d\omega. \tag{10}$$

The predictive distribution becomes:

$$p(y^*|x^*,Y,Z^u,X) = \int p(y^*|x^*,\omega)p(\omega|Y,Z^u,X)d\omega. \tag{11}$$

Likewise, a variational distribution $q(\omega)$ is introduced to approximate the posterior $p(\omega|\, Y,Z^u,X)$. By repeating the steps from Eqs. (4)–(7), the KL divergence $KL(q(\omega)\|p(\omega|\, Y,Z^u,X))$ between $q(\omega)$ and $p(\omega|\, Y,Z^u,X)$ is derived. Then, the new ELBO $\mathscr{L}(q(\omega),Z^u)$ on the KL divergence becomes:

$$\mathscr{L}(q(\omega),Z^u) = \sum_{n=1}^{P}\int q(\omega)logp\left(y_n^u|x_n^u,\omega\right)p\left(z_n^u|y_n^u\right)d\omega + \sum_{n=1}^{O}\int q(\omega)logp\left(y_n^l|x_n^l,\omega\right)d\omega - KL(q(\omega)\|p(\omega)) + const. \tag{12}$$

Here, ${x_n}^{\{l,u\}}$, ${y_n}^{\{l,u\}}$ and ${z_n}^u$ denote individual images and their own corresponding masks in $X^{\{l,u\}}$, $Y^{\{l,u\}}$ and Z^u.

By applying the aforementioned dropout as a Bayesian approximation, we obtain:

$$\mathscr{L}(q(\omega),Z^u) = \sum_{n=1}^{P} logp\left(y_n^u|x_n^u,\widehat{\omega}\right)p\left(z_n^u|y_n^u\right) + \sum_{n=1}^{O} logp\left(y_n^l|x_n^l,\widehat{\omega}\right) + \|\omega\|^2 + const, \tag{13}$$

where we have two unknown variables to learn: optic disc mask ${y_n}^u$ and network weights ω.

3.4. Learning with expectation-maximization

The Expectation-Maximization (E-M) algorithm (see Fig. 2, step 4) is applied to alternately estimate ${y_n}^u$ and ω in Eq. (13). Suppose ω' is the previous estimation of network weights ω and $\mathscr{L}(q(\omega),Z^u) > \mathscr{L}(q(\omega'), Z^u)$. Then, the expected complete-data log-likelihood $\mathscr{Q}(\omega;\omega')$ given ω' is:

$$\mathscr{Q}(\omega;\omega') = \sum_{n=1}^{P}\sum_{y_n^u} p\left(y_n^u|x_n^u,z_n^u,\omega'\right)logp\left(y_n^u|x_n^u,\widehat{\omega}\right) + \sum_{n=1}^{O} logp\left(y_n^l|x_n^l,\widehat{\omega}\right) + \|\omega\|^2. \tag{14}$$

By adopting a hard-EM approximation, we obtain:

$$\mathscr{Q}(\omega;\omega') \approx \sum_{n=1}^{P} logp\left(\widehat{y}_n^u|x_n^u,\widehat{\omega}\right) + \sum_{n=1}^{O} logp\left(y_n^l|x_n^l,\widehat{\omega}\right) + \|\omega\|^2, \tag{15}$$

where $\widehat{y}_n^u$ is the optimal estimation of the disc mask. We proceed to the details of the E-step and M-step.

3.4.1. E-step

The E-Step aims to infer the optic disc mask $\widehat{y}_n^u$ using the posterior from Eq. (14), $p({y_n}^u|\, {x_n}^u,{z_n}^u,\omega') \propto p({y_n}^u|\, {x_n}^u,\omega')p({z_n}^u|\, {y_n}^u)$:

$$\widehat{y}_n^u = \underset{y_n^u}{argmax} logp\left(y_n^u|x_n^u,\omega'\right)p\left(z_n^u|y_n^u\right) = \underset{y_n^u}{argmax}\sum_{m=1}^{M} f_m\left(y_{nm}^u|x_n^u,\omega'\right) + logp\left(z_n^u|y_n^u\right). \tag{16}$$

Here, $f_m\left(y_{nm}^u|x_n^u,\widehat{\omega}\right)$ is the output of the Bayesian U-Net at pixel m. However, mask ${y_n}^u$ estimated only by $p({y_n}^u|\, {x_n}^u,\omega')$ may be noisy (see Fig. 3). Thus, $p({z_n}^u|\, {y_n}^u)$ is another term helping to remove the noise, is defined as:

$$logp\left(z_n^u|y_n^u\right) = \sum_{m=1}^{M}\phi\left(y_{nm}^u,z_n^u\right), \tag{17}$$

where $\phi({y_{nm}}^u,{z_n}^u)$ is defined as:

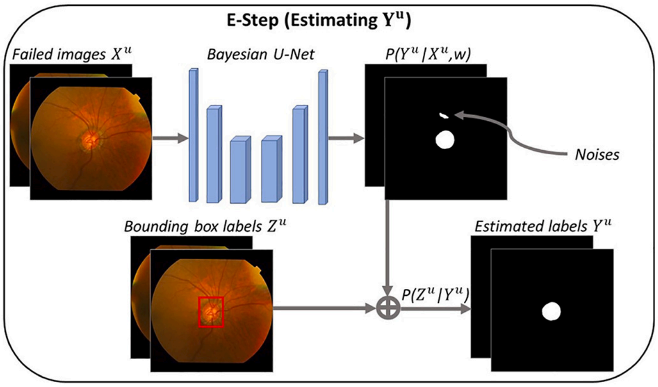

Fig. 3. E-Step. In the E-Step, the estimation of the Bayesian U-Net may contain background noises. Hence, the bounding box label Z^u is introduced to enforce the area outside the bounding box as non-optic disc area.

$$\phi(y^u_{nm}, z^u_n) = \left(\begin{array}{ll} 0 & \text{if } m^{th} \text{ pixel inside bounding box} \\ -1 & \text{if } m^{th} \text{ pixel outside bounding box} \end{array} \right. \quad (18)$$

Note that $f_m(y_{nm}{}^u | x_n{}^u, \omega') \in [0,1]$. By combining Eqs. (16) and (18), we observe that the area inside the bounding box is the prediction of the Bayesian U-Net ($f_m(y_{nm}{}^u | x_n{}^u, \omega')$), while the area outside bounding box is the background non optic disc area. Eventually, a clean mask estimation is generated by combining $p(y_n{}^u | x_n{}^u, \omega')$ and $p(z_n{}^u | y_n{}^u)$, as shown in Fig. 3.

3.4.2. M-step

In the M-step, the optimal estimation $\hat{y}^u_n$ of the disc mask obtained by the E-Step is fed into Eq. (15) and combined with the masks $y_n{}^l$ labeled by the Hough transform, to optimize the model parameters ω. The E-steps and M-steps iterate until convergence.

3.5. Inference

For the predictions, the variational distribution $q(\omega)$ is utilized as a replacement for the posterior $p(\omega | Y, Z^u, X)$ in Eq. (11), which becomes:

$$p(y^* | x^*) = \int p(y^* | x^*, \omega) q(\omega) d\omega. \quad (19)$$

Using the Monte Carlo dropout [12], Eq. (19) is approximated as:

$$p(y^* | x^*) \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* | x^*, \hat{\omega}_t). \quad (20)$$

It can be seen that the final prediction is obtained by averaging the outputs sampled from the trained Bayesian U-Net. Likewise, $\hat{\omega}_t$ are the weights of the Bayesian U-Net in the t^{th} sampling. In every sampling, $\hat{\omega}_t$ is different because the network weights ω are randomly dropped with dropout. The inference process is illustrated in Fig. 2(b).

4. Experiments

We evaluated the proposed method on three datasets using standard metrics, such as sensitivity, specificity, and IOU. We also qualitatively and quantitatively compared our method with the state-of-the-arts methods. Finally, we performed a series of ablation studies to evaluate the necessity and contribution of the key components of the proposed model.

4.1. Data

DRISHTI-GS [57]: DRISHTI-GS is a public dataset containing 50 training images and 51 testing images. All the images were collected at the Aravind Eye Hospital, Madurai. The DRISHTI-GS subjects were 40–80 years of age, with a similar number of males and females. The data acquisition protocol was centered on the optic disc, with a field-of-view of 30°, saved in the PNG uncompressed image format, with the 2045 × 1752 resolution.

RIM-ONE [58]: RIM-ONE is a public dataset, which consists of 159 images collected at the Hospital Universitario de Canarias. The number of male and female subjects is similar. All the images were captured by a non-mydriatic Kowa WX3D stereo fundus camera. The images were taken by centering on the optic nerve head, with a field-of-view of 34°, then saved in the JPEG format, with the 2144 × 1424 resolution.

REFUGE [59]: REFUGE is a public dataset containing 1200 annotated fundus images, split into three subsets: 400 training images, 400 validation images, and 400 testing images. The training images were captured by a Zeiss Visucam 500 fundus camera, with the resolution of 2124 × 2056 pixels. The validation and testing images were captured by a Canon CR-2 device, with the 1634 × 1634 pixels resolution. These images correspond to Chinese patients only and they were collected at multiple hospitals and eye clinics. The number of male and female patients is similar.

Before the experiments, all the images from the above three datasets were zero-padded to equalize their width and height, and then resized to the 640 × 640 resolution.

4.2. Training and implementation

The proposed model was trained in two phases. In the **first phase**, only images, the masks of which could be successfully identified by the Hough transform, were fed into the Bayesian U-Net for training. In the **second phase**, also the images, the masks of which could not be identified by the Hough transform were taken into account and used to train the model. The E-Step using the Bayesian U-Net trained in the first phase was deployed to estimate the optic disc masks for those images. Then, the M-Step optimized the Bayesian U-Net based on the complete training dataset, including the images estimated by the E-Step.

For both training phases, we used mini batches of size 5. The learning rate of the stochastic gradient descent optimizer was 0.002 and the optimal dropout rate was 0.2. For all three datasets, the first and second training phases required no more than 300 and 30 epochs, respectively. The number of samples from the Bayesian U-Net used for inference was 50.[1]

4.3. Metrics and evaluation

For disc segmentation evaluation purposes, we adopted accuracy (Acc), sensitivity (Sen), specificity (Spe), intersection-over-union (IOU), disc similarity coefficient (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD) metrics:

$$\begin{gathered} Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad Sen = \frac{TP}{TP + FN}, \\ Spe = \frac{TN}{TN + FP}, \quad IOU = \frac{TP}{TP + FP + FN}, \\ DSC = \frac{2 \times TP}{2 \times TP + FP + FN}, \end{gathered} \quad (21)$$

Here, TP and FP refer to the true and false positives, while TN and FN – to the true and false negatives, respectively. We computed HD and ASD using existing Python library[2] and code.[3] These image-based metrics were computed on a pixel basis for each testing image. The values reported below are averaged across all the testing images.

[1] The experiments were performed with one GPU. The computation time for one testing image was on average 2.26 s with a Tesla V100-SXM2-16GB.

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.directed_hausdorff.html.

[3] https://mlnotebook.github.io/post/surface-distance-function/.

We also used the receiver operating characteristic (ROC) curve to analyze the quality of the optic disc segmentation. In the ROC curve plots, the X and Y axes represent the FP and TP rate, respectively. Thus, top-left corner of the plot is an ideal point where the FP rate is 0 and the TP rate is 1. The area under the ROC curve (AUC) is another metric quantifying the performance of the optic disc segmentation. In general, larger AUC values mean a more accurate segmentation.

For the small-sized DRISHTI-GS and RIM-ONE datasets, we utilized 5-fold cross validation. As REFUGEE had already been split into training, validation, and testing subsets, we trained the models using the training set, and evaluated on the validation and testing sets.

4.4. Results

We evaluate the performance of our approach with respect to three aspects: (i) comparison with baseline methods, (ii) ablation study of the E-Step, and (iii) comparison with a non-Bayesian variant.

4.4.1. Comparison with baseline methods

We compared our method to four state-of-the-art baselines: fully-supervised AG-Net [54] and U-Net [53], weakly-supervised optic disc segmentation (WSODS) [10], weakly-supervised image segmentation (WSIS) [55], Global Constraint [56], and the Hough Transform. The results with respect to the seven metrics listed in Eq. (21) are given in Tables 1, 2 and 3.

Table 1
Performance comparison of our method and baselines (incl. two fully-supervised and four weakly-supervised methods) with the DRISHTI-GS dataset. (The best result is indicated in boldface)

Segmentation	Images	Acc	Sen	Spe	IOU	DSC	HD	ASD
Fully-supervised								
U-Net [53]	101	0.9982	0.9740	0.9988	0.9302	0.9632	3.2354	2.7083
AG-Net [54]	101	0.9987	0.9771	0.9993	0.9515	0.9749	3.0747	1.2598
Weakly-supervised								
Hough transform	101	0.9933	0.8998	0.9958	0.7780	0.8435	**3.2639**	34.6487
WSODS [10]	101	0.9962	0.9759	0.9968	0.8744	0.9303	4.5725	4.9554
WSIS [55]	101	0.9944	0.8873	0.9973	0.8099	0.8923	5.0390	8.9534
Global constraint [56]	101	0.9935	0.9326	0.9951	0.7909	0.8814	5.9280	7.4616
Proposed	101	**0.9970**	**0.9831**	**0.9974**	**0.8951**	**0.9436**	3.6149	**3.3944**

Table 2
Performance comparison of our method and baselines (incl. two fully-supervised and four weakly-supervised methods) with the RIM-ONE dataset. (The best result is indicated in boldface).

Segmentation	Images	Acc	Sen	Spe	IOU	DSC	HD	ASD
Fully-supervised								
U-Net [53]	159	0.9965	0.9482	0.9982	0.9008	0.9455	4.1203	4.5952
AG-Net [54]	159	0.9963	0.9470	0.9981	0.8936	0.9422	4.1015	3.6140
Weakly-supervised								
Hough transform	159	0.9857	0.8746	0.9899	0.6797	0.7845	5.0342	14.8896
WSODS [10]	159	0.9866	**0.9847**	0.9868	0.7031	0.8215	5.7206	12.6978
WSIS [55]	159	0.9871	0.8218	0.9906	0.6744	0.8008	6.7603	13.8021
Global constraint [56]	159	0.9882	0.8880	0.9916	0.7159	0.8275	6.0823	9.7739
Proposed	159	**0.9916**	0.9735	**0.9923**	**0.7831**	**0.8756**	**4.9929**	**8.8861**

Table 3
Performance comparison of our method and baselines (incl. two fully-supervised and four weakly-supervised methods) with the REFUGEE validation and test datasets. (The best result is indicated in boldface).

Segmentation	Dataset	Images	Acc	Sen	Spe	IOU	DSC	HD	ASD
Fully-supervised									
U-Net [53]	Validation	400	0.9975	0.9001	0.9992	0.8611	0.9235	3.7664	11.1502
	Test	400	0.9977	0.9227	0.9990	0.8712	0.9290	3.8932	13.4530
AG-Net [54]	Validation	400	0.9970	0.9650	0.9976	0.8463	0.9110	3.9116	11.4670
	Test	400	0.9968	0.9747	0.9972	0.8355	0.9036	3.8708	9.2733
Weakly-supervised									
Hough transform	Validation	400	0.9905	0.9310	0.9914	0.6224	0.7544	5.3554	15.1709
	Test	400	0.9904	0.9496	0.9910	0.6203	0.7549	5.4413	12.9623
WSODS [10]	Validation	400	0.9954	0.9222	0.9968	0.7754	0.8696	4.8411	11.9440
	Test	400	0.9953	0.9344	0.9964	0.7711	0.8643	4.8166	11.2019
WSIS [55]	Validation	400	0.9945	0.8931	0.9964	0.7339	0.8333	5.2291	12.5056
	Test	400	0.9943	0.9102	0.9957	0.7206	0.8277	5.4567	12.4187
Global constraint [56]	Validation	400	0.9936	0.8585	0.9958	0.6897	0.8077	5.5624	12.6585
	Test	400	0.9933	0.8797	0.9951	0.6829	0.8044	5.6899	14.9573
Proposed	Validation	400	**0.9967**	**0.9652**	**0.9973**	**0.8289**	**0.9034**	**4.0429**	**7.3427**
	Test	400	**0.9965**	**0.9727**	**0.9969**	**0.8185**	**0.8963**	**4.0494**	**6.9535**

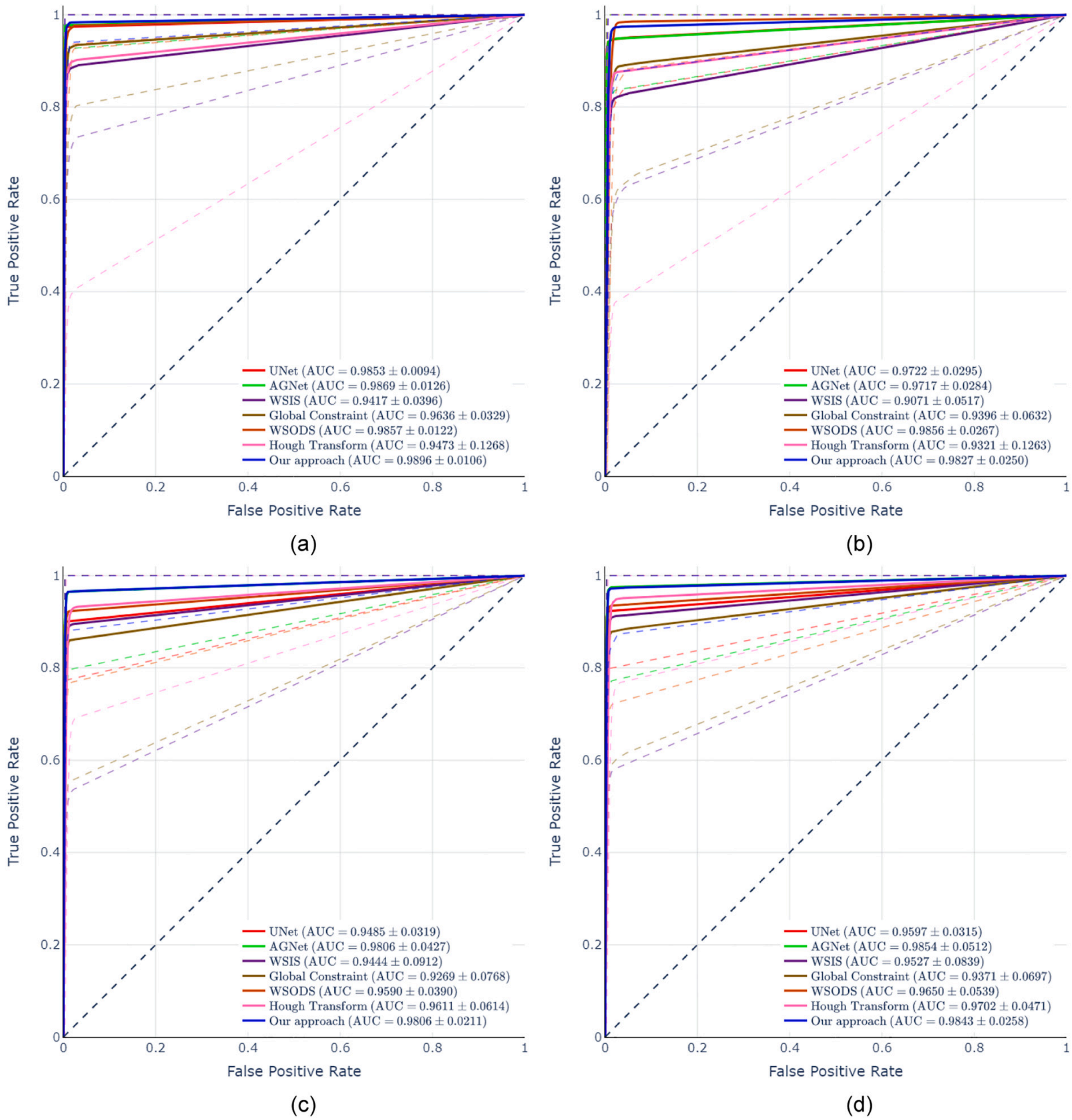

Fig. 4. ROC curves (solid lines) of all the compared methods with AUC values ± standard deviation on (a) DRISHTI-GS, (b) RIM-ONE, (c) REFUGEE Validation and (d) REFUGEE Test datasets. The dashed lines refer to the upper and lower bounds of the confidence interval.

We observe that our method outperforms the Hough transform, the weakly supervised WSODS, WSIS and Global Constraint with respect to all the metrics on the REFUGEE dataset, all but HD on DRISHTI-GS, and all but Sen on RIM-ONE. Recall that the Hough transform may fail to label optic disc masks in the fundus images, which dramatically lowers its performance. In addition, WSODS requires a rough disc segmentation as a pseudo ground truth for training, which leads to a lower performance than our method. In similar to pseudo ground truth, the bounding box labels utilized by WSIS and Global Constraint are not reliable for learning the segmentation task and thus result in inferior performance to our method. The superiority of our model over these methods lies in the following factors: 1) our Hough Transform based labels are more accurate than the bounding box labels and pseudo ground truths (comparisons are shown in Fig. 1). As can be seen in Fig. 1, a Hough Transform based label is a circular mask that conforms to the shape of the optic disc; 2) Despite being more reliable, Hough Transform labels are more noisy and less accurate than the ground truth, and thus tend to introduce uncertainty in the segmentation learning. Our Bayesian U-Net considers such an uncertainty by making U-Net a probabilistic model, from which several segmentation probability maps can be sampled at the inference stage. The final segmentation is generated by averaging these probability maps and is shown to be more accurate than those produced by

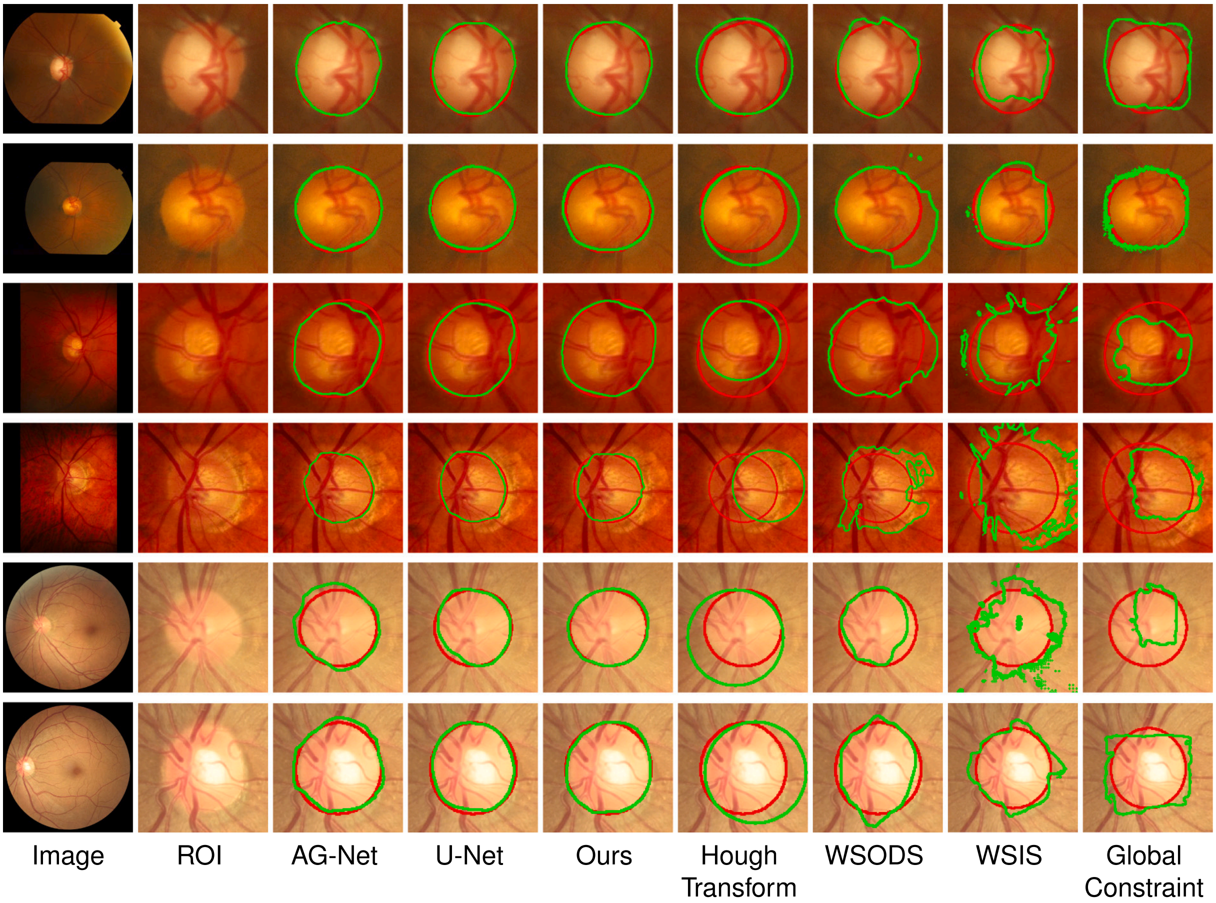

Fig. 5. Comparison of the proposed method with six baseline methods on two DRISHTI-GS, two RIM-ONE and two REFUGEE images (top to bottom). The red and green contours indicate the boundaries of the ground truth and predicted optic discs. The second column illustrates the enlarged ROI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

weakly supervised methods. However, the performance of our method is inferior to the benchmarked fully-supervised AG-Net and U-Net methods, i.e., our method generally achieved lower Acc, Spe, IOU, DSC and higher HD, ASD scores than the fully-supervised methods. The ROC curves of these methods are shown in Fig. 4, where our method achieves an AUC of 0.9896, 0.9827, 0.9806 and 0.9843 on the DRISHTI-GS, RIM-ONE and REFUGEE Validation/Test datasets, respectively. It can be seen that the ROC curves of our method are above the ROC curves of the Hough Transform, WSODS, WSIS, and Global Constraint methods, except for the RIM-ONE dataset.

Fig. 5 exemplifies the results of six disc segmentations across the evaluated methods. As can be seen, the segmentation produced by our method is more accurate than that of the weakly-supervised WSODS, WSIS, Global Constraint and Hough Transform, and visually similar to that of the fully-supervised methods. Notably, unlike weakly-supervised methods, the fully-supervised methods require annotations of the optic discs as an input. Such annotations are produced by human experts, imposing a strong constraint on practical disc segmentation applications. In practice, many clinical tasks require the cropped rectangular disc region rather than the exact disc for examination by a clinician. Hence, despite being inferior to the fully-supervised methods, our weak label based method may meet practical needs, while not requiring the expensive manual annotation.

Table 4
p-Values produced by a t-test comparing our method vs. four weakly-supervised methods with the DRISHTI-GS, RIM-ONE and REFUGEE datasets. (N.S. refers to *not significant*).

	DRISHTI-GS	RIM-ONE	REFUGEE	
			Validation	Test
Ours vs Hough transform	0.0097	0.0060	<0.001	<0.001
Ours vs WSODS	0.0448	<0.001	<0.001	<0.001
Ours vs WSIS	0.0024	<0.001	<0.001	<0.001
Ours vs global constraint	<0.001	N.S.	<0.001	<0.001

We also report the results of a *t*-test comparing our method with the other weakly-supervised methods: WSODS, WSIS, Global Constraint and Hough Transform. Here, the *t*-test is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. We assess statistical significance with respect to ASD. The other metrics evaluate performance on the whole mask, for which the central areas of the optic disc are relatively easy to predict (see examples in

Fig. 5). Therefore, evaluating using the whole mask cannot clearly illustrate the differences between the methods. In contrast, the ASD metric better demonstrates the differences by comparing the outermost contours of two masks. As can be seen in Table 4, all the *p*-values but ours vs global constraint on RIM-ONE are smaller than the significance level 0.05, which indicates that our findings are statistically significant, reliable and, not due to chance.

4.4.2. *Ablation study*

To demonstrate how the E-Step of the E-M algorithm affects the disc segmentation performance, we designed two scenarios. We denote the subset of images, for which the mask was successfully identified by the Hough Transform as *detected*, while the images, for which the mask was not identified – as *failed*. In the first **M-Step**$_d$ scenario, the Bayesian U-Net is trained with the detected images only. In the second **M-Step**$_{d+f}$ scenario, in addition to the images, where the mask was detected, also the failed images are used for training the Bayesian U-Net. Note that in both scenarios, the Bayesian U-Net is trained and optimized directly, without the E-step. In DRISHTI-GS, for 87 images the mask was successfully detected and for 14 images this failed; while in RIM-ONE there were 144 and 15 detected and failed images, respectively; lastly, REFUGEE contained 389 detected and 11 failed images.

We evaluate the performance of our approach with the above two scenarios. The results obtained for the DRISHTI-GS, RIM-ONE and REFUGEE datasets are reported in Table 5. All the metrics of our model, trained with both the E- and M-steps are better than those without the E-step for DRISHTI-GS and RIM-ONE. For REFUGEE, Acc, Sen, IOU, DSC and HD of the proposed model trained with E- and M-steps are better than their counterparts without them. The reason might be that E-step accurately estimates the segmentation masks of the images. Following this, the re-estimated masks augment the training data and enhance the segmentation accuracy. The ROC curves of our approach are shown in Fig. 6. The AUC values on DRISHTI-GS, RIM-ONE and REFUGEE Validation/Test datasets are 0.9896, 0.9827, 0.9806 and 0.9843, respectively, all higher than those of M-Step$_d$ and M-Step$_{d+f}$. In addition, Fig. 7 exemplifies the segmentation results of these methods. As can be seen, our method with both the E- and M-Steps produces more accurate predictions.

4.4.3. *Bayesian vs. non Bayesian*

The Bayesian variant of U-Net differs from the traditional U-Net because a prior that is put on the network weights, thus, the objective

Table 5

Ablation study comparing the performance of training with/without the E-Step on the DRISHTI-GS, RIM-ONE, and REFUGEE datasets. The **detected** column denotes the number of images, for which the mask was successfully detected by the Hough Transform, while **failed** refers to the number of images, for which the mask was not detected. (The best result is indicated in boldface).

(a) Evaluation on DRISHTI-GS dataset

	Detected ($n = 87$)	Failed ($n = 14$)	E-step	M-step	Metrics						
					Acc	Sen	Spe	IOU	DSC	HD	ASD
M-step$_d$	✓			✓	0.9968	0.9759	0.9974	0.8864	0.9384	3.7830	5.3997
M-step$_{d+f}$	✓	✓		✓	0.9967	0.9714	0.9974	0.8830	0.9365	3.8672	4.0233
Proposed	✓	✓	✓	✓	**0.9970**	**0.9831**	**0.9974**	**0.8951**	**0.9436**	**3.6149**	**3.3944**

(b) Evaluation on RIM-ONE dataset

	Detected ($n = 144$)	Failed ($n = 15$)	E-step	M-step	Metrics						
					Acc	Sen	Spe	IOU	DSC	HD	ASD
M-step$_d$	✓			✓	0.9914	0.9675	0.9923	0.7803	0.8729	5.0672	9.5819
M-step$_{d+f}$	✓	✓		✓	0.9909	0.9651	0.9919	0.7691	0.8666	5.1595	9.5397
Proposed	✓	✓	✓	✓	**0.9916**	**0.9735**	**0.9923**	**0.7831**	**0.8756**	**4.9929**	**8.8861**

(c) Evaluation on REFUGEE dataset

	Database	Detected ($n = 389$)	Failed ($n = 11$)	E-step	M-step	Metrics						
						Acc	Sen	Spe	IOU	DSC	HD	ASD
M-step$_d$	Validation	✓			✓	0.9962	0.9639	0.9969	0.8094	0.8908	4.1894	7.7410
	Test	✓			✓	0.9958	0.9727	0.9962	0.7938	0.8797	4.1962	10.2126
M-step$_{d+f}$	Validation	✓	✓		✓	0.9965	0.9455	**0.9975**	0.8199	0.8975	4.1783	**6.5323**
	Test	✓	✓		✓	0.9964	0.9613	**0.9971**	0.8125	0.8932	4.1655	**6.9445**
Proposed	Validation	✓	✓	✓	✓	**0.9967**	**0.9652**	0.9973	**0.8289**	**0.9034**	**4.0429**	7.3427
	Test	✓	✓	✓	✓	**0.9965**	**0.9727**	0.9969	**0.8185**	**0.8963**	**4.0494**	6.9535

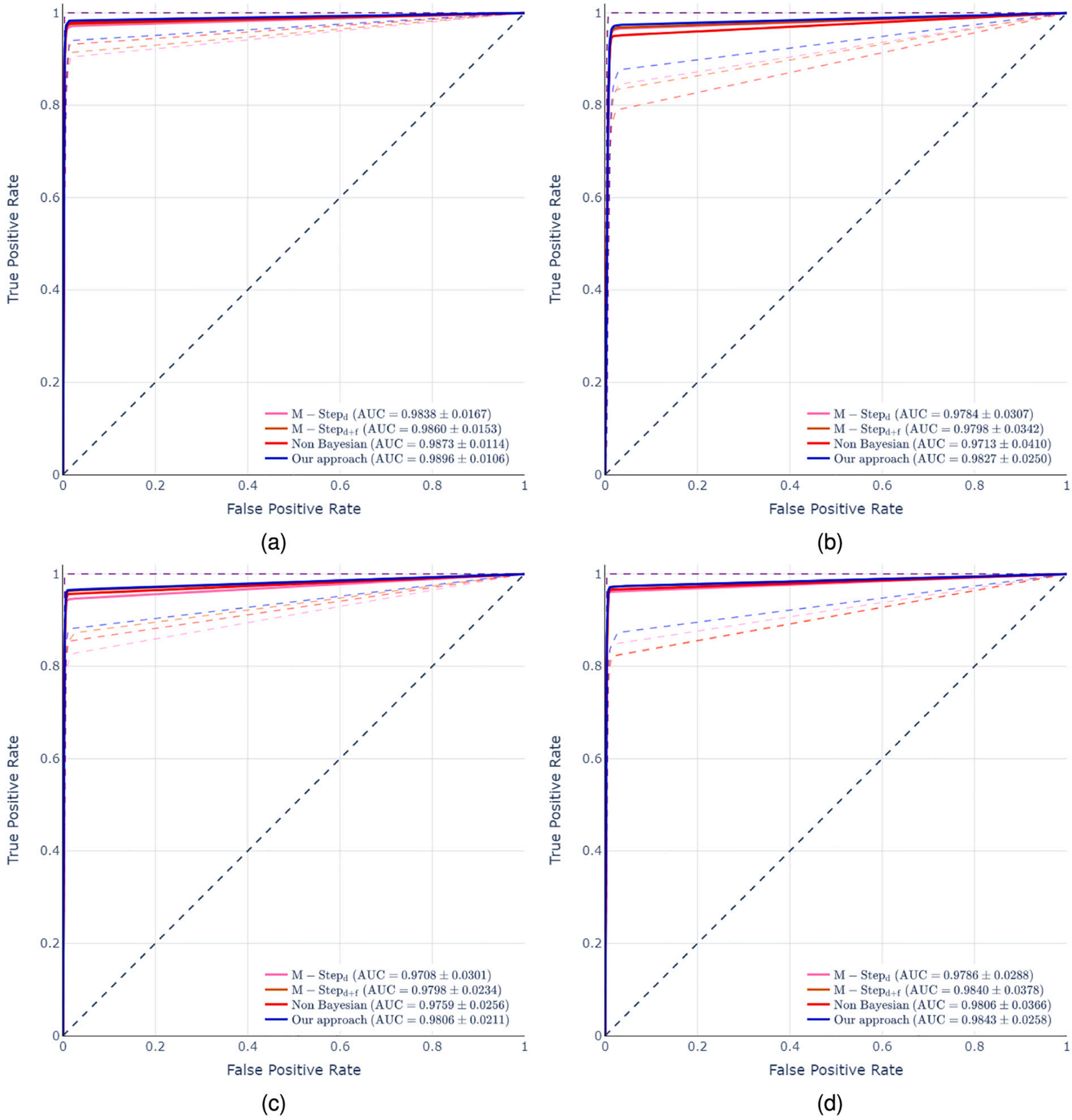

Fig. 6. Ablation study ROC curves (solid lines) of all the compared methods with AUC values ± standard deviation on (a) DRISHTI-GS, (b) RIM-ONE, (c) REFUGEE Validation and (d) REFUGEE Test datasets. The dashed lines refer to the upper and lower bounds of the confidence interval.

function is defined differently. As described in previous section, we approximate the objective function using dropout, which has been shown to be able to model uncertainty, such as in the Bayesian models [12]. Hence, the study of the Bayesian vs. Non-Bayesian U-Net is equivalent to the study of Dropout U-Net vs. U-Net without dropout (traditional U-Net).

As shown in Table 6 and Fig. 6, U-Net without dropout (traditional U-Net) is inferior to the Dropout U-Net (Bayesian U-Net) across most metrics. It is also evident in Fig. 7 that the Non-Bayesian traditional U-Net predicts less accurate masks than the Bayesian one. The latter enhances the segmentation accuracy due to the fact that the dropout itself reduces overfitting in the training.

We also compare the calculation time between the Bayesian and Non-Bayesian U-Nets (shown in Table 7). As can be seen, the Bayesian U-Net takes longer to predict one image, on average. This is primarily because Bayesian U-Net is a probabilistic model; it samples at the inference stage to generate the segmentation result, which takes longer than simply predicting the result from an ordinary network like the Non-Bayesian U-Net. However, practical clinical applications do not require the segmentation to be done in real time and the average calculation

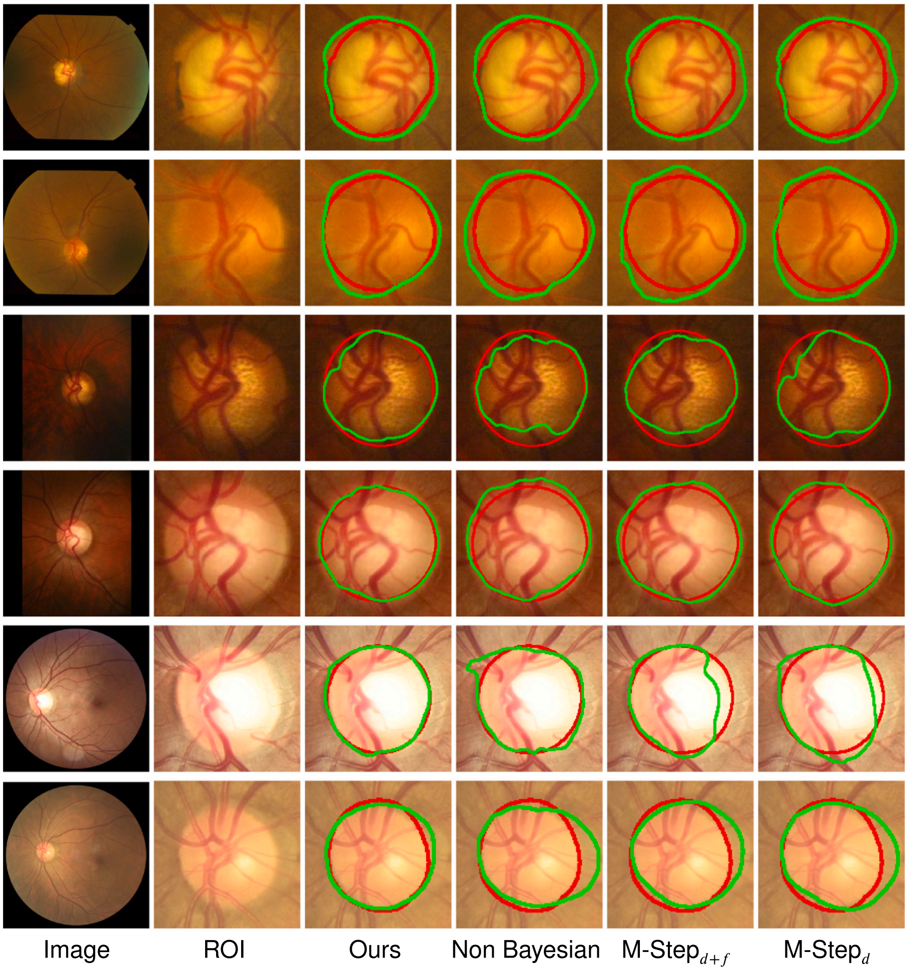

Fig. 7. Comparison of different variants of the proposed method on two DRISHTI-GS, two RIM-ONE and two REFUGEE images (top to bottom). The red and green contours indicate the boundaries of the ground truth and predicted optic discs. The second column illustrates the enlarged ROI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Ablation study comparing the performance of the Bayesian and non-Bayesian U-Nets on the DRISHTI-GS, RIM-ONE, and REFUGEE datasets. (The best result is indicated in boldface).

(a) Evaluation on DRISHTI-GS dataset

	Acc	Sen	Spe	IOU	DSC	HD	ASD
U-Net	0.9967	0.9786	0.9972	0.8870	0.9389	3.8311	4.4243
Bayesian U-Net	**0.9970**	**0.9831**	**0.9974**	**0.8951**	**0.9436**	**3.6149**	**3.3944**

(b) Evaluation on RIM-ONE dataset

	Acc	Sen	Spe	IOU	DSC	HD	ASD
U-Net	0.9912	0.9502	**0.9928**	0.7751	0.8702	5.2735	8.9966
Bayesian U-Net	**0.9916**	**0.9735**	0.9923	**0.7831**	**0.8756**	**4.9929**	**8.8861**

(c) Evaluation on REFUGEE dataset

	Database	Acc	Sen	Spe	IOU	DSC	HD	ASD
U-Net	Validation	0.9960	0.9561	0.9968	0.7984	0.8846	4.5228	11.1744
	Test	0.9957	0.9658	0.9962	0.7847	0.8749	4.5635	13.1779
Bayesian U-Net	Validation	**0.9967**	**0.9652**	**0.9973**	**0.8289**	**0.9034**	**4.0429**	**7.3427**
	Test	**0.9965**	**0.9727**	**0.9969**	**0.8185**	**0.8963**	**4.0494**	**6.9535**

Table 7
Average prediction times of the Bayesian and non-Bayesian U-Nets (in seconds) for the DRISHTI-GS, RIM-ONE, and REFUGEE datasets.

	DRISHTI-GS	RIM-ONE	REFUGEE	
			Validation	Test
U-Net	0.046 s	0.044 s	0.040 s	0.040 s
Bayesian U-Net	1.828 s	1.882 s	1.940 s	1.939 s

time under 2 s across the four datasets is sufficiently fast. Therefore, the longer calculation time of the Bayesian U-Net is traded off by its higher segmentation accuracy.

5. Discussion

Optic disc segmentation is a vital process for screening of eye diseases. Therefore, it is extensively used in clinical practice and attracts numerous research attempts. The existing methods primarily focus on fully- and semi-supervised learning of the optic disc segmentation. For these methods, pixel-level annotations are required for training purposes. However, obtaining pixel-level annotations necessitates manual annotation, which is time consuming and renders non-scalable in practice.

Since fully- and semi-supervised methods require laborious pixel-level annotations, it is worth turning the attention to weakly-supervised learning of the optic disc segmentation. To the best of our knowledge, the work of Lu et al. is so far the only one focusing on the weakly-supervised segmentation of optic disc [10]. There, the learning of optic disc segmentation relied on weak labels, such as image-level and bounding box labels. Compared to pixel-level annotations, such weak labels are easier to annotate and obtain. Instead of annotating the bounding box labels for all images, our proposed weak label based approach exploits the Hough transform based labels as pseudo masks for most training images. For a small number of images that cannot be detected by the Hough transform, we annotate the bounding box label to estimate the pseudo masks. Evidently, our approach further reduces annotation time. Besides, by exploiting dropout as a Bayesian approximation, our Bayesian network does not have more weights to learn than the ordinary U-Net. As such, the training is effective and the predictions at the inference stage are fast. The evaluation on three public datasets proves that our method is superior to several baseline methods.

It is important to note that at this stage, our approach is unable to segment the optic cup. This is mainly due to the fact that the optic cup and the optic disc areas share similar textures, color, and brightness. Thus, the Hough transform cannot accurately detect the optic cup area and generate the labels used by our method for the optic cup segmentation. To address this issue, the semi-supervised learning exploiting partial ground truth optic cup masks offer a solid alternative. Likewise, these ground truth masks may be used by our model in the first training phase. In the second training phase, the images that do not have masks can be re-estimated by E-M algorithm. Then, all the masks, including the ground truth and re-estimated ones, can be utilized for the optic cup segmentation learning. In the future, we intend to improve our model so that it can simultaneously segment the optic disc and the optic cup.

6. Conclusions

In this work, we proposed and evaluated a weak label based Bayesian U-Net to segment the optic disc in fundus images. Notably, our method does not rely on manually annotated optic disc masks, but only requires the Hough transform based annotations. This method has the potential to considerably reduce the time required for annotating optic disc masks and simplifies the segmentation pipeline. Our method was shown to outperform baseline weakly-supervised methods, although it was inferior to two fully-supervised baselines. However, analysis of several sample segmentations showed that the performance of the proposed method may be sufficient for practical clinical needs. We also report an ablation study and evaluation of a non-Bayesian variant of the method.

The segmentation accuracy achieved by our approach is still inferior to the models trained with ground truth labels manually annotated by experts. Despite this, our method demonstrates high accuracy that can satisfy clinical needs when screening eye diseases. The state-of-the-art screening tools often leverage deep learning models, such as CNN. Rather than the exact optic disc area, these models take as an input the ROI containing the optic disc area. The segmentation of the optic disc generated by our model is accurate enough to crop such an ROI for eye disease screening. As a probabilistic variant of U-Net, the proposed method has a strong potential to be used in other medical image segmentation tasks, which we intend to explore in the 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.

References

[1] Chen X, Xu Y, Yan S, Wing D, Wong T, Liu J. Automatic feature learning for glaucoma detection based on deep learning. In: Proc. MICCAI; 2015. p. 669–77.

[2] Orlando JI, Prokofyeva E, Fresno MD, Blaschko MB. Convolutional neural network transfer for automated glaucoma identification. In: Proc. SPIE; 2016. p. 1–10.

[3] Fu H, Cheng J, Xu Y, Zhang C, Wong DWK, Liu J, Cao X. Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans Med Imaging 2018;37(11):2493–501.

[4] Liu S, Graham SL, Schulz A, Kalloniatis M, Zangerl B, Cai W, Gao Y, Chua B, Arvind H, Grigg J, Chu D, Klistorner A, You Y. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. Ophthalmol Glaucoma 2018;1(1):15–22.

[5] Haleem MS, Han L, Li B, Nisbet A, Hemert JV, Verhoek M. Automatic extraction of the optic disc boundary for detecting retinal diseases. In: Proc. of the IASTED International Conference on Computer Graphics and Imaging; 2013. p. 40–7.

[6] Patton N, Aslam TM, MacGillivray T, Deary IJ, Dhillon B, Eikelboom RH, Yogesan K, Constable IJ. Retinal image analysis: concepts, applications and potential. Prog Retin Eye Res 2006;25(1):99–127.

[7] Zhou W, Wu C, Chen D, Yi Y, Du W. Automatic microaneurysm detection using the sparse principal component analysis-based unsupervised classification method. IEEE Access 2017;5:2563–72.

[8] Osareh A, Mirmehdi M, Thomas B, Markham R. Classification and localisation of diabetic-related eye disease. In: Proc. ECCV; 2002. p. 502–16.

[9] Saha O, Sathish R, Sheet D. Learning with multitask adversaries using weakly labelled data for semantic segmentation in retinal images. In: Proc. of Machine Learning Research; 2019. p. 414–26.

[10] Lu Z, Chen D, Xue D, Zhang S. Weakly supervised semantic segmentation for optic disc of fundus image. J Electron Imaging 2019;28(3):033012.

[11] Sekhar S, Al-Nuaimy W, Nandi AK. Automated localisation of retinal optic disk using hough transform. In: IEEE International Symposium on Biomedical Imaging; 2008. p. 1577–80.

[12] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proc. ICML; 2016. p. 1050–9.

[13] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: 34th International Conference on Machine Learning; 2017. p. 1183–92.

[14] Kendall A, Badrinarayanan V, Cipolla R. Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proceedings of the British Machine Vision Conference (BMVC); 2017. p. 1–12.

[15] Orlando JI, Seebock P, Bogunovic H, Klimscha S, Grechenig C, Waldstein S, Gerendas BS, Schmidt-Erfurth U. U2-net: a Bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In: IEEE 16th International Symposium on Biomedical Imaging; 2019. p. 1441–5.

[16] Graves A. Practical variational inference for neural networks. In: Proceedings of the 24th International Conference on Neural Information Processing Systems; 2011. p. 2348–56.

[17] Mnih A, Gregor K. Neural variational inference and learning in belief networks. In: Proceedings of the 31st International Conference on Machine Learning; 2014. p. 1791–9.

[18] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15 (1).

[19] Youssif AAA, Ghalwash AZ, Ghoneim AASA. Optic disc detection from normalized digital fundus images by means of a vessels' direction matched filter. IEEE Trans Med Imaging 2008;27(1):11–8.

[20] Zou BJ, Zhang SJ, Zhu CZ. Automatic localization and segmentation of optic disk in color fundus images. Opt Precis Eng 2015;23(4):1187–95.

[21] Lowell J, Hunter A, Steel D, Basu A, Ryder R, F. E., Kennedy L. Optic nerve head segmentation. IEEE Trans Med Imaging 2004;23(2):256–64.

[22] Joshi GD, Sivaswamy J, Krishnadas S. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. IEEE Trans Med Imaging 2011;30(6):1192–205.

[23] Zheng SH, Chen J, Pan L, Yu L. Optic disc detection on retinal images based on directional local contrast. Chin J Biomed Eng 2014;33(3):289–96.

[24] Lu S. Accurate and efficient optic disc detection and segmentation by a circular transformation. IEEE Trans Med Imaging 2011;30(12):2126–33.

[25] Priyadharsini R, Beulah A, Sharmila TS. Optic disc and cup segmentation in fundus retinal images using feature detection and morphological techniques. Curr Sci 2018;115(4):748–52.

[26] Aquino A, Gegúndez-Arias ME, Marín D. Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques. IEEE Trans Med Imaging 2010;29(11):1860–9.

[27] Chakravarty A, Sivaswamy J. Joint optic disc and cup boundary extraction from monocular fundus images. Comput Methods Prog. Biomed. 2017;147:51–61.

[28] Wong DWK, Liu J, Tan NM, Yin F, Lee BH, Wong TY. Learning-based approach for the automatic detection of the optic disc in digital retinal fundus photographs. In: Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc; 2010. p. 5355–8.

[29] Mittapalli PS, Kande GB. Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma. biomed. Signal processControl 2016;24:34–46.

[30] Cheng J, Liu J, Xu Y, Yin F, Wong DWK, Tan NM, Tao D, Cheng C, Aung T, Wong TY. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Trans Med Imaging 2013;32(6):1019–32.

[31] Cheng J, Liu J, Xu Y, Yin F, Wong DWK, Tan NM, Tao D, Cheng C, Aung T, Wong TY. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Trans Med Imaging 2013;32(6):1019–32.

[32] Abràmoff MD, Alward WLM, Greenlee EC, Shuba L, Kim CY, Fingert JH, Kwon YH. Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features. Invest Ophthalmol Vis Sci 2007;48(4):1665–73.

[33] Juneja M, Singh S, Agarwal N, Bali S, Gupta S, Thakur N, Jindal P. Automated detection of glaucoma using deep learning convolution network (g-net). Multimed Tools Appl 2019;79:1–23.

[34] Al-Bander B, Williams BM, Al-Nuaimy W, Al-Taee MA, Pratt H, Zheng Y. Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. Symmetry 2018;10(4):87–91.

[35] Yuan X, Zhou L, Yu S, Li M, Wang X, Zheng X. A multi-scale convolutional neural network with context for joint segmentation of optic disc and cup. Artif Intell Med 2021;113:102035.

[36] Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans Med Imaging 2018;37(7):1597–605.

[37] Sevastopolsky A. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. Pattern Recognit Image Anal 2017;27(3):618–24.

[38] Jin B, Liu P, Wang P, Shi L, Zhao J. Optic disc segmentation using attention-based u-net and the improved cross-entropy convolutional neural network. Entropy 2020;22(8):844.

[39] Tan JH, Acharya UR, Bhandary SV, Chua KC, Sivaprasad S. Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. J Comput Sci 2017;20:70–9.

[40] Zilly J, Buhmann JM, Mahapatra D. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Comput Med Imaging Graph 2017;55:28–41.

[41] Wang S, Yu L, Yang X, Fu C-W, Heng P-A. Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans Med Imaging 2019; 38(11):2485–95.

[42] Wang S, Yu K, Li L, Yang X, Fu C-W, Heng P-A. Boundary and entropy-driven adversarial learning for fundus image segmentation. In: MICCAI; 2019. p. 102–10.

[43] Wang S, Yu L, Li K, Yang X, Fu C-W, Heng P-A. Dofe: domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans Med Imaging 2020;39(12):4237–48.

[44] Bian X, Wang C, Liu W, Lin X. Unsupervised optic disc segmentation for cross domain fundus image based on structure consistency constraint. In: International Conference on Image and Graphics; 2019. p. 724–34.

[45] Norouzifard M, Dehkordi AA, Dehkordi MN, Gholamhosseini H, Klette R. Unsupervised optic cup and optic disk segmentation for glaucoma detection by icica. In: 15th International Symposium on Pervasive Systems, Algorithms and Networks; 2018. p. 209–14.

[46] Rajchl M, Lee MC, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Damodaram M, Rutherford MA, Hajnal JV, Kainz B, Rueckert D. Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans Med Imaging 2017;36(2):674–83.

[47] G. Yang C. Wang J. Yang Y. Chen L. Tang P. Shao J.-L. Dillenseger H. Shu L. Luo Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal cta images, BMC Med Imaging 20 (37).

[48] Kervadec H, Dolz J, Tang M, Granger E, Boykov Y, Ayed IB. Constrained-cnn losses for weakly supervised segmentation. Med Image Anal 2019;54:88–99.

[49] Rajchl M, Lee MC, Schrans F, Davidson A, Passerat-Palmbach J, Tarroni G, Alansary A, Oktay O, Kainz B, Rueckert D. Learning under distributed weak supervision. 2016. arXiv: 1606.01100.

[50] Girum KB, Créhange G, Hussain R, Lalande A. Fast interactive medical image segmentation with weakly supervised deep learning method. Int J Comput Assist Radiol Surg 2020;15:1437–44.

[51] Illingworth J, Kittler J. The adaptive hough transform. In: IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI. 5; 1987. p. 690–8.

[52] Green B. Canny edge detection tutorial. 2002.

[53] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proc. MICCAI; 2015. p. 234–41.

[54] Zhang S, Fu H, Yan Y, Zhang Y, Wu Q, Yang M, Tan M, Xu Y. Attention guided network for retinal image segmentation. In: Proc. MICCAI; 2019. p. 797–805.

[55] Wang J, Xia B. Bounding box tightness prior for weakly supervised image segmentation. In: Medical Image Computing and Computer Assisted Intervention; 2021. p. 526–36.

[56] Kervadec H, Dolz J, Wang S, Granger E, Ayed IB. Bounding boxes for weakly supervised segmentation: global constraints get close to full supervision. In: Medical Image with Deep Learning; 2020. p. 365–80.

[57] Sivaswamy J, Krishnadas SR, Joshi GD, Jain M, Tabish AUS. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: Proc. IEEE 11th International Symposium on Biomedical Imaging (ISBI); 2014. p. 53–6.

[58] Fumero F, Sigut J, Alayón S, González-Hernández M, de la Rosa MGonzález. Interactive tool and database for optic disc and cup segmentation of stereo and monocular retinal fundus images. In: Short Papers Proceedings - WSCG; 2015. p. 91–7.

[59] Orlando JI, Fu H, Breda JB, Keer KV, Bathula DR, Diaz-Pinto A, Fang R, Heng P-A, Kim J, Lee J, Lee J, Li X, Liu P, Lu S, Murugesan B, Naranjo V, Phaye SSR, Shankaranarayana SM, Sikka A, Son J, Hengel AVD, Wang S, Wu J, Wu Z, Xu G, Xu Y, Yin P, Li F, Zhang X, Xu Y, Bogunovića H. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 2020;59:101570.