Biomedical Signal Processing and Control xxx (xxxx) xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Diabetic retinopathy lesion segmentation using deep multi-scale framework

Tianjiao Guo [a,b,c], Jie Yang [b,*], Qi Yu [d,e,f,**]

[a] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[b] Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[c] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[d] Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[e] National Clinical Research Center for Eye Diseases, Shanghai, China
[f] Shanghai Clinical Research Center for Eye Diseases, Shanghai, China

ARTICLE INFO

Keywords:
Retinal image
Convolutional neural network (CNN)
Segmentation
Diabetic retinopathy (DR)

ABSTRACT

The segmentation of diabetic retinopathy (DR) lesions is important for large-scale screening using color fundus photography (CFP) images. The difficulty of this task is that the DR lesions have various sizes, shapes, and intensities. Traditional handcrafted feature-based approaches are unsatisfactory, and recent deep-learning-based approaches ignore the features of DR lesions. In this paper, we propose a segmentation approach that segments four typical DR lesions simultaneously based on convolutional neural networks (CNN). The raw CFP image is first pre-processed and resized to different sizes. A set of fully convolutional neural networks (FCN) with various input sizes are then trained to extract the lesions from different scales. An auxiliary CNN is then introduced to fuse the output from these FCN and refine the segmentation result. We conducted our experiments on one local dataset and four public datasets including IDRiD, DDR, E-ophtha, and DIARETDB1. The results of the area under the precision–recall-curve (AUPR) and the dice similarity coefficient (DSC) show that our approach achieves competitive performance. The improvement in performance indicates that this approach is beneficial to DR lesion segmentation and has potential in other segmentation tasks.

1. Introduction

The number of people with diabetes worldwide will be more than 590 million by 2035 according to the predictions of the World Health Organization (WHO) and the International Diabetes Federation (IDF) [1]. The presence of diabetes can cause chronic damage in the human body such as diabetic retinopathy (DR), which is one of the leading causes of blindness in the labor force.

In the early stage of DR, the dilation of capillaries can lead to microaneurysms (MA). The leaking of lipoproteins can lead to hard exudates (EX). The arteriole occlusion can lead to the ischemia of the retinal nerve fiber layer and further leads to the cotton wool spots (CW, which is also called soft exudates and is abbreviated as SE in some research). The broken of abnormal blood vessels and microaneurysms leads to hemorrhages (HE). MA, EX, CW, and HE are four typical Non-proliferative Diabetic Retinopathy (NPDR) lesions. If not treated timely, the NPDR will develop into Proliferative Diabetic Retinopathy (PDR) [2,3], which is more severe and will lead to irreversible damage to vision. Early screening of these NPDR lesions is the most effective method to slow the progress of DR and prevent vision loss.

As one of the screening modalities, color fundus photography (CFP) is commonly utilized since it is non-invasive and cost-effective. In a CFP image, as shown in Fig. 1, MA and HE are similar in color. The former usually has the shape of a dot while the latter is usually blocky. EX and CW are both bright lesions. The former is usually yellowish-white, well-defined, and waxy while the latter is cotton-wool-like. In clinical practice, human experts usually diagnose or grade the DR by observing the area and quantity of lesions. Therefore, the segmentation of DR lesions is of great significance for DR diagnosis.

With the development of computer science, deep learning methods represented by a convolutional neural network (CNN) are widely used in medical image [6–9]. CNN-based methods have shown promising results in both accuracy and efficiency compared with humans. In the field of DR lesion segmentation, a number of researches based on CNN models have been made, and a number of databases with pixel-wise lesion annotations have been released to the public in recent years. They mainly pre-process the raw CFP to fix the field-of-view (FoV) and the image size, then train several networks to segment different lesions [10–14]. These approaches are straightforward and intuitive.

* Corresponding author.
** Correspondence to: No. 100 Haining Road, Shanghai, 200080, China.
E-mail addresses: 292725651@sjtu.edu.cn (T. Guo), jieyang@sjtu.edu.cn (J. Yang), yu.qi@sjtu.edu.cn (Q. Yu).

https://doi.org/10.1016/j.bspc.2023.105050
Received 10 February 2023; Received in revised form 22 April 2023; Accepted 23 May 2023

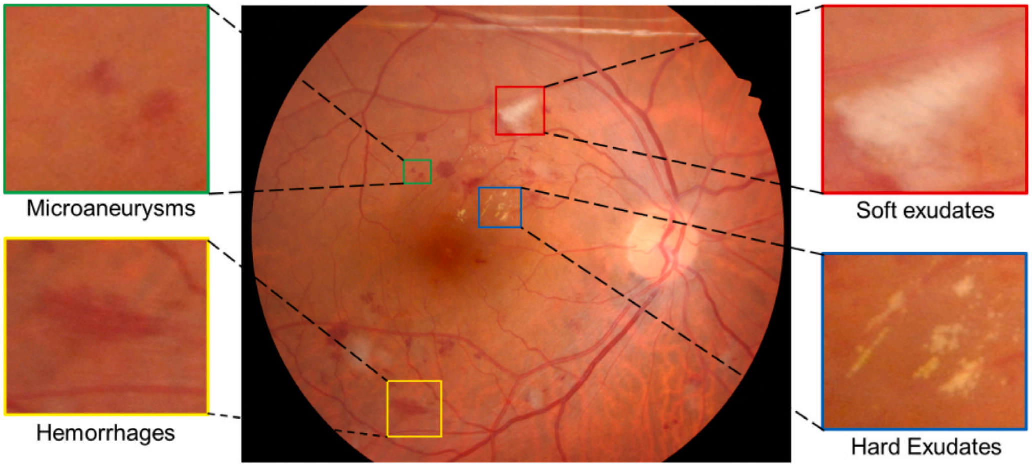

Fig. 1. The typical DR lesions in CFP image [4,5]. Microaneurysm (MA) is usually like a red small dot. Hemorrhage (HE) is usually a blocky dark-red lesion. Hard exudates (EX) is yellowish-white, well-defined, and waxy. And soft exudates (cotton wool spots or CW in this paper) is usually cotton-wool-like.

However, existing works have their limitations. First, single and fixed image size is unreasonable. The sizes of lesions are very different from each other. For example, the MA is very small compared with CW and may disappear when the image size is reduced. Assigning the feature extractors of the same scale to different lesions in CNN is obviously flawed. Second, existing works mostly treat each lesion separately. However, segmenting all these lesions simultaneously by a single network, i.e. multi-task segmentation, can make full use of the image to learn discriminative features, which can further increase the accuracy.

Besides, existing works only conduct their experiments on DR screening photographs, which are captured with 45~50-degree FoV cameras. The experiments on the CFPs with 30-degree FoV cameras, which are widely utilized in the protocol of the Early Treatment Diabetic Retinopathy Study (ETDRS) [15], have not been conducted.

To explicitly solve the issues mentioned above, in this paper, we propose a framework for DR lesion segmentation. First, to deal with the issue of variant lesion scale, we design multi-scale segmentation and fusion modules. The features of the lesion are extracted from different scales and weighted to fuse the information of various receptive fields. Second, to make full use of the features, we introduce the idea of multi-task learning to segment all these lesions simultaneously, which is more accurate and elegant. To demonstrate the effectiveness of our framework, we conduct our experiments on several datasets. We annotate some 30-degree-FoV CFP images captured from our local hospital. We also refine and publicly share the annotations from the DIARETDB1 [16] dataset. The results on these datasets show that our framework outperforms the state-of-the-art.

The rest of this paper is organized as follows: In Section 2, summaries of some related methods including traditional threshold segmentation and CNN-based segmentation are made. In Section 3, we introduce the detailed procedure of our framework. The experimental settings and evaluation approaches are illustrated in Section 4 followed by the analysis of the results in Section 5. The conclusion is summarized at last in Section 6.

2. Related work

Traditionally, DR lesion segmentation approaches are based on hand-craft features as intensity since the HE and MA show low intensity while the exudates are the opposite. Antal et al. [17] designed a context-aware approach to classify the MAs based on their intensity and spatial location. Kaur et al. [18] designed an exudates segmentation pipeline. They extracted and removed the vessels and optic disk firstly by using hand-craft features and then segmented the exudates by dynamic threshold. Similar is the work of Imani et al. [19]. These traditional threshold-based approaches only take the intensity of lesions into consideration. Generally, the feature selection and the hyper-parameters of hand-craft-based approaches have a great influence on the performance and are difficult to be decided, which leads to a lack of the generalized ability [10].

With the development of computer science, deep neural networks, especially the CNN, have gradually entered people's vision. After giving supervision to the output, the CNN can automatically learn to extract the most discriminative features by gradient descent. On account of their promising performance and convenience, in recent years, CNN is widely utilized in computer vision and medical image. A lot of researchers have introduced CNN approaches to DR lesion segmentation.

Van et al. [20] proposed an approach to speed up CNN training. They increased the probability to select the false positive samples to improve the identify HE lesions. Huang et al. [21] proposed a bounding box refining network. They generated fine annotations and simulated corresponding coarse annotations and then trained a network to refine the coarse annotations. Sarhan et al. [13] proposed a two-stage deep learning MA detection approach. In the first stage, a fully convolutional network was utilized to propose the region of interest (ROI). In the second stage, the ROI candidates were further classified by a CNN with triplet loss to refine the MA detection. Adem et al. [22] used circular Hough transformation to remove the optic disk, then used a CNN to decide if the exudates exist in the retinal image. Guo et al. [23] proposed a top-k loss and a bin loss to solve the class unbalances in exudates segmentation. Xie et al. [12] designed a four-stage segmentation pipeline including segmentation, emendation, re-segmentation, and verification to detect and refine the MA. Yan et al. [11] designed a framework with global image and local patch inputs. These works treat each DR lesion separately, which is less efficient and practical value.

Some researchers noticed the advantages of simultaneous multi-lesion segmentation recently [24]. Guo et al. [10] designed a multi-lesion segmentation network. They added one side extraction layer to each convolutional layers group of VGG-net [25] to extract and fuse the features of different scales, and segmented DR lesions simultaneously. Saranya et al. [26] designed a diabetic retinopathy detection framework. They introduced a set of pre-process approaches including resizing, binarization, contrast enhancement, and morphological transformations, then they segmented MA and HE simultaneously by UNet [27]. Huang et al. [14] developed a novel network by introducing self-attention and cross-attention blocks to UNet for better interaction between lesions and vessels. However, they ignored the impact of different lesion scales. He et al. [28] noticed the lesion scales. They extracted feature maps from the blocks with different depths and

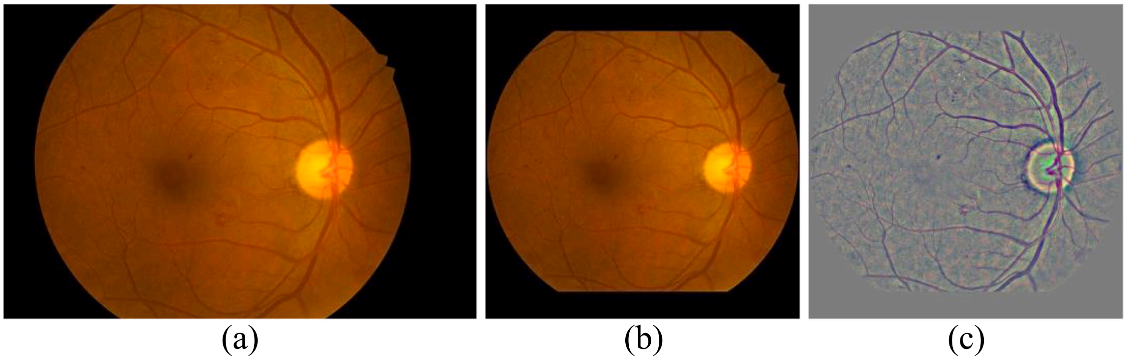

Fig. 2. One example of image pre-processing. (a) is the RAW CFP image, i.e. χ, with a large and random resolution. (b) is the reshaped RGB image $\boldsymbol{x}^{RGB}$ as defined in Eq. (2). (c) is the pre-processed image $\boldsymbol{x}^{eh}$ as formulated in Eq. (1).

designed attention blocks to fuse them. Liu et al. [29] also noticed the lesion scales. They aimed at segmenting tiny DR lesions and designed many-to-many feature reassembly mechanisms to segment them.

3. Methods

3.1. Image pre-processing

The retinal images acquired from different cameras may have different backgrounds, which are reflected in color, illumination, and contrast. Normalizing these differences can promote the deep learning process. [30] The normalizing process is described as follows [31]. The background information is first calculated by using a Gaussian filter on the original RGB image. The original RGB image then subtracts the calculated background to normalize the color information while preserving the anatomical structure. Two factors are then multiplied and added to tune the contrast to restrict the total intensity. And we crop, pad, and resize the image to an appropriate size finally. This process can be formulated as Eq. (1) [31].

$$\boldsymbol{x}^{eh} = f_{CPR}[\alpha(\chi - G_\sigma * \chi) + \gamma] \tag{1}$$

where χ denotes the original RGB fundus image; $\boldsymbol{x}^{eh}$ denotes the pre-processed image; G_σ is a Gaussian filter kernel with a variance of σ; $*$ denotes the convolution operation; f_{CPR} denotes the reshaping process including cropping, padding, and resizing; Factors α and γ are to tune the contrast and restrict the total intensity [31]. The three factors α, γ, and σ are empirically set to be 4, 0.5, and $r/30$ respectively, where r is the radius of the FoV [9,30].

The anatomical structure is enhanced while the color information is lost after Eq. (1). To preserve the color information, we also reshape (crop, pad, and resize) the RGB image in the same way as above, which can be formulated as Eq. (2).

$$\boldsymbol{x}^{RGB} = f_{CPR}[\chi] \tag{2}$$

where χ denotes the original RGB fundus image; f_{CPR} denotes the reshaping process including cropping, padding, and resizing. And $\boldsymbol{x}^{RGB}$ is the reshaped RGB fundus image. One example is given in Fig. 2.

3.2. Framework designing

The framework of our approach consists of two stages as shown in Fig. 3. In the first stage, we train a set of fully convolutional networks (FCN, as the 'Net 1, …, Net N' shown in Fig. 3) with different input scales to acquire multi-scale lesion segmentation masks. In the second stage, we train a CNN (Net W) to learn the appropriate fusion weight, which is subsequently used to dynamically fuse the segmentation masks to get refined results.

3.2.1. FCN for multi-scale lesion segmentation

The anatomical structure is enhanced after pre-processing by Eq. (1). However, the color information, which is also important in lesion identification, is damaged. To fully leverage the valuable information from both structure and color, both pre-processed and RGB images should be learned by the networks.

Introducing the network parameters which had been well trained on the large-scale dataset into other tasks, as known as transfer learning, has been proven to be effective in accelerating convergence and improving the performance [32]. However, the state-of-the-art deep neural networks are mainly trained on the natural 3-channel, i.e. RGB, image domain, which is not suitable to be introduced directly in our task. In this paper, we use the same network structure as our previous work [9], which is shown below in Fig. 4. Concretely, the FCN of this work is changed based on CE-Net [32]. The input block of CE-Net is composed of a convolutional (Conv) layer, a batch normalization (BN) layer, a Relu layer, and a max pooling (MP) layer. We add a duplicate input block. The pre-processed image $\boldsymbol{x}^{eh}$ and RGB image $\boldsymbol{x}^{RGB}$ are fed into two input blocks respectively. The output feature maps from two input blocks are then added and fed to subsequent layers which are the same as CE-Net. Note that all models in this work take the combination of $\boldsymbol{x}^{eh}$ and $\boldsymbol{x}^{RGB}$ as input. For convenience, we denote the combination of them as x, the same below.

In this work, we train N FCN (Net 1, …, Net N) with the same architecture by using N different input scales. We denote the nth scale as $s(n)$, and the input image with a scale of $s(n)$ as $x_{s(n)}$. The nth FCN, i.e. Net n, can thus be denoted as $F_{s(n)}$. The N FCN will predict N lesion segmentation maps. This process can be formulated as Eq. (3)

$$P_{s(n)} = F_{s(n)}(x_{s(n)}), (n = 1, 2, \ldots, N) \tag{3}$$

where $P_{s(n)} \in \mathbb{R}^{C \times s(n) \times s(n)}$ denotes the prediction map with a scale of $s(n)$ and C is the number of channels. Note that we temporarily treat the HE and MA as the same lesion named 'dark lesion' during network training and testing, and thus the number of segmentation map's channel C equals three in this stage.

3.2.2. Separation of MA and HE

After the segmentation of dark lesion, EX, and CW in the first stage, we need to separate MA and HE before the second stage. Take the scale $s(n)$ as an example, the prediction map $P_{s(n)}$ is resized to a fixed scale of S (is set to be 1024 in this work), which is denoted as $P'_{S(n)}$. The difference between HE and MA is mainly reflected in size, where the former is larger than the latter most time. Based on this, threshold segmentation is firstly introduced on the 'dark lesion' (the first in our work) channel of $P'_{S(n)}$ to get a binary mask $P_{BS(n)}$.

$$P_{BS(n)} = [P'_{S(n)} \langle 1, : \rangle > T] \tag{4}$$

where $\langle 1, : \rangle$ denotes the first channel of a tensor. We calculate the area of each binary region in $P_{BS(n)}$ secondly. We denote the mask with

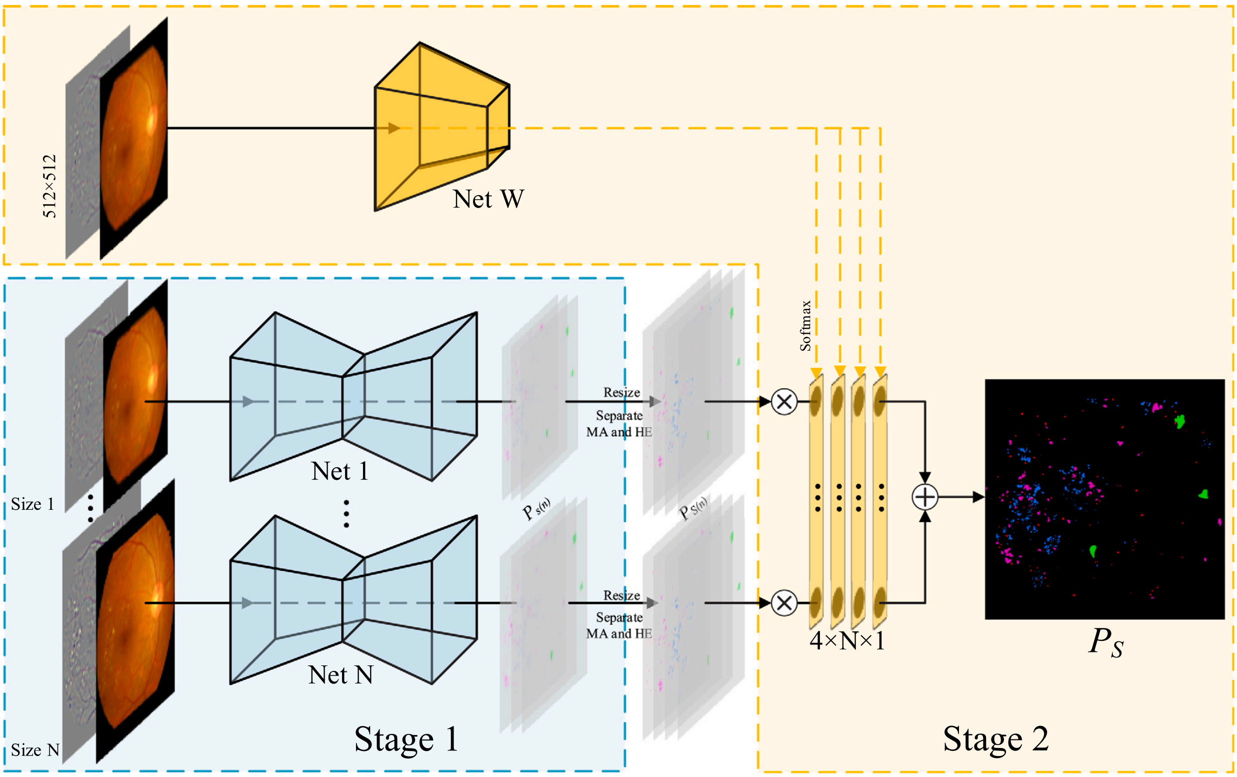

Fig. 3. The whole process of our framework. In the first stage, we train N FCN to segment DR lesions from different scales. In the second stage, we train a CNN to fuse these N predictions of FCN.

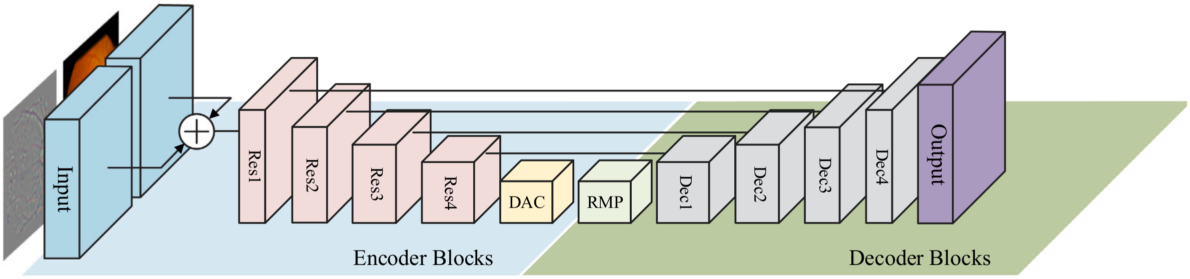

Fig. 4. The structure of FCN in this work. The input block contains a Conv layer, a BN layer, a Relu layer, and an MP layer. We add the same input block to CE-Net [32] to receive the pre-processed image. 'Res*' denotes the basic block of ResNet; 'DAC' denotes the dense atrous convolution block [32]; 'RMP' denotes the residual multi-kernel pooling block [32]; 'Dec' denotes the decoder block. 'Res*', 'DAC', 'RMP', and 'Dec' blocks are all the same as CE-Net [32]. The Encoder blocks are composed of the blocks on the left side of the DAC block while the Decoder blocks are composed of the blocks on the right side of the RMP block.

all regions larger than 20 pixels as $P_{R1(n)}$. Similarly, the mask with all regions smaller than 20 pixels is denoted as $P_{R2(n)}$, and apparently, $P_{BS(n)} = P_{R1(n)} + P_{R2(n)}$. Finally, the MA and HE segmentation masks are calculated by $P'_{S(n)}$ dot multiplying $P_{R1(n)}$ and $P_{R2(n)}$ respectively as shown in Eq. (5).

$$\begin{cases} P_{MA(n)} = P_{R1(n)} \cdot P'_{S(n)} \langle 1, : \rangle \\ P_{HE(n)} = P_{R2(n)} \cdot P'_{S(n)} \langle 1, : \rangle \end{cases} \tag{5}$$

where $P_{MA(n)}$ and $P_{HE(n)}$ denote MA and HE segmentation masks respectively. $\cdot$ denotes the dot multiplying, i.e., element-wise multiplying.

As for the EX and CW segmentation masks $P_{EX(n)}$ and $P_{CW(n)}$, we have $P_{EX(n)} = P'_{S(n)} \langle 2, : \rangle$ and $P_{CW(n)} = P'_{S(n)} \langle 3, : \rangle$. Finally, the segmentation mask after the separation of MA and HE is the concatenation of $P_{MA(n)}$, $P_{HE(n)}$, $P_{EX(n)}$, $P_{CW(n)}$, and can be denoted as $P_{S(n)}$.

3.2.3. CNN for weight fusion

The segmentation masks from N scales perform differently between types of lesions. For example, the MA segmentation performs better when the input scale is large while the CW is the opposite. To dynamically fuse these N predicted lesion masks, i.e. $P_{S(n)}$, we train a CNN (Net W) to learn the weight for the ensemble. The first few layers of our CNN are the same as the Encoder blocks of our FCN, and the following layers are summarized in Table 1. Concretely, the input of CNN x is resized to 512 × 512, then it is fed into the Encoder blocks which are the same as that of our FCN, and then it is pooled and flattened to

Table 1
The architecture of CNN layers. The Encoder blocks are the same as that of our FCN.

Layer	Input shape	Output shape
Encoder blocks	3 × 512 × 512	512 × 16 × 16
AveragePooling	512 × 16 × 16	512 × 1 × 1
Flatten	512 × 1 × 1	512
FC	512	4N
Reshape	4N	4 × N
Softmax(dim = 2)	4 × N	4 × N

a 512-dimension embedding. A fully connected (FC) layer is used to reduce the embedding dimension to 4 N, where N is the number of FCN. To guarantee that the sum of one lesion's weights equals 1.0, the embedding is finally reshaped and activated by a Softmax function. The algorithm of Net W can be formulated as Eq. (6)

$$W = F_W(x) \tag{6}$$

where F_W denotes the Net W and $W \in \mathbb{R}^{4\times N}$ denotes the output tensor of Net W, i.e. fusion weight. The element of W is denoted as $W\langle c,n\rangle$. Apparently we have $\sum_{n=1}^{N} W\langle c,n\rangle = \mathbf{1}^4$. The final segmentation result $P_S \in \mathbb{R}^{4\times S\times S}$ is the weighted sum of N predicted lesion maps by using W as the weight, which can be formulated as Eq. (7)

$$P_S\langle c,i,j\rangle = \sum_{n=1}^{N} P_{S(n)}\langle c,i,j\rangle \cdot W\langle c,n\rangle \tag{7}$$

where $\langle c,i,j\rangle$ denote the element of corresponding tensors.

3.3. Training and testing process

We use $P_{s(n)}$ and $G_{s(n)}$ to denote the prediction map and ground-truth (GT) with a scale of $s(n)$ respectively and use $\langle c,i,j\rangle$ to denote the element (pixel) which locate at the ith row, jth row, and cth channel. Apparently, we have $1 \le c \le C$, where C is the number of segmentation map's channel and equals 3 and 4 in the first and second stages respectively. Besides, we have $1 \le i,j \le s(n)$ and $1 \le n \le N$.

Training: Our training process contains two stages. In the first stage, we train N FCN using N input scales respectively. The GT lesion masks are used as the supervision. The sum of the binary cross-entropy loss and Dice loss is used as the loss function. The binary cross-entropy loss and the Dice loss of the nth FCN can be formulated as Eqs. (8) and (9) respectively.

$$L_{BCE}(P_{s(n)},G_{s(n)}) = \frac{1}{Cs(n)^2}\sum_{c=1}^{C}\sum_{i,j=1}^{s(n)}\big[G_{s(n)}\langle c,i,j\rangle \log P_{s(n)}\langle c,i,j\rangle \\ +(1-G_{s(n)}\langle c,i,j\rangle)\log(1-P_{s(n)}\langle c,i,j\rangle)\big] \tag{8}$$

$$L_{Dice}(P_{s(n)},G_{s(n)}) = C - \sum_{c=1}^{C}\frac{2\sum_{i,j=1}^{s(n)} P_{s(n)}\langle c,i,j\rangle G_{s(n)}\langle c,i,j\rangle}{\sum_{i,j=1}^{s(n)} P_{s(n)}^2\langle c,i,j\rangle + \sum_{i,j=1}^{s(n)} G_{s(n)}^2\langle c,i,j\rangle} \tag{9}$$

The total segmentation loss of the nth FCN can thus be formulated as Eq. (10).

$$L_{seg}(P_{s(n)},G_{s(n)}) = L_{BCE}(P_{s(n)},G_{s(n)}) + L_{Dice}(P_{s(n)},G_{s(n)}) \tag{10}$$

We use Eq. (10) to update the parameters of 'Net 1 ... Net N'. After all the N FCN have been well trained, the second stage will be started. In the second stage, the final prediction map P_S is calculated by Eq. (7). We also calculate the loss between the final prediction map P_S and corresponding GT G_S by using the sum of the binary cross-entropy loss and Dice loss as Eq. (11).

$$L_{CNN}(P_S,G_S) = L_{BCE}(P_S,G_S) + L_{Dice}(P_S,G_S) \tag{11}$$

Where L_{BCE} and L_{Dice} have the same formulation as Eqs. (8) and (9) respectively. We use Eq. (11) to update 'Net W' in the second stage.

Testing: In the testing phase, an unseen image is resized to N scales, which are then fed into N FCN respectively to get a set of prediction maps $P_{s(n)}$. These N prediction maps are then resized to a size of S. At the same time, the unseen image is resized to 512 × 512 and fed into 'Net W' to get the fusion weight W. The final prediction map P_S is then calculated by Eq. (7). The difference between P_S and corresponding GT G_S is then evaluated to assess the performance.

Table 2
The components of the datasets utilized in this work. The number of cases with certain lesion is listed on the right side.

Dataset	Total images	Num of cases			
		MA	HE	EX	CW
IDRiD	81	81	80	81	40
DDR	608	438	488	416	153
E-ophtha	21	21	–	21	–
DIARETDB1	89	80	54	48	36
Local	332	279	269	228	41

4. Experiments

4.1. Materials

The publicly available databases including IDRiD [33], DDR [34], E-ophtha [35], and DIARETDB1 [16] are utilized in this work. IDRiD was acquired by a Kowa VX-10 alpha digital fundus camera with 50° FoV. It was used to hold a challenge and consisted of segmentation, grading, and localization sub-task [4]. The CFP images in DDR dataset were taken by multiple types of fundus cameras with 45° FoV. Both two datasets have been split into their own training and testing subsets by their authors. E-ophtha dataset contains 148 images with MA and 47 images with EX. 21 of them contain both MA and lesions [34]. DIARETDB1 dataset was acquired by a ZEISS FF 450 + fundus camera with Nikon F5 digital camera, 50° FoV. However, the annotations of them are coarse [36,37].

The images in the databases mentioned above are 45~50° FoV which is mainly used in screening. We also collect some fundus photographs which are mainly used in clinical examination. The study was approved by the institutional review board of Shanghai General Hospital and conducted in accordance with the tenets of the Declaration of Helsinki. The requirement for written consent was waived by the institutional review board because of the retrospective nature of the study. And all analyzed data were anonymized and de-identified. Concretely, 332 photographs from visitors to the Shanghai General Hospital, with their consent, were utilized in this work. Photographs were taken by Topcon TRC-50DX with 30° FoV.

To acquire fine and pixel-wise annotations, we asked three ophthalmologists (two junior ophthalmologists and one senior ophthalmologist) to annotate our local dataset and refine the DIARETDB1 dataset. The annotations are made by drawing contours along the boundary of each lesion. Two junior ophthalmologists first annotated all unlabeled data respectively, and then the senior ophthalmologist reviewed and fine-tuned the annotations. The refined annotations of the DIARETDB1 dataset are publicly shared on Google Drive.[1] The components of these four datasets is summarized in Table 2.

4.2. Evaluation metrics

To evaluate the performance of our method, we plot the precision and recall (PR) curves of each lesion, then we calculate the area under the curve (AUPR) of them. Besides, we also evaluate the dice similarity coefficient (DSC) which can be formulated as Eq. (12)

$$DSC = \frac{2TP}{2TP+FP+FN} \tag{12}$$

where TP, FP, and FN denote the numbers of true positive, false positive, and false negative cases. Since the goal is to segment all lesions, the TP is calculated using a low threshold of 0.1 in this task

[1] https://drive.google.com/drive/folders/18bsix3hbh3TUxaD3lOPXBM9yJOgW0JHt?usp=sharing.

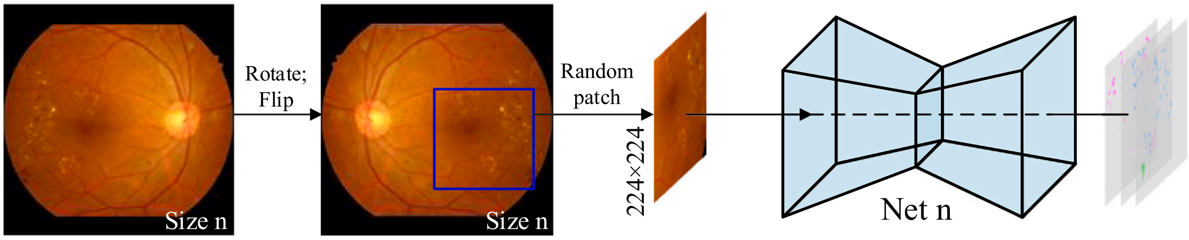

Fig. 5. The data augmentation approach in the first stage during training.

to get a relatively large sensitivity (recall). In fact, DSC equals F1-score which is the harmonic mean of precision and recall in this task. Note that some of the testing images may not have a certain lesion. Therefore, during evaluation, we regard each pixel as one case, regard the testing set as one large pixel set, then plot the curves, and then calculate the AUPR and DSC of all pixels.

4.3. Experiment detail

In this work, all experiments were conducted on an NVIDIA DGX workstation equipped with Intel Xeon(R) Platinum 8168 CPU and Tesla V100-SXM3 32 GB GPU. We use PyTorch as our deep-learning framework. The optimizer is Adam with a learning rate of (1e-4), and the network is initialized by loading the parameters of trained ResNet34 [38].

We train and test on each dataset separately using initialized network. To better augment the data, besides introducing the common augmentation methods including random rotation and flip, we extract patches randomly with the size of 224 from the image during training in the first stage, which is shown in Fig. 5. We use the sum of binary cross-entropy loss and the Dice loss as Eq. (10). The training epoch is set to be 10 000. Three scales (i.e. N = 3) including 384, 512, and 1024 are utilized in this work. In the second stage, the data augmentation approach is random rotation and flip. The loss function is Eq. (11), and the total epoch is set to be 1000.

5. Results and discussion

5.1. Comparisons with baselines

IDRiD and DDR datasets have been split into training and testing subsets by their proposers. E-ophtha dataset is split into 11 and 10 for training and testing respectively as suggested by Guo et al. [10] and He et al. [28]. We split one-fourth of the images in DIARETDB1 and our local datasets as testing sets. The summary of data usage is listed in the first column of Table 3 below dataset name. Extracting patches randomly means massive training data can be generated from one CFP image, and therefore the network can be trained on only dozens of CFP images.

The results on IDRiD, DDR, E-ophtha, DIARETDB1, and our local datasets are plotted from left to right in Fig. 6. The horizontal axis is the recall score while the vertical axis is the precision score. The red, pink, blue, and green curves correspond to the MA, HE, EX, and CW lesions, respectively, in each sub-figure. We can see that the MA curves drop significantly as the recall score increases, which means the AUPR of MA is low.

The AUPR and DSC (F1) of each lesion are calculated, and the comparisons with others are listed in Tables 3 and 4. We can observe from Table 3 that our approach shows competitive performance, especially in CW lesion segmentation.

Concretely, on the IDRiD dataset, our approach achieves the AUPRs of 49.75%, 68.84%, 89.61%, and 75.63% for the segmentation of MA, HE, EX, and CW respectively. The DSC of these four lesions reaches 52.68%, 63.59%, 79.47%, and 72.33% respectively. Our approach ranks first in the AUPR and DSC of CW and the average AUPR and DSC of all lesions. Table 4 shows the results of the top three teams in the IDRiD challenge. Our approach ranks first in the AUPR of HE, EX, CW, and the average AUPR of all lesions.

DDR is a recent released dataset which have only been experimented by a few researchers. The image quality of this dataset is relative lower than the IDRiD dataset, which lead to lower performance [14]. On the DDR dataset, our approach ranks first in the segmentation of MA, EX, and CW with the AUPRs of 11.94%, 60.14%, and 32.60% respectively. The DSC of four lesions are 15.22%, 33.17%, 54.06%, and 38.76% respectively.

E-ophtha dataset contains only MA and EX lesions. We achieve the best performance of EX with AUPR of 54.16% and DSC of 53.57%. The AUPR and DSC of MA are 30.38% and 33.61% respectively.

We also conduct our experiments on the datasets annotated or refined by our experts. Our approach achieves the AUPRs of 46.77%, 76.30%, 81.15%, and 63.43% for MA, HE, EX, and CW respectively, and achieves the DSC scores of 49.96%, 69.31%, 76.12%, and 62.23% respectively on the DIARETDB1 dataset. The AUPR scores on our local dataset reach 41.94%, 63.75%, 76.96%, and 36.75% respectively. And the DSC scores of them are 44.63%, 56.35%, 71.20%, and 38.79% respectively.

To show the effectiveness of our approach, we conduct two baseline models including UNet [27] and Swin-UNet [39] on all datasets mentioned in this paper. The input size of these two models is 512. The data augmentation approach includes random rotation and flip. And the results are also listed in Table 3. It can be seen that our approach outperforms the two baseline models significantly. Besides, UNet performs better than Swin-UNet of the most time.

Some samples of segmentation results from these datasets are given in Fig. 7, where the original RGB images, corresponding GTs, and our results are listed on the first, second, and third rows respectively. And the five columns show the samples from IDRiD, DDR, E-ophtha, DIARETDB1, and our local datasets respectively.

Xie et al. [12] and Sarhan et al. [13] proposed methods to segment the MA lesions only. These methods may achieve promising results. However, they are specific to one lesion and will show unstable performances when conducted on other lesions. Designing one specific method for one lesion lacks efficiency and generalization ability. Huang et al. [14] proposed a network with self-attention and cross-attention blocks. The self-attention block is designed to learn the relations between lesions while the cross-attention block is designed to interact between lesions and vessels. However, ignoring the impact of the lesion scale leads to limited performance. Besides, their network needs to learn to segment vessels first, during which the extra annotations are introduced.

Guo et al. [10], Yan et al. [11], He et al. [28], and Liu et al. [29] all took the impact of the lesion scale into consideration. Guo et al. [10] aimed at giving the last layer of VGG16 block supervision. The small lesion as MA will be well supervised at the first several blocks in this way. He et al. [28] utilized a similar idea. They extracted feature maps from each of the blocks and fused them. However, the network is also bound after giving too much supervision. Yan et al. [11] designed a network that can fuse global and local features. However, as they

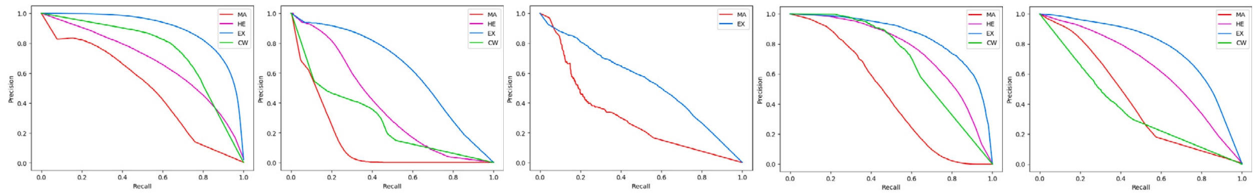

Fig. 6. The PR curves of IDRiD, DDR, E-ophtha, DIARETDB1, and our local datasets (from left to right). The red, pink, blue, and green curves denote the MA, HE, EX, and CW lesions respectively.

Table 3
Performance comparison with others for lesion segmentation. The number of training/testing images is listed in the first column below the dataset name. The AUPR and DSC (F1) of each lesion and the average AUPR and DSC of all lesions are reported. The best results are marked in bold.

Dataset (Train/Test)	Methods	AUPR%					DSC% (F1)				
		MA	HE	EX	CW	Ave	MA	HE	EX	CW	Ave
IDRiD(54/27)	Guo et al. [10]	46.27	67.34	79.45	71.13	66.05	–	–	–	–	–
	Yan et al. [11]	**52.50**	**70.30**	88.90	67.90	69.90	–	–	–	–	–
	Xie et al. [12]	50.99	–	–	–	–	–	–	–	–	–
	Sarhan et al. [13]	41.96	–	–	–	–	43.23	–	–	–	–
	Liu et al. [29]	48.80	68.69	82.16	69.32	67.24	48.63	**66.42**	**79.85**	67.92	65.71
	He et al. [28]	46.94	67.05	87.24	71.11	68.08	–	–	–	–	56.02
	Huang et al. [14]	48.97	68.80	**90.24**	75.02	70.76	–	–	–	–	–
	UNet [27]	42.39	50.23	75.74	64.40	58.19	44.71	49.57	67.73	62.22	56.06
	Swin-UNet [39]	41.81	52.56	76.91	59.14	57.61	43.06	50.49	69.01	58.37	55.23
	Ours	49.75	68.84	89.61	**75.63**	**70.96**	**52.68**	63.59	79.47	**72.33**	**67.02**
DDR(383/225)	Guo et al. [10]	10.52	35.86	55.46	26.48	32.08	–	–	–	–	–
	Huang et al. [14]	11.76	**36.56**	56.71	29.43	33.62	–	–	–	–	–
	UNet [27]	9.32	19.74	49.68	17.03	23.94	14.23	26.15	44.56	23.74	27.17
	Swin-UNet [39]	10.11	19.77	54.36	16.85	25.27	**16.86**	24.81	47.69	22.11	27.87
	Ours	**11.94**	36.52	**60.14**	**32.60**	**35.30**	15.22	**33.17**	**54.06**	**38.76**	**34.80**
E-ophtha(11/10)	Guo et al. [10]	16.87	–	41.71	–	29.29	–	–	–	–	–
	He et al. [28]	**30.60**	–	51.20	–	40.90	–	–	–	–	**45.43**
	UNet [27]	15.02	–	33.98	–	24.50	22.63	–	31.49	–	27.06
	Swin-UNet [39]	12.34	–	32.75	–	22.55	18.21	–	31.97	–	25.09
	Ours	30.38	–	**54.16**	–	**42.27**	**33.61**	–	**53.57**	–	43.59
DIARETDB1(67/22)	UNet [27]	41.72	67.11	76.47	55.53	60.21	44.53	58.01	69.18	56.26	56.99
	Swin-UNet [39]	37.88	65.81	74.26	53.57	57.88	42.97	58.47	67.82	55.69	56.24
	Ours	**46.77**	**76.30**	**81.15**	**63.43**	**66.91**	**49.96**	**69.31**	**76.12**	**62.23**	**64.41**
Local(249/83)	UNet [27]	36.59	56.28	70.10	30.64	48.40	39.21	50.17	60.98	31.44	45.45
	Swin-UNet [39]	37.02	55.96	70.18	23.44	46.65	38.64	51.55	61.11	27.06	44.59
	Ours	**41.94**	**63.75**	**76.96**	**36.75**	**54.85**	**44.63**	**56.35**	**71.20**	**38.79**	**52.74**

Table 4
Performance comparison with top three teams in IDRiD challenge [4]. The AUPRs of each lesion are reported, and the averages of all lesions' AUPRs are listed in the last column. The best results are marked in bold.

Author/Team	MA	HE	EX	CW	Ave
VRT	49.51	68.04	71.27	69.95	64.69
PATech	47.40	64.90	88.50	–	–
iFLYTEK-MIG	**50.17**	55.88	87.41	65.88	64.84
Ours	49.75	**68.84**	**89.61**	**75.63**	**70.96**

analyzed, the performance of large and compact lesions as CW dropped after fusing local features [11]. Liu et al. [29] aimed at segmenting small lesions. The DSC scores of HE and EX reach 66.42% and 79.85% on the IDRiD dataset. However, their performance for the relatively larger lesion as CW is not promising.

5.2. Ablation analysis

We conduct ablation studies in this subsection to show the impact of the input scale, weighting strategy, multi-task learning, and network structure. The first variable in this work is the input scale. As illustrated in Section 3, we use three scales including 384 × 384 (Size 1), 512 × 512 (Size 2), and 1024 × 1024 (Size 3) to train a set of FCN respectively. The results of individual FCN and the combinations of two FCN are listed in the 'Scale' block of Table 5. Two examples of three individual FCN' segmentation results are visualized in the 'Size 1', 'Size 2', and 'Size 3' sub-figures of Fig. 8 respectively.

From Table 5 we can see that the input scale affects the segmentation result a lot. Concretely, the individual FCN with the input scale of 'Size 1' achieves 30.15%, 62.80%, 78.65%, and 73.11% AUPR scores for the segmentation of MA, HE, EX, and CW respectively. The FCN with the input scale of 'Size 3' achieves 48.88%, 64.01%, 87.50%, and 58.60% AUPR scores for the segmentation of four lesions respectively. And the results of the FCN with the input scale of 'Size 2' are mostly between that of 'Size 1' and 'Size 3'. Apparently, the FCN with a large input scale is good at the segmentation of MA and EX while the FCN with a small input scale is good at the segmentation of CW. The HE segmentation results do not show too many differences between these scales. These characteristics of FCN are visualized in Fig. 8, where the 'Size 3' has detected more MA and EX lesions but fewer CW lesions compared with 'Size 1' and 'Size 2'. The main reason is that the FCN with a small input scale has a large receptive field, and the small lesions as MA and EX are lost due to downsampling. The FCN with a large input scale, on the contrary, has a small receptive field, and the small lesions take up more areas, which means that the details can be fully learned by the network. However, the large lesions as CW take up too many areas when the receptive field is small. At this time, the small input scale shows its advantage of learning more useful and global features. The shape of HE is diverse since it contains both large and small lesions. Therefore, the AUPRs of HE does not show too much difference when the input scale is changed.

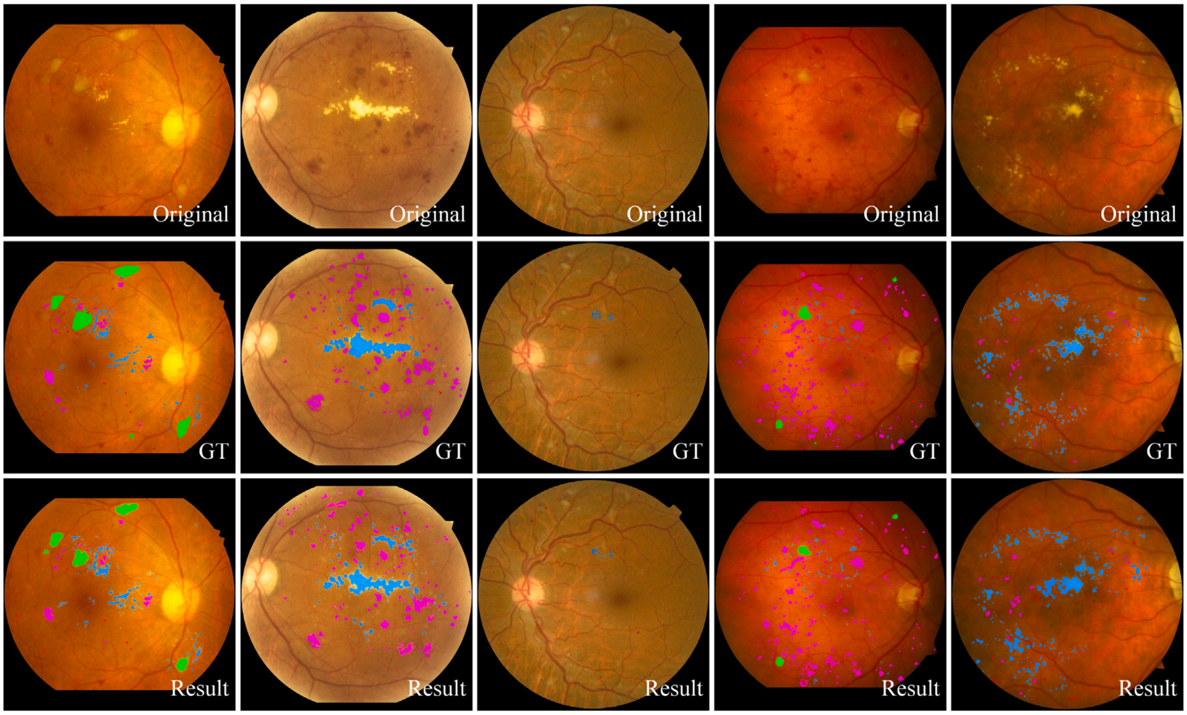

Fig. 7. The samples of segmentation results with original images from IDRiD, DDR, E-ophtha, DIARETDB1, and our local datasets respectively (from left to right). The first row shows the original images. The second row shows the corresponding GTs. And the segmentation results are shown in the last row. The red, pink, blue, and green masks denote the MA, HE, EX, and CW lesions respectively.

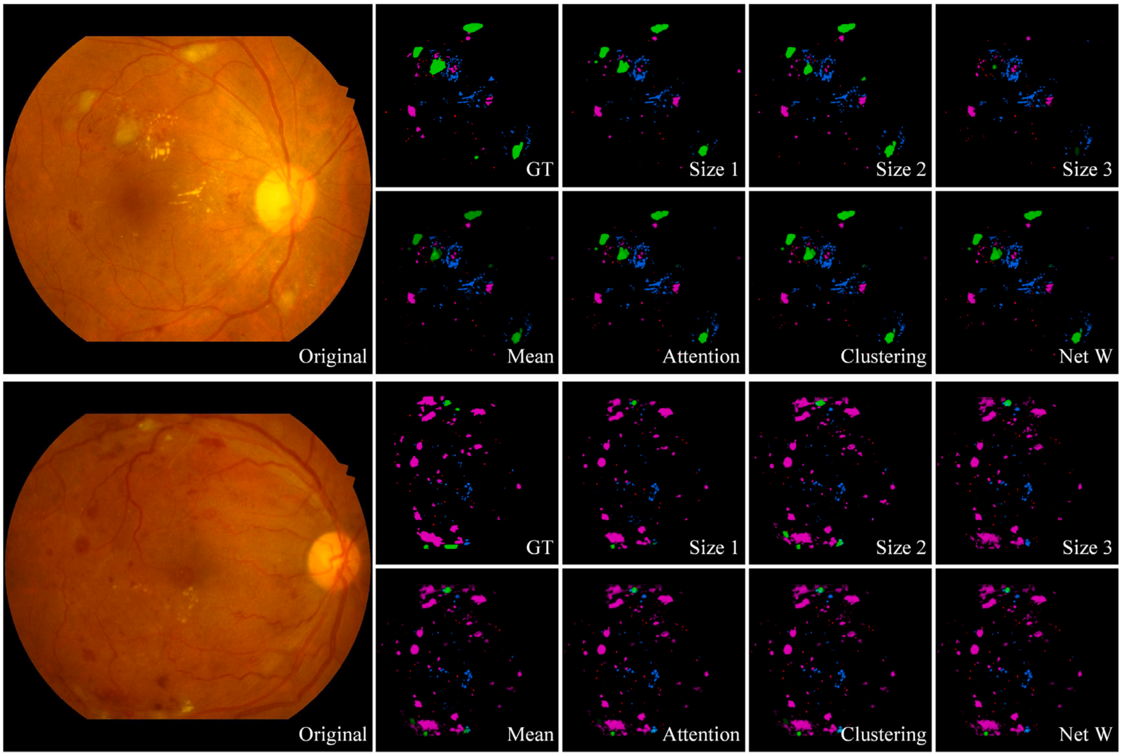

Fig. 8. The visualized results of different scale combinations and weighting strategies. Two samples from the IDRiD dataset are given here. ‘Size 1’, ‘Size 2’, and ‘Size 3’ denote the segmentation maps of the FCN with the input size 384, 512, and 1024 respectively. ‘Mean’ denotes the average of these three segmentation maps. ‘Net W’ denotes using the ensemble weight calculated by our CNN. ‘GT’ denotes the corresponding ground truth. The red, pink, blue, and green masks denote the MA, HE, EX, and CW lesions respectively.

Table 5
Performance comparison of different scale combinations and weighting strategies. The best results in each group are marked in bold.

Variable	Strategy	AUPR%					DSC% (F1)				
		MA	HE	EX	CW	Ave	MA	HE	EX	CW	Ave
Scale	Size 1	30.15	62.80	78.65	**73.11**	61.18	35.62	56.51	71.04	**70.29**	58.37
	Size 2	35.20	60.95	79.79	62.77	59.68	39.03	55.72	71.15	61.68	56.90
	Size 3	**48.88**	**64.01**	**87.50**	58.60	**64.75**	**51.56**	**60.00**	**78.49**	55.97	**61.51**
	Size 1, 2	36.16	64.49	81.71	**74.81**	64.29	41.13	60.95	73.24	**71.19**	61.63
	Size 1, 3	46.52	**68.52**	87.30	72.75	**68.77**	49.58	63.44	**78.91**	70.06	65.50
	Size 2, 3	**49.44**	67.73	**87.36**	70.31	68.71	**52.70**	**63.63**	78.76	67.89	**65.75**
Weight	Mean	48.84	68.19	87.98	74.31	69.78	51.82	63.20	78.68	71.29	66.25
	Attention	49.36	68.42	89.57	75.54	70.72	51.81	63.43	**79.49**	72.18	66.73
	Clustering	49.17	68.56	89.60	75.47	70.70	51.95	**63.60**	79.36	72.31	66.81
	Net W	**49.75**	**68.84**	**89.61**	**75.63**	**70.96**	**52.68**	63.59	79.47	**72.33**	**67.02**
Others	Single task	49.12	68.06	88.13	73.22	69.63	51.35	63.43	79.30	70.11	66.05
	Multi-task	**49.75**	**68.84**	**89.61**	**75.63**	**70.96**	**52.68**	**63.59**	**79.47**	**72.33**	**67.02**
	Swin-UNet	47.91	65.63	86.72	71.25	67.88	49.22	59.78	75.74	66.09	62.71

The advantages of small and large receptive fields inspire us that we can fuse the segmentation results from different input scales for better performance. We conduct the experiments of two-scale combinations, the results of which are listed in the second group of Table 5. From the results of ‘Size 1&2’, ‘Size 1&3’, and ‘Size 2&3’ we can see that the fusion of two scales is better than each of them the most time. And the fusion of three scales achieves the best performance. The fusion of four or more scales may perform better. However, after balancing the performance against the efficiency and complexity, we use three scales in this work.

Since the network can benefit from the fusion of multi-scale, the next key point is how to decide the fusion weight. In this work, we use a CNN, i.e. the Net W, to extract the feature of the retinal image, and calculate the fusion weights dynamically according to the input image. We explore three more fusion approaches. The first is assigning the same weight to each component, i.e. mean approach; The second is introducing attention mechanism [40] to these segmentation maps; The last one is clustering. We also make a comparison of our ‘Net W’ and other fusion approaches, the results of which are listed in the ‘Weight’ of Table 5. The visualized samples are shown in the ‘Mean’, ‘Attention’, ‘Clustering’, and ‘Net W’ sub-figures of Fig. 8. From Table 5 we can see that the ‘Attention’, ‘Clustering’, and our ‘Net W’ perform better than the mean approach in all indicators since these three approaches can maximize the advantage of each scale. The results of ‘Attention’ and ‘Clustering’ do not show too many differences since the fusion weights they have learned are all fixed. Our ‘Net W’ performs a little better than ‘Attention’ and ‘Clustering’ mostly since the fusion weights change dynamically according to the input image.

The effect of multi-task learning is another variable in our work. We make a comparison between multi-task and single-task learning, and the results of them are listed in the third group of Table 5. The results of multi-task learning are a little better than that of single-task learning in all indicators. It is worth noting that the performance of CW shows an apparent decrease. One of the possible reasons is that the number of CW’s pixels is significantly less than dark lesion and EX. The network is hard to extract representative features with insufficient foreground pixels. Learning together with other lesions can increase the ratio of foreground pixels, which means that the network can learn more information and get a full understanding of the lesion and retinal anatomic structure.

We also make a comparison between our FCN and Swin-UNet [39]. In this experiment, we replace our FCN with Swin-UNets [39] and train them using the same strategies as Section 4.3. The results are listed in the third group of Table 5. From Table 5 and the baselines in Table 3 we can see that the Swin-UNet is not suitable for this task. The possible reason is that the lesions distribute randomly. The spatial information which can be captured by Swin-UNet is not discriminative in this task. The data augmentation strategies including random cropping, rotation, and flip also modify the spatial information. Besides, transformers are not suitable for small dataset [41]. Therefore, the performance of Swin-UNet is limited.

5.3. Discussion

The DR lesion segmentation plays an important role in computer-aided diagnosis of the retinal image, which can be used in early diagnosis and treatment to prevent the progression of the disease and to avoid vision loss [14]. A lot of approaches had been proposed on this topic. Traditionally, DR segmentation was done by hand-craft feature as the intensity and size [17–19]. With the development of deep learning, CNN-based approaches have been gradually proposed in recent years [10–14]. They designed different network structures, resize the retinal image into a smaller and fixed size to save memory, and then train the network end-to-end. However, one problem is that the difference in features between the DR lesions, as MA and CW, is significant, which means that using the same input size may be inappropriate.

In this work, we propose a DR lesion segmentation framework to solve the problem of existing CNN-based lesion segmentation works. A set of parallel FCN with different input scales in this work are designed to better segment the DR lesions with different sizes. A weighting CNN is designed to tune the ensemble weights of FCN. The multi-task training strategy makes full use of lesion features, which increase the performance of each sub-task. This work has experimented on four public datasets and one dataset from our local hospital. The data were collected from both 45~50-degree-FoV screening and 30-degree-FoV clinical examination, which increases the diversity of distribution. And compared with other works, as shown in Tables 3 and 4, our framework shows competitive performance.

As a DR lesion segmentation framework, our work can be utilized in the large-scaling early screening of DR. The segmentation mask can give the patients a visualized warning of their disease, and give the ophthalmologists guidance on DR grading. Our framework can detect lesions with different scales simultaneously, which means it can be easily introduced in other types of lesion segmentation. Besides, the FCN of our framework can also be replaced by other network backbones, which may further increase the segmentation performance.

Though achieves promising results, our approach is also limited. The two-stage training and the separation of MA and HE may partly limit the performance since both of them are not completely end-to-end modes. The end-to-end segmentation of all lesions will be developed in our future works.

6. Conclusion

In this paper, we propose a novel DR lesion segmentation framework based on Convolutional Neural Network. Utilizing a set of parallel FCN with different input scales can better extract the features of lesions with different sizes while introducing the weight CNN can better fuse the advantages of these FCN, both of which are beneficial to segmentation. Besides, multi-task learning can make full use of lesion features.

Experiments were conducted on four public datasets and one local dataset which contain multiple degrees of FoVs, and the results are promising compared with other recent works. This framework can be easily introduced in other segmentation tasks.

CRediT authorship contribution statement

Tianjiao Guo: Contributed the central ideas, Analyzed data, Wrote the initial draft. **Jie Yang:** Refined the ideas, Revised the manuscript. **Qi Yu:** Collected data, Labeled data, Refined the ideas.

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.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research is partly supported by Shanghai Agriculture Applied Technology Development Program, China (Grant No. 2019-02-08-00-08-F01118), Committee of Science and Technology, Shanghai, China (No. 19510711200), National Key R&D Program of China (2019YFB1311503), Shanghai Sailing Program, China (20YF1420800), NSFC, China (61661010, 61977046, 62003208, 81600776, 81970846).

References

[1] https://www.who.int/publications/i/item/world-report-on-vision/.

[2] R.P. Danis, M.D. Davis, Proliferative diabetic retinopathy, in: Diabetic Retinopathy, Springer, 2008, pp. 29–65.

[3] G.S. Crabtree, J.S. Chang, Management of complications and vision loss from proliferative diabetic retinopathy, Curr. Diabetes Rep. 21 (9) (2021) 1–8.

[4] https://idrid.grand-challenge.org/Localisation/.

[5] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, T. Wu, J. Xiao, F. Wang, B. Yin, Y. Wang, G. Danala, L. He, Y.H. Choi, Y.C. Lee, S.-H. Jung, Z. Li, X. Sui, J. Wu, X. Li, T. Zhou, J. Toth, A. Baran, A. Kori, S.S. Chennamsetty, M. Safwan, V. Alex, X. Lyu, L. Cheng, Q. Chu, P. Li, X. Ji, S. Zhang, Y. Shen, L. Dai, O. Saha, R. Sathish, T. Melo, T. Araújo, B. Harangi, B. Sheng, R. Fang, D. Sheet, A. Hajdu, Y. Zheng, A.M. Mendonça, S. Zhang, A. Campilho, B. Zheng, D. Shen, L. Giancardo, G. Quellec, F. Mériaudeau, IDRiD: Diabetic retinopathy – Segmentation and grading challenge, Med. Image Anal. 59 (2020) 101561.

[6] Y. Wu, M. Zhang, W. Yu, H. Zheng, J. Xu, Y. Gu, LTSP: long-term slice propagation for accurate airway segmentation, Int. J. Comput. Assist. Radiol. Surg. 17 (5) (2022) 857–865.

[7] T. Zhang, Y. Gu, X. Huang, J. Yang, G.-Z. Yang, Disparity-constrained stereo endoscopic image super-resolution, Int. J. Comput. Assist. Radiol. Surg. 17 (5) (2022) 867–875.

[8] M. Zhang, H. Pan, Y. Zhu, Y. Gu, Progressive attention module for segmentation of volumetric medical images, Med. Phys. 49 (1) (2022) 295–308.

[9] T. Guo, Z. Liang, Y. Gu, J. Yang, Q. Yu, Deep multi-task framework for optic disc and fovea detection, J. Electron. Imaging 30 (4) (2021) 1–18, http://dx.doi.org/10.1117/1.JEI.30.4.043002.

[10] S. Guo, T. Li, H. Kang, N. Li, Y. Zhang, K. Wang, L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images, Neurocomputing 349 (2019) 52–63.

[11] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong, S. Cui, Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 597–600.

[12] Y. Xie, J. Zhang, H. Lu, C. Shen, Y. Xia, SESV: Accurate medical image segmentation by predicting and correcting errors, IEEE Trans. Med. Imaging 40 (1) (2020) 286–296.

[13] M.H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, A. Eslami, Multi-scale microaneurysms segmentation using embedding triplet loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 174–182.

[14] S. Huang, J. Li, Y. Xiao, N. Shen, T. Xu, RTNet: Relation transformer network for diabetic retinopathy multi-lesion segmentation, IEEE Trans. Med. Imaging (2022).

[15] Early Treatment Diabetic Retinopathy Study Research Group, et al., Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics: ETDRS report number 7, Ophthalmology 98 (5) (1991) 741–756.

[16] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, J. Pietilä, The diaretdb1 diabetic retinopathy database and evaluation protocol, in: BMVC, Vol. 1, 2007, pp. 1–10.

[17] B. Antal, A. Hajdu, Improving microaneurysm detection in color fundus images by using context-aware approaches, Comput. Med. Imaging Graph. 37 (5–6) (2013) 403–408.

[18] J. Kaur, D. Mittal, A generalized method for the segmentation of exudates from pathological retinal fundus images, Biocybern. Biomed. Eng. 38 (1) (2018) 27–53.

[19] E. Imani, H.-R. Pourreza, A novel method for retinal exudate segmentation using signal separation algorithm, Comput. Methods Programs Biomed. 133 (2016) 195–205.

[20] M.J. Van Grinsven, B. van Ginneken, C.B. Hoyng, T. Theelen, C.I. Sánchez, Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images, IEEE Trans. Med. Imaging 35 (5) (2016) 1273–1284.

[21] Y. Huang, L. Lin, M. Li, J. Wu, P. Cheng, K. Wang, J. Yuan, X. Tang, Automated hemorrhage detection from coarsely annotated fundus images in diabetic retinopathy, in: 2020 IEEE 17th International Symposium on Biomedical Imaging, ISBI, IEEE, 2020, pp. 1369–1372.

[22] K. Adem, Exudate detection for diabetic retinopathy with circular Hough transformation and convolutional neural networks, Expert Syst. Appl. 114 (2018) 289–295.

[23] S. Guo, K. Wang, H. Kang, T. Liu, Y. Gao, T. Li, Bin loss for hard exudates segmentation in fundus images, Neurocomputing 392 (2020) 314–324.

[24] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, H. Fu, Applications of deep learning in fundus images: A review, Med. Image Anal. 69 (2021) 101971.

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[26] P. Saranya, R. Pranati, S.S. Patro, Detection and classification of red lesions from retinal images for diabetic retinopathy detection using deep learning models, Multimedia Tools Appl. (2023) 1–21.

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[28] A. He, K. Wang, T. Li, W. Bo, H. Kang, H. Fu, Progressive multiscale consistent network for multiclass fundus lesion segmentation, IEEE Trans. Med. Imaging 41 (11) (2022) 3146–3157.

[29] Q. Liu, H. Liu, W. Ke, Y. Liang, Automated lesion segmentation in fundus images with many-to-many reassembly of features, Pattern Recognit. 136 (2023) 109191.

[30] L. Zhou, Y. Zhao, J. Yang, Q. Yu, X. Xu, Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images, IET Image Process. 12 (4) (2017) 563–571.

[31] T. Guo, Z. Liang, Y. Gu, K. Liu, X. Xu, J. Yang, Q. Yu, Learning for retinal image quality assessment with label regularization, Comput. Methods Programs Biomed. 228 (2023) 107238.

[32] 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.

[33] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau, Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research, Data 3 (3) (2018) 25.

[34] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang, Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, Inform. Sci. 501 (2019) 511–522.

[35] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, et al., TeleOphta: Machine learning and image processing methods for teleophthalmology, Irbm 34 (2) (2013) 196–203.

[36] C. Playout, R. Duval, F. Cheriet, A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, Springer, 2018, pp. 101–108.

[37] C. Playout, R. Duval, F. Cheriet, A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images, IEEE Trans. Med. Imaging 38 (10) (2019) 2434–2444.

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

[39] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, Springer, 2023, pp. 205–218.

[40] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[41] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, H. Shi, Escaping the big data paradigm with compact transformers, 2021, arXiv preprint arXiv:2104.05704.