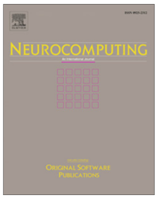
Neurocomputing 527 (2023) 100–109

Contents lists available at ScienceDirect

Neurocomputing

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



CLC-Net: Contextual and local collaborative network for lesion segmentation in diabetic retinopathy images

Xiyue Wang [a,b,1], Yuqi Fang [c,d,2], Sen Yang [d], Delong Zhu [c], Minghui Wang [a,b], Jing Zhang [a,*], Jun Zhang [d], Jun Cheng [e], Kai-yu Tong [f], Xiao Han [d]

[a] College of Biomedical Engineering, Sichuan University, Chengdu 610065, China
[b] College of Computer Science, Sichuan University, Chengdu 610065, China
[c] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
[d] Tencent AI Lab, Shenzhen 518057, China
[e] Institute for Infocomm Research, Agency for Science, Technology of Singapore, Singapore
[f] Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

ARTICLE INFO

Article history:
Received 30 September 2021
Revised 18 December 2022
Accepted 8 January 2023
Available online 10 January 2023
Communicated by Zidong Wang

Keywords:
Fundus lesion segmentation
Diabetic retinopathy
Contextual and local
Joint segmentation and classification
CNN

ABSTRACT

Diabetic retinopathy (DR) is the leading cause of blindness among people of working age. Fundus lesions are clinical signs of DR, and their recognition and delineation are important for early screening, grading, and monitoring of the disease. We propose in this work a fully automatic deep convolutional neural network method for simultaneous segmentation of four different types of DR-related fundus lesions. To exploit multi-scale image information, we propose a collaborative architecture that comprises a contextual branch and a local branch. An attention mechanism is designed to fuse feature maps from all decoding layers in order to effectively and fully combine informative features from the two branches. Moreover, an auxiliary classification task with a novel supervision scheme is introduced to reduce model overfitting and further improve the accuracy of lesion segmentation. Extensive experiments are conducted using three public fundus datasets, and our method produces a mean AUC value of 0.677, 0.629, and 0.581 on them respectively. The results demonstrate the advantages of the proposed method, outperforming alternative strategies and other state-of-the-art methods in the literature.

1. Introduction

Diabetic retinopathy (DR), a type of diabetes complication, is the leading cause of blindness and visual impairment in people aged 20 to 64 years [1]. It is estimated that by 2030, the number of cases with DR will reach 191 million, which will place a heavy burden on the global healthcare system [2]. Because DR does not show obvious symptoms in the early stages, it is difficult to get noticed until the condition becomes severe. Therefore, it is important to screen for DR on a regular basis to help detect the disease early and prevent it from developing into a vision-threatening stage. Fundus photography, an examination of the details of the retina, is commonly used to detect DR. There are certain lesions that are early clinical signs of DR and often indicate the severity level of DR, including microaneurysms (MAs), hemorrhages (HEs), hard exudates (EXs), and soft exudates (SEs), as shown in Fig. 1. Effective segmentation of them is essential for proper diagnosis and monitoring of the disease. In clinical practice, however, manually identifying these lesions from a large number of fundus images is very cumbersome and time-consuming for human experts. Moreover, even for ophthalmologists with extensive experience, recognition accuracy cannot be ensured. This difficulty is mainly caused by three factors: (1) The shape, size, and appearance of each lesion type can vary widely across different individuals; (2) different lesions may share the same characteristics, *e.g.*, MAs and HEs both have low image intensities while EXs and SEs are both much brighter; (3) some lesions, *e.g.*, MAs, are extremely small, ranging from one pixel to a few pixels. A fully automatic DR lesion segmentation system is thus highly desired that can significantly reduce the burden on ophthalmologists and improve consistency in the final diagnosis.

In recent years, numerous deep learning based methods have been developed for DR lesion segmentation [3–10]. Because typical fundus images have very high image resolution (*e.g.*, 4288×2848 pixels), many methods downsample the input image to a smaller

[*] Corresponding author.
E-mail address: jing_zhang@scu.edu.cn (J. Zhang).
[1] Authors contributed equally to this work.
[2] Authors contributed equally to this work.

https://doi.org/10.1016/j.neucom.2023.01.013
0925-2312/© 2023 Elsevier B.V. All rights reserved.

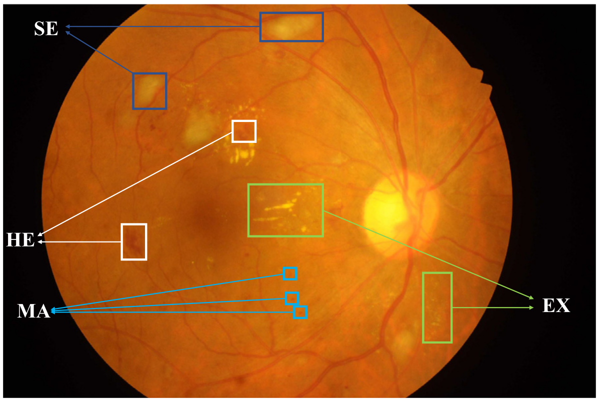


Fig. 1. An example fundus image with DR lesions, including hemorrhage (HE), microaneurysm (MA), hard exudate (EX), and soft exudate (SE).

size (*e.g.*, 640×640) in order to reduce the burden on computation resources [3,4,9]. Although global information can still be captured after downsampling, detailed image information is lost. In contrast, some other methods [5,6,11] crop the high-resolution images into small patches (*e.g.*, 64×64), and then perform patch-wise segmentation on them. Although detailed image information is preserved, global contextual information is missing, which makes lesion segmentation more difficult and may cause inconsistency across nearby patches. More recently, a few studies [7,8] adopt a collaborative strategy, which aggregates contextual and local information of fundus images, but multi-scale feature representations are not effectively fused and leveraged in them.

Based on these observations, we propose in this work a novel collaborative approach that integrates contextual and local information simultaneously and effectively. Our proposed model is comprised of two encoder-decoder branches (*i.e.*, a contextual branch and a local branch), which extract the contextual and local patch features respectively. Specifically, the inputs of the local branch are patches of size 256×256 cropped from the original high-resolution fundus images by a sliding window. Centered around each of these patches, a larger image patch of size 512×512 is correspondingly cropped and serves as the corresponding input of the contextual branch. The larger patches of the contextual branch not only take peripheral information around the local patches into consideration, but also avoid the drawbacks of downsampling as in [8,3]. Consequently, all detailed image information is preserved. Furthermore, to fully integrate the multi-scale features of the two branches, output feature maps of all the decoder layers in the contextual branch and those in the local branch are fused through an effective attention mechanism. Contrary to our method, only feature maps from the last decoder layers are fused in [8], and no concatenation between two branches is found in [7]. This novel concatenation strategy using attention as well as the utilization of different input scales constitutes the first contribution of our proposed method.

Another improvement proposed in this paper is applying a multi-task learning strategy for DR lesion segmentation, which is implemented by adding an auxiliary classification head at the bottom of the network encoder. Different from single-task learning that learns each task in isolation, multi-task learning simultaneously exploits commonalities and differences across multiple tasks, *e.g.*, a segmentation task and a classification task. Usually, multi-task learning can produce better results [12] by leveraging domain-specific information contained in the related tasks. Specifically, the model tends to find representative features shared by different tasks, which significantly reduces the risk of overfitting. In the area of fundus image processing, there exist some studies on multi-task learning, but most of them combine a fundus lesion segmentation task and a DR severity classification task [13,14]. In such a method, subnetworks corresponding to the two tasks share the same encoder for common feature extraction but learn task-specific features in their respective decoders. Since DR severity is directly determined by the presence of fundus lesions, better fundus lesion segmentation can result in better severity level prediction. Similarly, more accurate classification results also help improve the performance of lesion segmentation. However, there is a clear weakness in those methods, in that the severity label corresponding to the full fundus image is assigned to every input patch. This inaccurate label assignment will confuse the network training, making it difficult to distinguish patches with and without DR. To tackle this problem, in our method we generate accurate patch-wise labels to supervise a different classification task. Specifically, the ground truth label of our classification branch is the types of lesions that each input patch contains, which can be derived from the given segmentation masks. For example, a label of `0110` means that the patch has no MAs, has HEs and SEs, and has no EXs. Our method is the first one to adopt this type of classification supervision to assist a fundus lesion segmentation task. With this design, each input patch has an accurate patch-wise label rather than a weak image-wise label as used in the previous works. Obviously, more accurate labels lead to better classification network training, which better assist the multi-task learning.

Our contributions are summarized as follows:

- We propose a novel collaborative strategy that integrates contextual and local patch information using an attention mechanism to segment multiple types of DR-related fundus lesions simultaneously.
- A classification branch with a novel supervision scheme is proposed and jointly trained with the fundus segmentation branch. The new multi-task learning strategy further improves the segmentation accuracy and robustness.
- We validate our method using three different fundus segmentation datasets. Comprehensive experimental results demonstrate that our method achieves state-of-the-art performance.

The remainder of this paper is organized as follows. Section 2 briefly surveys recent studies on fundus lesion segmentation. Section 3 explains details of our proposed method with a collaborative network and a multi-task learning scheme. Experimental results that validate the proposed method are presented and discussed in Section 4. Lastly, the paper is concluded in Section 5.

2. Related work

In this section, we review some closely related studies in the literature, and also explain the major differences of our proposed method comparing to the existing approaches.

2.1. Contextual and local information integration

Benefiting from deep learning technology, research on fundus lesion classification and segmentation of regions of interest (e.g., retinal lesion and vessel) has been extensively studied in recent years [3–9,11,14–21]. Existing methods can be divided into three categories based on whether global or contextual information is integrated with local information.

Global Information Only. Since fundus images usually have very high resolution, it is not feasible to put the images directly into a deep learning model due to the computational burden. One commonly adopted solution is to downsample the original images first, and then use the downsampled images as model inputs. Guo et al. [3] proposed a deep supervised neural network (L-seg) that fused multi-scale features to segment four types of fundus lesions. In their work, the fundus images were downsampled to about one-ninth of the original size (from 4288×2848 to 1440×960). Image details including small lesions are easily lost due to the downsampling and max-pooling operations. Similar to [3], Li *et al.* [21] achieved fundus lesion classification by downsampling images to three different resolutions and using a lesion attention module to fuse these features. Some other studies [4,9] also took low-resolution images as the inputs of their deep learning models. Although global and multi-scale information can be captured well, the lack of detailed image information hinders the performance of these methods.

Local Information Only. To preserve image details, some studies first crop the high-resolution images into small patches, and then use the cropped patches as their model inputs. For example, a fully convolutional neural network architecture with inception modules was proposed in [5] for exudate segmentation and the input data were patches with a size of 32×32. A GoogLeNet architecture was utilized to predict five types of fundus lesions using patches of size 128×128 in [6]. Zheng et al. [11] fed patches of 48×48 pixels into a U-Net ensemble network for exudate detection, in which the predicted feature maps were extracted using four different kernel sizes, *i.e.*, 3×3, 4×4, 5×5, and 6×6. Xiao et al. [15] incorporated the HED network [22] into a conditional generative adversarial network and used a patch size of 128×128 for SE, EX, and HE segmentation, and 64×64 for MA segmentation. RGB patches of 480×480 pixels were taken as inputs for a cascaded deep residual network for exudate segmentation in [14]. Li et al. [20] achieved the retinal vessel segmentation by first segmenting the row images into patches of 64 × 64 pixels and then inserting an attention module into the basic U-Net. Although these methods keep detailed information as much as possible, global contextual information is lost, which makes the segmentation problem more difficult and often results in inconsistency across nearby patches.

Contextual Information Integrated with Local Patches. Some researches have adopted a collaborative strategy combining global contextual information and local patch information in order to exploit their respective advantages. Sarhan et al. [7] trained two separate networks, each corresponding to a specific image scale, *i.e.*, 1× and 0.5× magnifications. The multi-scale feature representations from the two networks, however, were not combined or exchanged during model training. To fully integrate global and local information, Yan et al. [8] proposed a collaborative network that trained two network branches together, taking the downsampled full image and cropped local patches as respective inputs. In this framework, feature maps from the global branch were cropped and concatenated with the feature maps derived from the local patches, and then sent to a sigmoid layer for segmentation prediction. Although features from multiple scales were combined, there was a lack of high-resolution information in the global branch because of the downsampling operation (from 2848×4288 to 640×640). In addition, the concatenation between the two branches was only performed at the end of the decoding path while feature maps from intermediate layers were not combined. To address these shortcomings and more effectively integrate contextual and local information, we propose a new collaborative strategy that differs from existing studies in two aspects: (1) an effective attention mechanism is leveraged to fuse feature maps from all decoder layers between the contextual and local branches; (2) the contextual branch uses patches of size twice as large as those of the local branch rather than performing any downsampling operations on the original full-resolution fundus images.

Note that there are a couple of other methods in the literature that also perform multi-scale data processing and adopt some attention mechanisms. In [23], a deep neural network with both global and local branches was designed for thorax disease classification. An attention map was computed by the global branch to select the most discriminative regions from the original image, which were further processed by the local branch. In [24], a global–local network was proposed for facial expression recognition. The global branch aimed to encode full face information whereas the local branches were targeted at various local regions. The attention mechanism was applied to each branch separately to focus the network on non-occluded regions of the face image. In contrast to these existing studies, our method applies the selective kernel module (SKM) attention unit to effectively aggregate feature information across the contextual and local scales, and the adaptive feature integration is computed at every resolution level of the network.

2.2. Classification combined with segmentation tasks

Many studies on fundus lesion segmentation have only used the lesion masks as ground truth labels [3,4,11,8], which belong to the category of *single-task learning*. Another popular approach is *multi-task learning*, which cooperatively uses both segmentation and classification labels to train deep neural networks. The two tasks can benefit from each other and lead to better segmentation (and classification) performance [13,14,9,16,10]. Since the presence of certain lesions, *e.g.*, MAs and HEs, directly reflects the disease severity level of DR, many studies have designed multi-task learning by combining a fundus lesion segmentation task and a disease severity classification task [13,14]. For example, He et al. [13] developed a diabetic macular edema (DME) grading model using an XGBoost classifier, which produced macular and EX segmentation as by-products. As closely related, Mo et al. [14] proposed a cascaded framework for DME recognition, where the final classification was generated based on highly accurate segmentation results from the previous stage.

Architectures introduced in [9,16,10] are more similar to our proposed method in this paper. Guo et al. [9] proposed a lightweight neural network for EX segmentation, in which a classification branch was added after the contextual information encoder to differentiate whether the input image had DR or not. A similar binary classification branch was also used in [16] to classify whether an input patch was pathological. Haloi et al. [10] suggested an end-to-end cooperative learning strategy to segment fundus lesions with the help of a classification of five DR severity levels.

Our method is in line with the previous studies and also introduces a classification branch sharing the same encoder with the segmentation branch. But different from the studies mentioned above, we propose to use the types of lesions that each input patch contains as the supervision label instead of using the disease severity grade information. The disease grading depends on an entire fundus image, which is not accurate as a per-patch label since not every patch is DR-related. On the contrary, our classification label is directly applicable for every patch, which can be simply derived from the lesion segmentation masks. From this perspective, the classification sub-task in our method can be considered as accurately or fully supervised while the previous methods are weakly supervised. Obviously, the performance of a deep learning model highly depends on the accuracy of the training labels. Our fully supervised labels provide more precise supervision information than the weak labels; thus better segmentation results can be expected, as verified by our experimental results.

3. Methods

In this section, we first detail the design of the proposed contextual branch and local branch. Then, the collaboration of the two branches through an attention mechanism is described (see Fig. 2(b)), which is the major contribution of our method. Finally, an auxiliary classification branch is explained, which further improves segmentation accuracy through multi-task learning.

3.1. Collaboration of contextual and local branches

Contextual Branch. A deep encoder-decoder-like architecture [26–29] is adopted as the backbone for the contextual branch, as shown in Fig. 2(a). The encoder aims to capture low-level features (*e.g.*, edges and textures) in shallow layers and high-level features (*e.g.*, shapes and objects) in deeper layers, whereas the decoder inversely maps back the extracted features to the same size of the input data to reconstruct the segmentation mask. In our settings, SE-ResNeXt-50 [30] (see Fig. 3) serves as the encoder of the contextual branch. ResNeXt-50 [31] is a highly modular architecture that aggregates a number of residual transformations with the same topology. The multi-pathway inception modules inside ResNeXt-50 make the network wider and significantly enlarge its capacity without increasing the model complexity. Based on ResNeXt-50 [31], the squeeze-and-excitation modules are additionally added to SE-ResNeXt-50 [30], which enable the network to recalibrate channel-wise features dynamically, thereby strengthening the representation power of the total network.

For the decoder of the contextual branch, we introduce a new deconvolution operation, named *Conv2d-BN-SKNet-TransConv2d* (bold red arrows in Fig. 2(a)), between adjacent resolution blocks. With this operation, the size of the feature map is doubled while the number of channels is reduced by half. Within each deconvolution operation, a SKNet [25] is embedded to adaptively learn informative features using different sizes of kernels, which effectively extracts feature representations from multiple receptive fields. Following [26], features from each layer of the encoder are concatenated with features from the corresponding decoder layer (dashed lines in Fig. 2(a)). Leveraging such skip connections, feature maps containing low-level spatial information from the encoder can be effectively integrated into the decoder path. Additionally, inspired by the deep supervision strategy [22], we apply multiple supervisions for intermediate layers of the decoder. Specifically, bilinear upsampling is first applied to resize the output of each intermediate layer to the same size as the ground truth mask, and then all intermediate outputs are concatenated and mapped to the final segmentation output.

Local Branch. The local branch shares a similar architecture with the contextual branch. The only difference between the two branches lies in the input data. The inputs for the local branch are patches cropped from the original high-resolution fundus images with a size of 256×256 pixels while the inputs for the contextual branch are bigger patches centered around the local patches with a size of 512×512 pixels, as illustrated in Fig. 2(b).

Branch Collaboration. Branch collaboration is the major novelty proposed in this work, which aims to combine the contextual information extracted from the contextual branch and the detailed local information from the local branch. As mentioned above, the input of the local branch carries fine-gained fundus details, whereas, the input of the contextual branch contains a larger receptive field and is reduced to a lower resolution, which helps capture features with context information. The aggregation of features generated in both branches facilitates multi-scale feature representations. To this end, we design an attention mechanism to aggregate feature maps from the decoder layers in the contextual branch with the corresponding feature maps in the local branch (see Fig. 2(b)). We adopt the SKM [25] to learn the attentions for better feature map fusion. Applying SKM with its multi-scale kernels can dynamically learn the most relevant and informative features, *i.e.*, the attentions, from the contextual branch, which are then fed into the local branch to improve the final segmentation accuracy. SKM assigns different importance weights for feature maps generated by different sized kernels, *i.e.*, 3×3, 5×5, and 7×7.

To train the collective contextual and local segmentation network, we use a loss function that is a sum of the *Dice* loss [32] and a *weighted multi-class cross-entropy (wMCE)* loss, which are defined as follows:

$$\mathcal{L}_{Dice} = \sum_{c=1}^{N}\left(1 - \frac{2\sum_{c} y_{o,c}p_{o,c} + \varepsilon}{\sum_{c} y_{o,c} + \sum_{c} p_{o,c} + \varepsilon}\right), \tag{1}$$

$$\mathcal{L}_{wMCE} = -w_c\sum_{c=1}^{N+1} y_{o,c}log(p_{o,c}), \tag{2}$$

where $y_{o,c}$ is 1 if class label c is the correct classification label for observation o, and $y_{o,c}$ is 0 otherwise. $p_{o,c}$ represents the predicted probability of observation o belonging to class c, N denotes the number of lesion types, and N + 1 is used to indicate the background class. ε is a small positive number for improving numerical stability. w_c is a hyperparameter that controls the weights of each lesion type as well as the background. Empirically, the weights are set to 1, 2, 2, 2, 2 for the background, HE, MA, EX, and SE, respectively.

3.2. Classification branch incorporation

In our proposed method, a classification head is further incorporated into the segmentation network in both the contextual and the local branches to construct a multi-task learning paradigm, which is implemented by hard parameter sharing. Specifically, the segmentation branch and the classification head share the same encoder (*i.e.*, SE-ResNeXt-50). We choose image classification as an auxiliary task for the following two reasons. First, semantic segmentation and image classification are highly correlated tasks [33,34], which have proper feature dependencies in the latent space. For example, the shape properties of lesions are beneficial for both the classification and segmentation tasks [34]. Thus, the feature representations jointly trained by the two tasks help the network to learn features common to both tasks and reduce the problem of overfitting. Second, the classification task enables the segmentation network to focus on the surrounding anatomy for the feature learning of each pixel, thus improving the segmentation performance.

But different from existing studies, the types of lesions that each input patch contains are used as the supervision label. Specifically, we use one binary variable to represent one lesion type, and the supervision label is thus a 4D binary vector. For instance, if an input patch contains MA, EX, and HE but no SE, its ground-truth classification label will be the vector of `1110`. Here, the three `1`s denote the presence of MA, EX, and HE, and `0` represents the non-existence of SE. With this approach, each input patch has an accurate patch-wise label rather than a weak image-wise label as in previous studies. The standard binary cross-entropy loss is used to train the classification branch.

4. Experimental results and discussions

In this section, we first summarize three publicly available datasets that are used to validate the proposed method. We then show detailed lesion segmentation results comparing our method with alternative strategies and demonstrate the benefits of the proposed improvements.

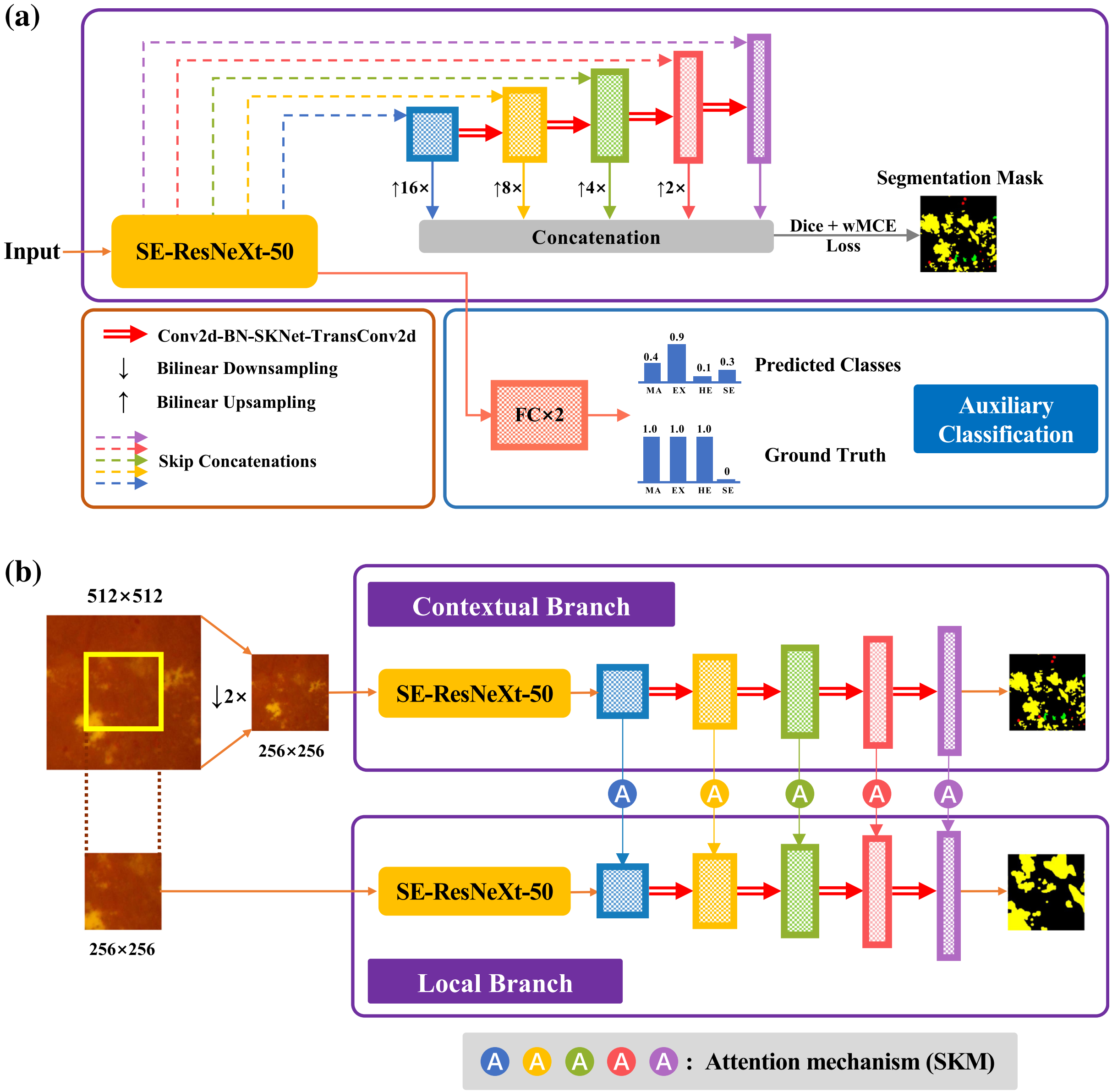


Fig. 2. Illustrations of the network architecture of the contextual and local branches (as shown in (a)) and their collaboration scheme (as shown in (b)). The contextual and local branches are designed with the same network structures but different inputs and parameters. In (a), SE-ResNeXt-50 is used as the encoder, and feature maps from different encoder layers are concatenated with corresponding ones in the decoder layers. In the auxiliary classification part, the predicted label (lesion types each input patch contains) is generated by a subnetwork with two fully-connected (FC) layers. In (b), feature maps from decoder layers in the contextual branch are concatenated with corresponding feature maps in the local branch through an attention mechanism (SKM [25]). For clarity, we omit in this figure the classification branch, deep supervision, and the skip concatenations that are detailed in (a).

4.1. Datasets and experimental setup

Indian Diabetic Retinopathy Image Dataset (IDRiD). The IDRiD [35] was released for the segmentation and disease grading of fundus images in a 2018 ISBI grand challenge. In this work, we focus only on the fundus lesion segmentation part. The dataset consists of 81 images with four types of retinal lesions, *i.e.*, MA, HE, EX, and SE, and each image has a resolution of 4288×2848 pixels. The split of the training and testing sets follows the same settings as in the challenge. Since the ground truth labels for the test set have been made publicly available, we can use them directly to evaluate the performance of different methods.

E-ophtha Dataset. This dataset [36] consists of two separate sub-datasets named e-ophtha-MA and e-ophtha-EX with lesion masks annotated by expert ophthalmologists. The e-ophtha-EX dataset contains 47 images with exudates and 35 normal images. The e-ophtha-MA dataset has 148 images with MAs and 233 normal images. There are total of 21 images containing both EXs and MAs combining these two sub-datasets, which are used for testing our multi-class lesion segmentation method. The reason why we only consider images containing both MAs and EXs is that our proposed method aims to segment multi-class lesions simultaneously, same as [3]. We randomly split the 21 images into a training set of 15 images and a testing set of 6 images.

DDR Dataset. The DDR dataset[3] is a high-quality dataset released in [3]. There are total of 757 fundus images with lesion segmentation annotations, in which 601 images have HEs, 570 images

[3] https://github.com/nkicsl/DDR-dataset

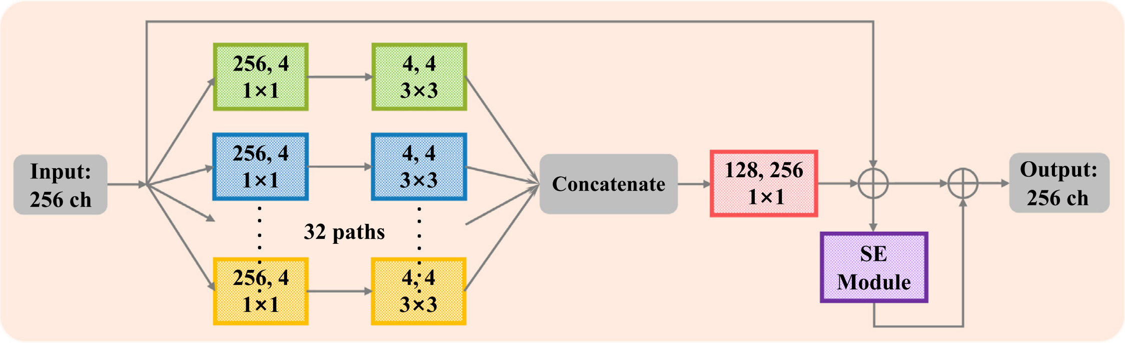


Fig. 3. A block of SE-ResNeXt-50 [30]. Numbers in each box represent the number of input channels, the number of output channels, and the convolutional kernel size, respectively. SE Module: Squeeze-and-Excitation modules; ch: channel.

have MAs, 486 images have EXs, and 239 images have SEs. The full dataset is divided into three subsets of 383, 149, and 225 fundus images for training, validation, and testing, respectively, following the same data split criterion of [3].

Experimental Setup. Images from all three datasets are first pre-processed with the following steps.

- The Otsu method [37] is applied to the original fundus images to remove non-tissue regions.
- All images are cropped into local patches of size 256×256 with an overlap of 128 pixels between adjacent patches. These small patches are used as the inputs to the local branch of our model. Centered around each of these patches, a larger image patch of size 512×512 is correspondingly cropped to serve as the inputs to the contextual branch.
- Real-time data augmentation methods including horizontal flip, vertical flip, random rotation, shift, rescaling, and cropping are performed at model training to reduce overfitting and make the models robust with respect to these geometric perturbations.

In this work, a mini-batch of size 16 is adopted for model training and the Adam optimizer [38] is used as the optimization method. The initial learning rate is set to 3e-4 and reduced by a factor of 10 at the 20th and the 50th epoch, with a total of 60 training epochs. All models are implemented using the PyTorch package [39] and all experiments are performed on a workstation equipped with a 24 GB memory NVIDIA Tesla P40 GPU card. We compute the computational complexity using the model parameters and FLOPs (floating-point operations) as they are widely used in the field of deep learning [31,40]. The number of FLOPs and parameters are 20.58×10^9 and 92.19 M, respectively.

During inference, the collaboration scheme and multi-task learning are still required, thus, the used network structure keeps the same as that used in the training stage as shown in Fig. 2(b). The final segmentation results take that generated in the local branch. The *area under the precision and recall curve* (AUC) is adopted as the evaluation metric, following the same settings as the IDRiD challenge [41]. Higher AUC values mean higher precision and recall, and better segmentation accuracy.

4.2. Ablation study

We conduct ablation studies to evaluate the contribution of each component in our method. The experimental results are summarized in Table 1 and some corresponding visualized segmentation results are presented in Fig. 4. The same set of hyper-parameters are used to compare different models, including the dataset setup, the batch size, the optimizer, and the training schedule, as summarized in the previous section.

We first investigate the performance of the local and contextual branches, respectively. In Table 1 and Fig. 4, these methods marked with $*$ denote the segmentation results obtained from the contextual branch, and results of the other methods are obtained from the local branch. By comparing the results of six methods (i.e., *loc* and *cont**, *loc + cl* and *cont + cl**, and *loc + cont + att + cl** and *loc + cont + att + cl*), it is seen that AUC scores of the local network are higher than those of the contextual network across all three datasets. For example, the mean AUCs obtained from the local branch are higher around 2% in IDRiD and 4% in E-ophtha than those in the contextual branch. This shows that detailed image information is very important for fundus lesion segmentation, since a large number of fundus lesions are small in size, mostly occupying only one to several pixels. Therefore, capturing detailed image information enables more precise localization and segmentation of fundus lesions.

We then investigate the advantage of combining the contextual and local branches. We first test a conventional strategy that simply concatenates corresponding feature maps from the two branches without using an attention scheme. It can be seen that this combination (*loc + cont*) does not consistently outperform the *loc* or *cont* network even though the mean AUCs are higher. This shows that the simple concatenation cannot effectively aggregate informative features from the local and contextual branches. In contrast, the employment of the attention mechanism as proposed in our method (*loc + cont + att*) leads to more significant improvements in AUC across all fundus lesions for all three datasets. In particular, the AUC scores of HE and MA in IDRiD and the AUC score of HE in DDR increased by around 2.5%, and the AUC scores of SE in IDRiD, MA in E-ophtha, MA and EX and SE in DDR increased by more than 1%. The results not only verify the effectiveness of the collaborative strategy, but also indicate the importance of the attention scheme, *i.e.*, dynamically assigning different weights to the contextual and local features. A similar phenomenon can also be seen in the segmentation results from the contextual branch (*i.e.*, *cont + cl** and *loc + cont + att + cl**), which indicates that our collaborative model also improves the segmentation performance of the contextual branch.

The benefit of the classification branch can be observed by comparing the results of six methods (i.e., *loc* and *loc + cl*, *cont** and *cont + cl**, and *loc + cont + att* and *loc + cont + att + cl*). It can be seen that the incorporation of the classification branch further improves the accuracy of lesion segmentation (approximately 1% in each paired comparison), which is in line with our expectation that multi-task learning can improve learning efficiency and help reduce the problem of overfitting. As we know, classification and

Table 1
Ablation study. loc: only the local branch is used; loc + cl: only the local branch with auxiliary classification are used; cont*: only the contextual branch is used; cont + cl*: only the contextual branch with auxiliary classification are used; loc + cont: collaborative model with both contextual and local branches, but corresponding feature maps are simply concatenated; loc + cont + att: collaborative model with both contextual and local branches, and feature maps from both branches are integrated using the SKM attention mechanism; loc + cont + att + cl*: collaborative model with the auxiliary classification (the prediction results are obtained from the contextual branch); loc + cont + att + cl: collaborative model with the auxiliary classification (the prediction results are obtained from the local branch). AUC is used as the evaluation metric.

Method	IDRiD					E-ophtha			DDR				
	HE	MA	EX	SE	mean	MA	EX	mean	HE	MA	EX	SE	mean
loc	0.629	0.528	0.822	0.621	0.650	0.494	0.693	0.594	0.550	0.375	0.686	0.619	0.557
loc + cl	0.638	0.532	0.825	0.637	0.658	0.510	0.705	0.608	0.557	0.386	0.698	0.621	0.566
cont*	0.614	0.495	0.812	0.606	0.632	0.432	0.669	0.550	0.545	0.377	0.678	0.613	0.553
cont + cl*	0.619	0.505	0.820	0.618	0.641	0.448	0.680	0.564	0.550	0.383	0.685	0.617	0.559
loc + cont	0.633	0.505	0.821	0.648	0.652	0.508	0.717	0.613	0.548	0.367	0.694	0.615	0.556
loc + cont + att	0.660	0.529	0.824	0.659	0.668	0.518	0.720	0.619	0.575	0.378	0.705	0.625	0.571
loc + cont + att + cl* (Ours)	0.642	0.499	0.824	0.652	0.654	0.467	0.690	0.578	0.570	0.349	0.698	0.629	0.561
loc + cont + att + cl (Ours)	**0.661**	**0.546**	**0.827**	**0.672**	**0.677**	**0.536**	**0.722**	**0.629**	**0.586**	**0.386**	**0.713**	**0.638**	**0.581**

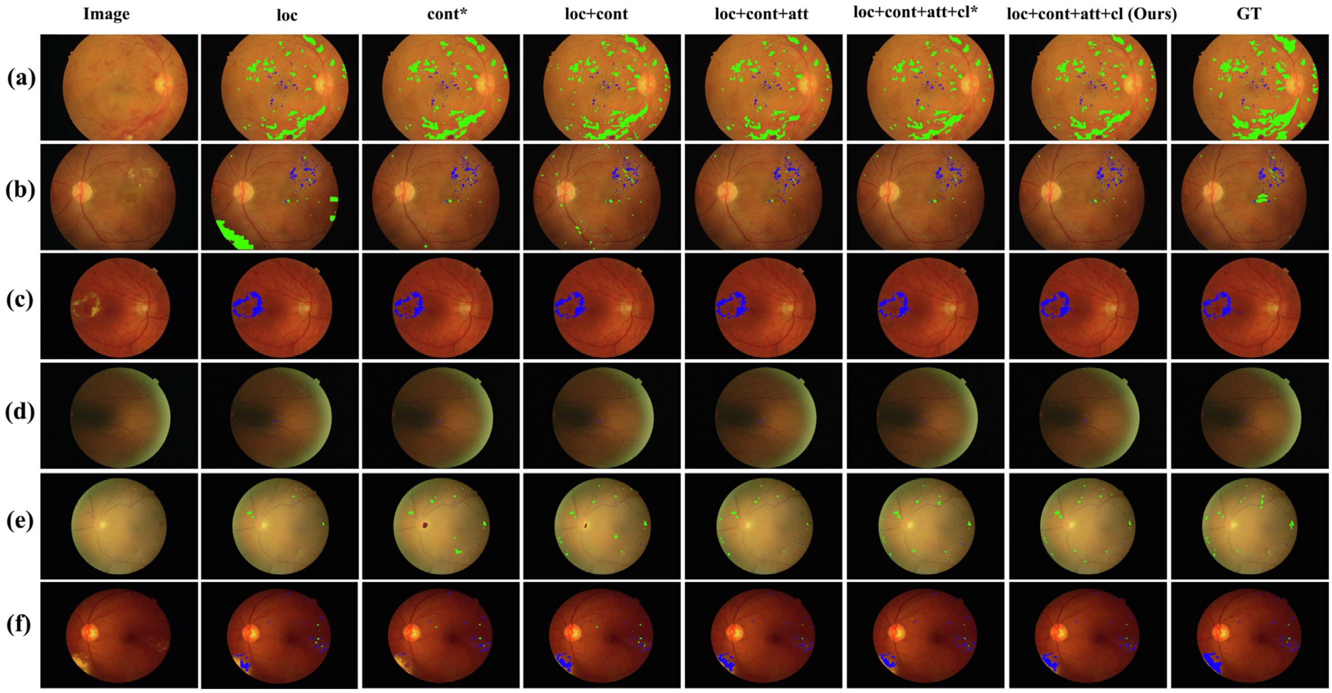


Fig. 4. Comparison of visualized segmentation results among different methods in the ablation study. Two examples are taken from each dataset: (a, b) from IDRiD, (c, d) from E-ophtha, and (e, f) from DDR. The green, red, blue, and brown colors represent HE, MA, EX, and SE, respectively.

segmentation are good complementary tasks since the former normally relies on high-level semantic information and the latter focuses more on local details. We acknowledge that there could be alternative ways to design the auxiliary task to assist the segmentation problem. Exploring this would be an interesting future research direction.

It can be seen that while each of the components results in a relatively low-performance gain, their combination brings a much larger improvement. For example, compared to the general method using only the local branch (*loc*), our full method *loc + cont + att + cl* improves the average AUC by about 2.7% in IDRiD, 3.5% in E-ophtha, and 2.4% in DDR.

4.3. Comparison with some baseline networks

In this experiment, eight deep learning image segmentation models, *i.e.*, SegNet [27], Attention U-Net [42], DeepLab V3+ [28], UNet++ [43], DUpsample [44], CE-Net [45], CPFNet [46], and L-seg [3], are adopted as the baseline models to compare our method against. SegNet is one of the earliest proposed deep encoder-decoder models for multi-class pixel-wise image segmentation [27]. Attention U-Net proposed an attention gate module to scale input features according to the specific task [42]. DeepLab V3+, UNet++, and CE-Net considered multi-scale information. To achieve this, DeepLab V3 + and CE-Net both used multiple convolution branches with different respective fields, and UNet++ inserted a series of nested dense convolutional blocks to integrate multi-scale features across the encoder and the decoder. DUpsample has been proposed more recently to improve the upsampling operation in decoding layers, which can recover pixel-wise prediction from low-resolution ones. This method has shown much superior performance than many state-of-the-art algorithms [47–49]. CPFNet designed a global pyramid guidance module and a scale-aware pyramid fusion to achieve global feature capture and multi-scale context information integration, respectively [46]. It has presented a good generalization ability on different medical image segmentation tasks. L-seg [3] is the first and latest method to segment all four types of DR lesions simultaneously using an end-to-end framework, and thus we can conduct a direct comparison with the L-seg method. Its segmentation results on the three datasets are directly copied from its corresponding publication [3].

Table 2
Comparison with some baseline models.

Method	IDRiD					E-ophtha			DDR				
	HE	MA	EX	SE	mean	MA	EX	mean	HE	MA	EX	SE	mean
SegNet [27]	0.416	0.324	0.655	0.405	0.450	0.164	0.454	0.309	0.419	0.279	0.565	0.427	0.422
Attention U-Net [42]	0.466	0.476	0.780	0.475	0.549	0.332	0.511	0.422	0.343	0.287	0.586	0.342	0.389
DeepLab V3+ [28]	0.513	0.479	0.788	0.511	0.573	0.431	0.461	0.446	0.224	0.327	0.574	0.259	0.346
UNet++ [43]	0.478	0.484	0.801	0.525	0.572	0.258	0.553	0.405	0.385	0.318	0.581	0.354	0.409
DUpsample [44]	0.556	0.518	0.814	0.600	0.622	0.373	0.580	0.477	0.491	0.359	0.663	0.525	0.510
CE-Net [45]	0.588	0.526	0.810	0.616	0.635	0.457	0.607	0.532	0.507	0.340	0.653	0.514	0.504
CPFNet [46]	0.600	0.523	0.816	0.590	0.632	0.444	0.652	0.548	0.520	0.373	0.677	0.540	0.528
L-Seg [3]	0.637	0.463	0.795	**0.711**	0.652	0.169	0.417	0.293	0.359	0.105	0.555	0.265	0.321
Ours	**0.661**	**0.546**	**0.827**	0.672	**0.677**	**0.536**	**0.722**	**0.629**	**0.586**	**0.386**	**0.713**	**0.638**	**0.581**

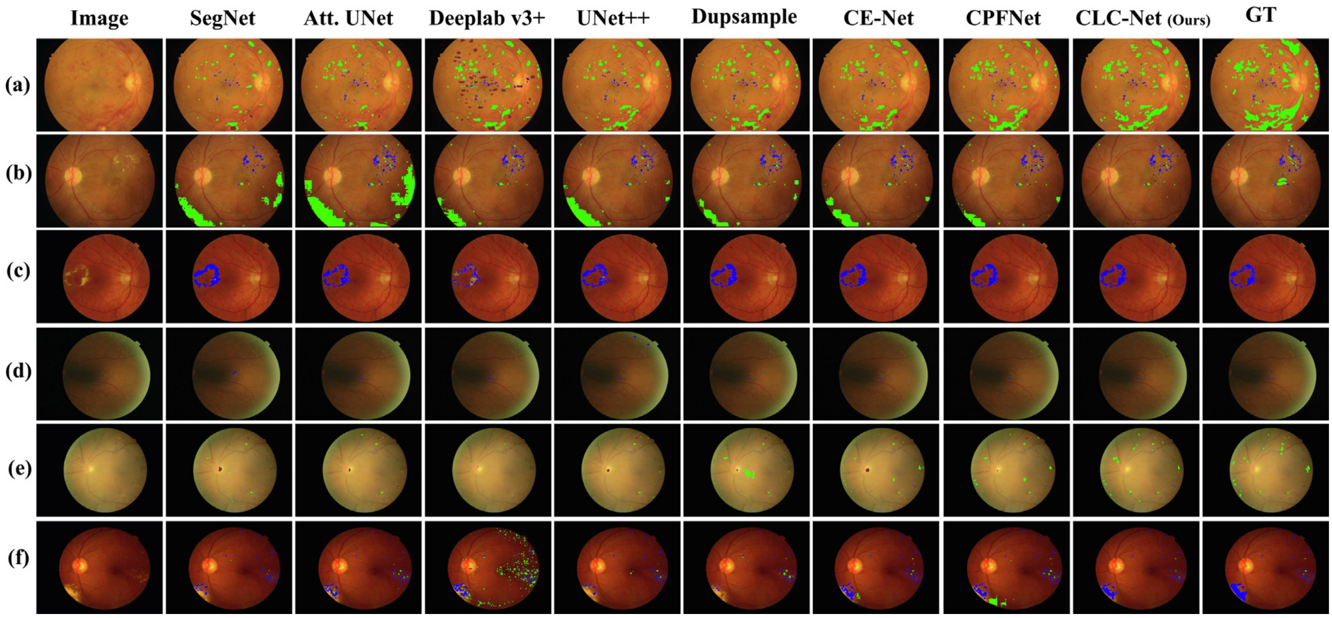


Fig. 5. Visualization of sample DR lesion segmentation results comparing our proposed method with seven baseline models, including, SegNet, Attention U-Net, DeepLab V3+, UNet++, DUpsample, CE-Net, and CPFNet (L-seg has not released their source code). Two examples are taken from each dataset: (a, b) from IDRiD, (c, d) from E-ophtha, and (e, f) from DDR. The green, red, blue, and brown colors represent HE, MA, EX, and SE, respectively.

The experimental results for this comparison study are summarized in Table 2, and some examples are visualized in Fig. 5. Overall, SegNet performs the worst on two datasets. The main reason is that SegNet does not use skip connections to help propagate high-frequency features to the decoder, whereas spatial information of small lesions (*e.g.*, the MAs) can easily get lost at deeper layers due to the convolution and max-pooling operations. The loss of high-frequency features and the use of un-pooling operations in SegNet are also likely the causes that the predicted EX segmentations appear to be dilated comparing to the ground truth. In addition, we also find that MAs are sometimes mis-recognized as HEs by SegNet, as both lesions have low image intensities and the differences between them cannot be well distinguished by the network. Compared with SegNet, Attention U-Net adds an attention weight matrix to suppress irrelevant regions and enhance salient features, which improves the segmentation performance by around + 9.9% on the IDRiD and + 11.3% on the E-ophtha datasets.

DeepLab V3+, UNet++, and CE-Net aggregate multi-scale information from the image, which enables them to better capture features from both large and small lesions. The benefits of multi-scale feature aggregation can be clearly seen in Rows #3, 4, and 6 of Table 2. It is seen that DeepLab V3 + and UNet++ show more than 12% higher mean AUC values than SegNet on IDRiD, and 13.7% and 9.6% higher respectively on E-ophtha dataset. CE-Net further improves the segmentation performance by around 18.5% higher mean AUC values than SegNet on IDRiD. These observations prompt us to fully integrate multi-scale information at all intermediate layers in our collaborative network design. We notice that a much lower accuracy for DeepLab V3 + is reported in [3] than we get in Table 2. Through our experiment, we find that this is mainly due to the downsampling operation performed in [3], where an original fundus image is first resized to about one-ninth of the original size before being fed into the network. As explained earlier in Section 2, image details are easily lost due to downsampling, which can limit the accuracy that can be achieved.

DUpsample model replaces upsampling with a trainable deconvolution to exploit the correlation among pixel-wise features in each decoder layer. The lower performance on MA is observed, which is likely due to the lack of skip connections between the encoder and the decoder. Different from DUpsample, our method utilizes a multi-scale feature aggregation strategy to enhance feature extraction of small lesions (*i.e.*, the MAs), and presents even superior performance. CPFNet fully captures context information by reconstructing the skip connection and parallel dilated convolutional filtering, which represents the suboptimal performance across the three datasets. L-seg [3] reports a related work as us, thus it is used for direct comparison. It is seen that, except for SE segmentation in the IDRiD dataset, our method outperforms L-

Table 3
Comparison with the top ten teams and another new segmentation method on IDRiD challenge.

Model (rank)	HE	MA	EX	SE	mean
VRT (1st)	**0.680**	0.495	0.713	**0.700**	0.647
PATech (2nd)	0.649	0.474	**0.885**	-	-
iFLYTEK-MIG (3rd)	0.559	0.502	0.874	0.659	0.648
SOONER (4th)	0.540	0.400	0.739	0.537	0.554
SAIHST (5th)	-	-	0.858	-	-
lzyuncc_fusion (6th)	-	-	0.820	0.626	-
SDNU (7th)	0.457	0.411	0.502	0.537	0.477
CIL (8th)	0.489	0.392	0.755	0.502	0.535
MedLabs (9th)	0.371	0.340	0.786	0.264	0.440
AIMIA (10th)	0.328	0.379	0.766	0.273	0.437
SMA [16]	0.530	-	0.821	-	-
Ours	0.661	**0.546**	0.827	0.672	**0.677**

seg in all four lesions and all three datasets. Aside from network design, this is likely due to the input scheme, i.e., the downsampling of the original images to a smaller size.

Overall, as shown in Table 2, our proposed method outperforms all eight baseline models to a large extent. In particular, compared with the previous best-performed method, the mean AUC is improved by more than 2.5% for the IDRiD dataset, 8.1% for E-ophtha, and 5.3% for DDR. Fig. 5 also shows that the results of our method match the ground truth much better. However, from these visualized segmentation results, it can be observed some failure cases, which miss some MA regions, *e.g.*, case (a), (b), (d), and (e). The reason can be explained by their extremely small lesion sizes, scattered distribution, and coexisting multiple lesions. That is, each MA region usually occupies a few pixels and these MA regions are scattered and distributed spanning the entire image. Moreover, the MA is usually accompanied by other lesions (*e.g.*, HE and EX), which further increases the difficulty of accurate segmentation.

4.4. Comparison with the state-of-the-art methods on the IDRiD dataset

Since the IDRiD dataset comes from an open competition, we can compare our method with the top-performing solutions in the same competition, and the results are summarized in Rows #1-#10 of Table 3. When comparing each lesion type separately, our method competes favorably, ranking 1st for MA segmentation, 2nd for HE and SE segmentation, and 4th for EX segmentation. In addition, according to published reports, the top two solutions, *VRT* and *PATech*, target at single type lesion segmentation. Different models are used for the segmentation of different types of lesions and a lot of hyper-parameter tuning is required. In contrast, our method can segment four types of lesions simultaneously, which is much more efficient and easy to use.

Another recently proposed method [16] also reports results on the IDRiD dataset, which is included in Table 3 for comparison. In [16], an auxiliary classification task is also designed, which targets at predicting the presence of lesions instead of precise lesion type classification. Only segmentation accuracies for HE and EX are reported in [16]. Clearly, the performance of our proposed method is superior to that of [16]. Specifically, the AUC scores of HE and EX are improved by + 13.1% and + 0.6% respectively using our method.

5. Conclusion

In this study, we propose a new collaborative neural network method that effectively aggregates contextual and local image information using an attention mechanism for automatic DR lesion segmentation in retinal images. Multi-task learning using a novel classification design is also introduced to further boost lesion segmentation accuracy and robustness. Extensive experiments have demonstrated that our proposed method produces superior segmentation results comparing to other state-of-the-art methods in the literature. Despite the improved performance, this work still has some limitations. First, these tiny lesions are still difficult to identify as shown in our visualization results. For these hard cases, a large number of fine-grained annotations can help improve performance, however, pixel-level annotations are costly to obtain. Future work will seek to further improve the accuracy of lesion segmentation by integrating self-supervised learning with unlabeled images for better feature representations. Second, our method lacks clinical validation on large-scale data to verify its generalization across different cohorts. Future work will collect more fundus images to evaluate the segmentation algorithm in prospective clinical trials.

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.

Acknowledgments

This research was in part funded by the National Natural Science Foundation of China (No. 61571314) and Science & technology department of Sichuan Province (No. 2020YFG0081).

References

[1] M.M. Engelgau, L.S. Geiss, J.B. Saaddine, J.P. Boyle, S.M. Benjamin, E.W. Gregg, E. F. Tierney, N. Rios-Burrows, A.H. Mokdad, E.S. Ford, et al., The evolving diabetes burden in the united states, Ann. Intern. Med. 140 (11) (2004) 945–950.

[2] Y. Zheng, M. He, N. Congdon, The worldwide epidemic of diabetic retinopathy, Indian J. Ophthalmol. 60 (5) (2012) 428–431.

[3] 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.

[4] S. Guo, K. Wang, H. Kang, T. Liu, Y. Gao, T. Li, Bin loss for hard exudates segmentation in fundus images, Neurocomputing 392 (2019) 314–324.

[5] P. Chudzik, S. Majumdar, F. Caliva, B. Al-Diri, A. Hunter, Exudate segmentation using fully convolutional neural networks and inception modules, Medical Imaging 2018: Image Processing, Vol. 10574, International Society for Optics and Photonics, 2018, p. 1057430.

[6] C. Lam, C. Yu, L. Huang, D. Rubin, Retinal lesion detection with deep learning using image patches, Invest. Ophthalmol. Visual Sci. 59 (1) (2018) 590–596.

[7] 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.

[8] 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, arXiv preprint arXiv:1901.06047 (2019).

[9] S. Guo, T. Li, K. Wang, C. Zhang, H. Kang, A lightweight neural network for hard exudate segmentation of fundus image, International Conference on Artificial Neural Networks, Springer (2019) 189–199.

[10] M. Haloi, Rethinking convolutional semantic segmentation learning, arXiv preprint arXiv:1710.07991 (2017).

[11] R. Zheng, L. Liu, S. Zhang, C. Zheng, F. Bunyak, R. Xu, B. Li, M. Sun, Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network, Biomed. Opt. Exp. 9 (10) (2018) 4863–4878.

[12] J. Baxter, A bayesian/information theoretic model of learning to learn via multiple task sampling, Mach. Learn. 28 (1) (1997) 7–39.

[13] X. He, Y. Zhou, B. Wang, S. Cui, L. Shao, DME-Net: Diabetic macular edema grading by auxiliary task learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention Springer, 2019, pp. 788–796.

[14] J. Mo, L. Zhang, Y. Feng, Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks, Neurocomputing 290 (2018) 161–171.

[15] Q. Xiao, J. Zou, M. Yang, A. Gaudio, K. Kitani, A. Smailagic, P. Costa, M. Xu, Improving lesion segmentation for diabetic retinopathy using adversarial learning, in: International Conference on Image Analysis and Recognition Springer, 2019, pp. 333–344.

[16] 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.

[17] L. Zhang, S. Feng, G. Duan, Y. Li, G. Liu, Detection of microaneurysms in fundus images based on an attention mechanism, Genes 10 (10) (2019) 817–823.

[18] P. Prentašić, S. Lončarić, Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion, Comput. Methods Programs Biomed. 137 (2016) 281–292.

[19] J.H. Tan, H. Fujita, S. Sivaprasad, S.V. Bhandary, A.K. Rao, K.C. Chua, U.R. Acharya, Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network, Inf. Sci. 420 (2017) 66–76.

[20] X. Li, Y. Jiang, M. Li, S. Yin, Lightweight attention convolutional neural network for retinal vessel image segmentation, IEEE Trans. Industr. Inf. 17 (3) (2020) 1958–1967.

[21] X. Li, Y. Jiang, J. Zhang, M. Li, H. Luo, S. Yin, Lesion-attention pyramid network for diabetic retinopathy grading, Artif. Intell. Med. 126 (2022).

[22] S. Xie, Z. Tu, Holistically-nested edge detection, in, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.

[23] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification, arXiv preprint arXiv:1801.09927 (2018).

[24] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2018) 2439–2450.

[25] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

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

[27] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[29] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147 (2016).

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

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[32] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: in: 2016 Fourth International Conference on 3D Vision (3DV) IEEE, 2016, pp. 565–571.

[33] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).

[34] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, D. Shen, Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images, Med. Image Anal. 70 (2021).

[35] 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) 25–30.

[36] E. Decencière, 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.

[37] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst., Man, Cybernet. 9 (1) (1979) 62–66.

[38] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[39] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch (2017).

[40] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321.

[41] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, et al., IDRiD: Diabetic retinopathy–segmentation and grading challenge, Med. Image Anal. 59 (2020).

[42] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018).

[43] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[44] Z. Tian, T. He, C. Shen, Y. Yan, Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3126–3135.

[45] 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 38 (10) (2019) 2281–2292.

[46] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.

[47] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[48] G. Lin, A. Milan, C. Shen, I. Reid, RefineNet: Multi-path refinement networks for high-resolution semantic segmentation, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.

[49] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.