Neurocomputing 290 (2018) 161–171

Contents lists available at ScienceDirect

Neurocomputing

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

Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks

Juan Mo[a,b], Lei Zhang[a,*], Yangqin Feng[a]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] School of Science, Inner Mongolia University of Science and Technology, Baotou 014010, China

ARTICLE INFO

Article history:
Received 12 August 2017
Revised 15 December 2017
Accepted 8 February 2018
Available online 12 February 2018

Communicated by Dr Bo Du

Keywords:
Diabetic macular edema recognition
Fully convolutional network
Residual learning
Retinal image
Exudate segmentation

ABSTRACT

Diabetic macular edema (DME), one of the leading causes of visual impairment and blindness, is usually diagnosed by the presence of exudates. However, exudate detection is challenging due to the large intraclass variation and high interclass similarity. To overcome these challenges, we propose the cascaded deep residual networks to recognize DME. Specifically, we first design a fully convolutional residual network that fuses multi-level hierarchical information to segment exudates accurately with a fast speed. Compared with previous methods, our approach avoids a wide range of preprocessing or postprocessing steps, reducing the impact of subjective factors. Then based on the segmentation results, the region centered on the pixel with the maximum probability is cropped and fed into the other deep residual network (for classification) to distinguish DME from its hard mimics. This makes the classification network to extract more representative features based on the segmentation results instead of the original images, further reducing the influence of complicated background. We evaluate the proposed method on two publicly available databases, the HEI-MED and e-ophtha EX databases. Extensive experiments demonstrate that our approach achieves better performance than the state-of-the-art methods with a fast processing speed, making it suitable for real-world clinical applications.

1. Introduction

Diabetic retinopathy (DR) is one of the leading causes of blindness in the working-age population of the developed world. The World Diabetes Foundation estimates that there will be 438 million people worldwide suffering from diabetes by 2030 [1]. Diabetic macular edema (DME), an important complication of DR, is also the most common cause of visual impairment and blindness [1]. DME refers to the retina thickening or hard exudate accumulation because of the leakage of fluid within the central macula from abnormal blood vessel or aneurysms [2]. In clinical practice, ophthalmologists usually diagnose DME by the presence of exudates, which appear as bright lesions with well defined border and variable shapes, as shown in Fig. 1.

If DME is diagnosed in time, progression to vision impairment can be slowed or averted. In general, the diagnosis of DME is usually based on the presence of exudates, thus it is very important to locate the exudates. However, detecting exudates is a time-consuming manual screening that requires a trained clinician to identify the exudates and evaluate digital color fundus photographs of the retina. Therefore, in view of the increasing prevalence of diabetes on the one hand, and the limited or even decreasing number of specialists on the other hand, automated exudates detection methods are urgently demanded to reduce the burden on specialists. However, exudate detection is a challenging problem. The main difficulties stem from the following aspects [3]: (1) there is a large variety in the size of the exudates: small lesions can be as small as microaneurysm (only a few pixels on a retinal image), and big ones can be as large as the optic disc. (2) The shape and intensity values of exudates can vary hugely, increasing the difficulty of exudate detection. (3) Some anatomical structures, e.g., optical artefacts and vessels reflections, which may share similar information (intensity, texture etc) as the exudates (see Fig. 1 and tend to mislead exudate segmentation methods. (4) The images from patients of different ethnicity and age group make the color and tissue pigmentation varies greatly, which further complicate the lesion segmentation and diagnosis algorithms.

Hence, the automated exudate segmentation has received significant attention over recent decades. All existing algorithms for exudate segmentation can be broadly categorized as thresholding methods [4], region growing methods [5], morphology based approaches [6,7] and classification methods [1,3]. The technique

[*] Corresponding author.
E-mail addresses: mojuan@stu.scu.edu.cn (J. Mo), leizhang@scu.edu.cn (L. Zhang), yangqinfeng@stu.scu.edu.cn (Y. Feng).

https://doi.org/10.1016/j.neucom.2018.02.035

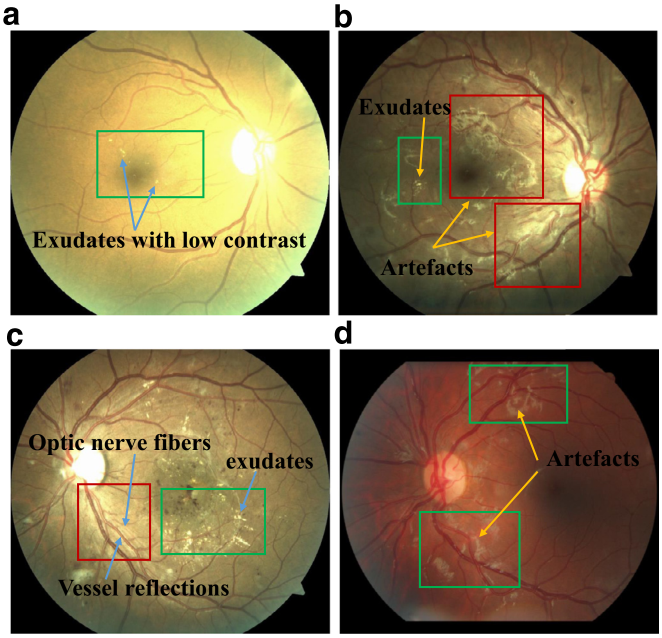


Fig. 1. Examples of fundus images in the database. (a) The low contrast of intensity values between exudate and optic disc, (b),(c) Some anatomical structures with high similarity to exudates. (d) Healthy image with many artefacts.

by Ali et al. [8] is one of the few exceptions where a retinal statistical atlas based on ethnicity is built and the lesions on a diseased eye image are identified depending on the chromaticity differences between the mean atlas image and the diseased image.

Thresholding methods segment exudates based on a global or adaptive grey level analysis. Sánchez et al. [4] proposed a mixture model to fit the histogram, and then the estimated mean and variance of a normal distribution are used to dynamically threshold the images in order to separate exudates from background. The algorithm obtained a sensitivity of 90.2% and a positive predictive value of 96.8%. However, it is prone to misclassify some bright structures (such as vessel reflections and optical artefacts) as exudates.

Region growing algorithms segment retinal images based on the homogeneity of the exudates illumination. For example, Sinthanayothin et al. [5] proposed a recursive region growing segmentation algorithm to automatically detect features of non-proliferative diabetic retinopathy. This technique has the drawback of being computationally expensive when it is used to the whole image.

Mathematical morphology based exudate detection methods use grayscale morphological operators to recognize all structures with predictable shapes (such as vessels and optic disc), then these structures are removed from the image so that exudates can be identified. Walter et al. [7] firstly removed the optic disc by means of morphological filtering techniques and the watershed transformation, then used grey level variation to find exudates and detected their contours by means of morphological reconstruction techniques. Similarly, Sopharak et al. [6] proposed an exudate detection techniques based on mathematical morphology on retinal images of non-dilated pupils. They combined morphological closing reconstruction operator with thresholding to remove the optic disc and main vessels, then identified exudates based on the features such as standard deviation, hue, intensity and number of edge pixels. Due to the large variety of the size of the exudates, it is difficult to select appropriate parameters for morphology operators. The algorithm will fail to detect tiny exudates and distinguish exudates from other bright lesions such as artifacts, drusen etc.

Classification methods extract a feature vector for each pixel or candidate region, which then feed into a classifier such as support vector machine (SVM), random forest classifier,or artificial neural network (ANN) to predict whether the pixel or candidate region is an exudate or not. Giancardo et al. [1] proposed an image level classification method to classify an image into one of the two classes: "healthy" or "presence of DME". Firstly, they used the thresholding method to generate the exudate segmentation map, then the feature vector of each image including color, wavelet decomposition and the exudate segmentation likelihood is taken into a SVM. In [3], Zhang et al. proposed a novel preprocessing method, which not only removed vessels and dark lesions, but also get rid of bright artifacts and vessel reflections. Then candidate regions were generated by mathematical morphology method. Finally, for each candidate, a feature vector that consisting of intensity, geometric, textural, contextual and hybrid features was taken into a random forest classifier to identify the exudates among the candidates. [3] achieved excellent performance for exudate detection. However, these low-level hand-crafted features were application-specific and can not generalize well, they were incapable of distinguishing exudates from other bright structures including optic disc and vessel reflections etc. Therefore, the exudate segmentation methods discussed above require many complicated preprocessing steps to remove some anatomical structures which have similar intensity or texture as the exudates.

Recently, deep neural networks (DNNs) have outperformed the state-of-the-art methods in many image recognition tasks [9–11]. This success can be partially attributed to the ability of DNNs to automatically extract hierarchical features of data [12]. In terms of exudate segmentation, Prentašić and Lončarić [13] constructed a pixel-wise classifier based on convolutional neural network (CNN). The class label of each pixel (exudate or non-exudate) is predicted by providing a square window centered on that pixel as an input. They also combined the output of CNN with the output of the optic disc detection and vessel detection procedures to improve the segmentation results. Perdomo et al. [14] proposed a two-stage CNN model for DME grade assessment. Firstly, RGB patches of 48×48 pixels were taken as inputs of a 8-layers CNN for exudate detection. Then the grayscale mask generated by the previous detection model and the original fundus image feed into the off-the-shelf AlexNet model to obtain the classification of DME. However, previous patch-based methods impose a heavy burden on the computational efficiency, because each patch must be computed separately by the network. This is infeasible for large-scale retinal images. Moreover, these methods employ CNN with relatively shallow architecture, which has limited discriminative capability and can not well address the challenges in exudate recognition.

To overcome these shortcomings of previous methods, we propose a cascaded approach to segment exudate first and then based on the segmentation results to recognize the DME. Specifically, our method consists of two stages. The first stage is exudate segmentation stage, in which the segmented probability map is obtained by developing the deep fully convolutional residual network (FCRN) that fuses multi-level features with different receptive fields. The FCRN adopts an end-to-end way to obtain the precise segmentation results. Within a single forward propagation, the network can produce a probability map with the same size as the original input. Therefore, it is much faster than previous patch-based methods. By exploiting the newly developed residual learning technique, the network is capable of extracting rich and discriminative features of exudates and achieving accurate segmentation results. Therefore, compared with the previous exudate segmentation methods, our FCRN avoids a wide range of preprocessing and postprocessing steps. The second stage is the classification stage. For each image, according to the segmented probability map, the region centered on the pixel with the maximal probability value is cropped and

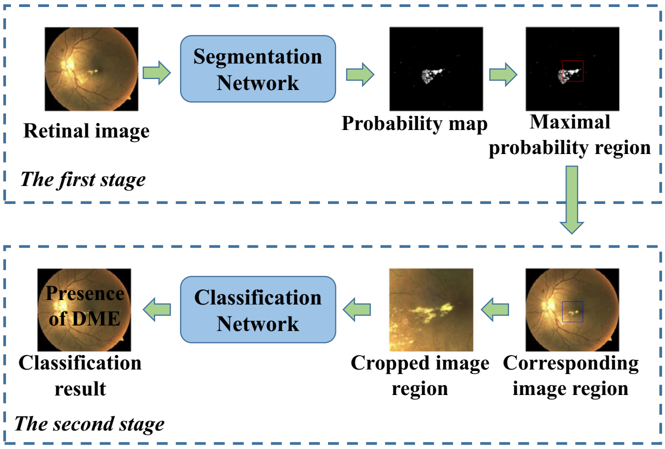


Fig. 2. An overview of the proposed cascaded framework for DME recognition.

fed into a deep residual network to classify an image into one of the two classes: "healthy" or "presence of DME". This stage allows the network to extract more specific and representative features on the most challenging region rather than the whole fundus image.

The remainder of this paper is organized as follows. The proposed method is presented in Section 2. In Section 3, performance metrics are defined and the experiments and results are reported. We further discuss our method in Section 4. Finally, our conclusions are given in Section 5.

2. Method

As mentioned above, exudates have large intraclass variation (size, shape, and intensity etc) and high interclass similarity (similar to optic disc and vessel reflection etc in intensity and texture). If we directly infer the presence the DME on the original fundus images, especially in the case of limited training data, it will seriously influence the recognition performance. In view of this situation, we propose a cascaded deep residual network to overcome these challenges. As shown in Fig. 2, we first obtain the probability map of exudate segmentation by the segmentation network, then based on the segmentation result, the classification network takes the region with maximal probability as the input and distinguish DME from its hard mimics. In this section, we first give a brief introduction to the residual learning, then describe in detail the segmentation and classification networks in the proposed method, respectively.

2.1. Deep residual networks

The importance of network depth has been verified theoretically [15] and empirically [16]. However, achieving significant performance gains from deep networks is not as easy as staking many layers to the networks, simple stacking will lead to the notorious problem of gradient vanishing and degradation issue. In order to easy the training of deep network and take advantage of its rich representation ability with the limited training data in medical images, we employ the newly developed residual learning in our DME recognition task. Here we give a brief introduction to the fundamentals of residual learning and readers can refer to [16,17] for more details.

Deep residual networks (DRNs) consist of a set of residual blocks and show compelling accuracy and nice convergence behaviors on several challenging recognition tasks, including ImageNet [16] and MS COCO competitions [18]. This success can be partially attribute to the skip connections in its residual blocks which allow the information to be directly propagated from one block to other blocks. As shown in Fig. 3, a typical residual block is composed of a few stacked layers, i.e., convolutional layers, rectified Linear Unit (ReLU) layers and batch normalization layers. In general, each residual block can be expressed by the following formula

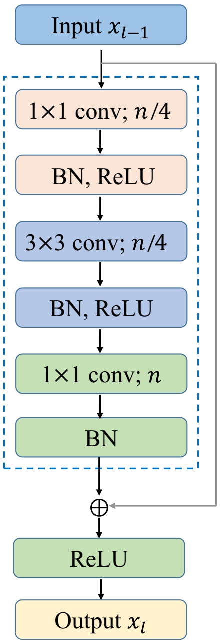


Fig. 3. The illustration of the lth residual block, both 1×1 and 3×3 are the kernel size of the convolutional layer, n represents the number of feature maps. BN and ReLU denote the batch normalization and ReLU activation layers, respectively.

$$\begin{cases} \mathbf{y}_l = \mathbf{x}_l + \mathcal{F}(\mathbf{x}_l, W_l), \\ \mathbf{x}_{l+1} = f(\mathbf{y}_l), \end{cases} \tag{1}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are input and output of the lth residual block, f is a ReLU function and $\mathcal{F}$ is a residual function that the lth block learns. The key idea of deep residual learning is to reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. This operation is implemented by using identity mapping as the skip connections and the element-wise addition operations, which make information propagate through the network smoothly [16]. It is noting that the dimensions of x and $\mathcal{F}$ should be equal in Eq. (1). If needing to change the size of the intermediate representation (such as downsampling operations), we can replace x with $\hat{x}$ obtained by suitably sub-sampling x or using a plain layer (without residual block) to change dimensionality.

2.2. Fully convolutional residual network for exudate segmentation

2.2.1. Network architecture

Exudate segmentation can be formulated as a per-pixel classification problem and the residual network was originally used for classification tasks. If we directly employ the residual network in our task, it will be implemented in sliding window way and impose heavy burden on the computational efficiency. In order to achieve efficient and precise segmentation results, inspired by the recent studies of fully convolutional network (FCN) [19], which converts the fully connected layers in the traditional CNN into convolutional kernels and thus can be applied to input image of arbitrary size, we design a fully convolutional residual network (FCRN) that input an image and directly output an equal-sized prediction score map within a single forward propagation. As shown in Fig. 4, the architecture of the proposed network consists of two

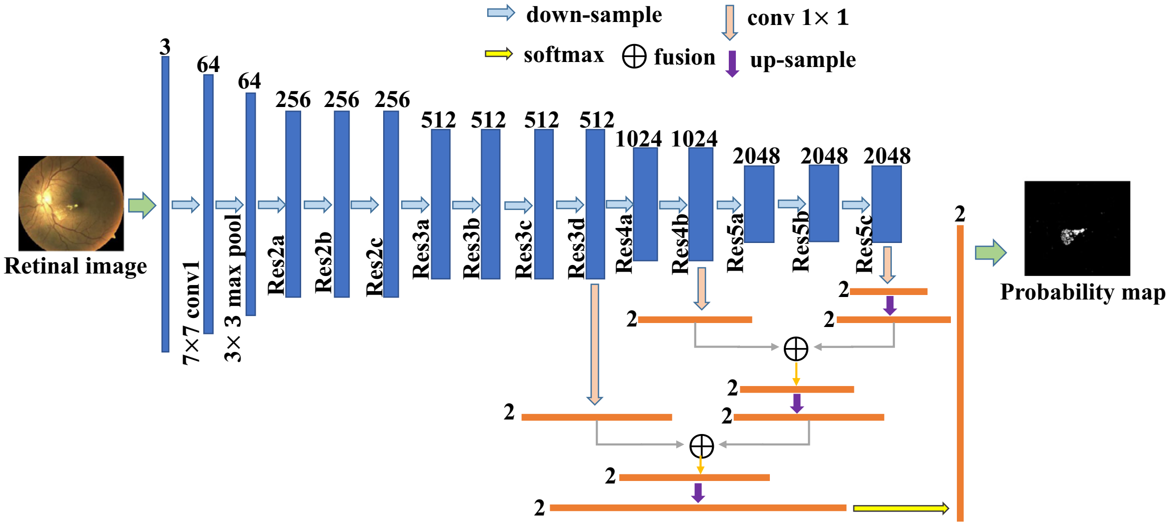


Fig. 4. The architecture of the proposed segmentation network in the first stage. Each box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The arrows denote the different operations, where the thin gray and orange arrows represent fusing the feature maps of different levels and producing the corresponding results, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parts, i.e., downsampling and upsampling modules. The downsampling module is composed of convolutional and max-pooling layers which capture discriminative semantic information and are robust to appearance changes. In the downsampling module, there are one 7×7 convolutional layer, one max pooling layer (both with stride of 2) and 12 residual blocks. Seen from Fig. 3, each residual block is composed of two 1×1 convolutional layers, and one 3×3 convolutional layer (with padding of 1 to preserve the spatial resolution after convolution). Each convolutional layer is followed by a batch normalization layer (BN) and rectifier linear Unit (ReLU). Previous studies [20] have demonstrated that BN can reduce the internal covariance shift, and thus accelerate the training process and improve performance.

After successive downsampling operations, the sizes of the feature maps are gradually reduced and become much smaller than that of the ground truth, which is unsuitable for the pixel-wise exudate segmentation. Therefore, the upsampling operations are implemented to ensure that the sizes of outputs are the same as that of the ground truth. The upsampling module comprises convolutional and deconvolutional layers, which obtain fine spatial resolution for precise localization. It is noting that the deconvolution is implemented by the backwards strided convolution [19] and its kernel can be learned during training.

Even though the upsampling operations followed by successive layers of downsampling can fast up-scale feature maps into the original input resolution, the large stride (32 pixels in our experiment) at the final prediction layer limits the scale of detail in the upsampled output, thus it may ignore some detailed location information and produce unsatisfactory results in some regions. Exudate segmentation is a pixel-level classification problem, so the hierarchical features and refined localization are required simultaneously. Global information from higher layers resolves the problem of what (classification ability) while local information from lower layers covers the problem of where (precise localization). Thus, we introduce skip connections between the downsampling and upsampling layers. Skip connections incorporate multi-lev el hierarchical feature and help the upsampling layer recover fine-grained information from the downsampling layers. Specifically, we first add 1×1 convolutional layers with channel dimension 2 to predict scores for class of exudate or non-exudate at three coarse output locations (i.e., residual blocks 3d, 4b,and 5c), which produces three levels of prediction maps, i.e., 8-pixel, 16-pixel and 32-pixel stride prediction maps, respectively. Then we fuse the information of three different hierarchical features with deconvolution and summing operations. Specifically, we first add a $2\times$ upsampling deconvolution layer on the top of 32-pixel stride prediction map to fuse it and the 16-pixel stride prediction map by using the element-wise summing operation, which generate a new 16-pixel stride prediction map. Please refer to the gray arrows and fusion symbols in Fig. 4. Similarly, we continue to fuse the new 16-pixel stride prediction map and the 8-pixel stride prediction map by adding a $2\times$ upsampling deconvolution layer on the top of the new 16-pixel stride prediction map, which produces the fused 8-pixel stride prediction map. Finally, we add a $8\times$ upsampling deconvolutional layer on the top of the fused 8-pixel stride prediction map to generate the final score map with the equal size as the input image. As a result, the final score map combine both global abstract features and local refine spatial information, generating accurate prediction results.

2.2.2. Details in training the segmentation network

Because exudates just account for a small proportion in retinal images, in the training process, to alleviate the class imbalance, for pathological images we cropped multiple regions containing as many exudates as possible with the size of 480×480 pixels, while for healthy images, we select one 480×480 challenging region with the similar information to exudates. As there are very few training data available for our tasks, to reduce overfitting in deep networks, the above cropped regions are augmented by different transformations, including rotation, translation, flipping, and mirroring. Although we include as many exudates as possible in constructing the training set, the distribution of exudate/non-exudate pixels is still heavily imbalanced: 96% of the ground truth is non-exudate. Therefore, the frequency-based balance strategy [21] is introduced to weight the loss differently according to the true class. we weight each class c by $\beta_c = mean\_f/f_c$, where the class frequency f_c is the number of pixels of class c divided by the total number of pixels in images where c is present, $c = 0, 1,$ and $mean\_f$ is the mean value of f_0 and f_1. Specifically, let I_c denote the set of images that contain the pixels of class c in the training set. $|X^c|$, $|X|$ represent the number of pixels of class c and the total number of pixels in image X, respectively. Then

$$f_c = \frac{\sum_{X \in I_c} |X^c|}{\sum_{X \in I_c} |X|}, c = 0, 1$$

Therefore, the exudate segmentation can be formulated as minimizing the following cross entropy loss function:

$$\mathcal{L}(W) = -\beta_1 \sum_{j \in Y^+} \log p\Big(y_j = 1 \Big| X; W\Big) \\ -\beta_0 \sum_{j \in Y^-} \log p\Big(y_j = 0 \Big| X; W\Big) + \frac{\lambda}{2} \|W\|_2^2, \quad (2)$$

where Y^+ and Y^- denote the class of exudate and non-exudate ground truth label sets in an image, respectively. $p(y_j = c \mid X; W), c = 0, 1$ is the output probability for class c at the pixel j. The last term is a regularization term that helps to prevent overfitting and the hyperparameter λ controls the relative importance of the data loss and weight decay terms. In the testing phase, given a image, the segmentation results are produced with the overlap-tile strategy to improve the robustness.

2.3. Residual network for DME recognition

2.3.1. Integration of the two stages

DME recognition is a classification problem, thus a intuitive way is to directly perform DME classification for the original retinal image. However, compared to the whole image, the size of exudate regions is extremely small and its distribution in the image has no fixed pattern. In this circumstance, if we directly recognize the DME from the whole retinal images, it will be disturbed by the complicated background of the original image and seriously influence the performance of the network, especially in the case of insufficient training data. In addition, the diagnosis of DME is usually based on the presence of exudates, i.e., an image is considered as presenting DME if it contains at least one exudate. Therefore, another straightforward way is whether we can recognize the DME according to the maximal value in the segmentation map. However, due to the huge variety of exudates in size, shape, intensity and contrast, some challenging mimics in healthy images also have high probabilities. Thus, directly using the maximal segmentation probability as the prediction score for recognizing DME will increase false positive rate. In this regard, we propose to first segment the exudates from retinal images and then crop the region centered on the position with maximal segmentation probability. Finally, we resize the cropped region into a fixed size and take it into the classification network to distinguish the DME from its hard mimics. It is worthwhile to point out that although having two stages, the whole recognition process is performed in a automated way.

The motivation of employing the above cropped region is to build a representative training database for the classification network. These cropped regions represent the location most likely to contain exudates or its challenging mimics. Therefore, in the classification process, for DME image, it makes the feature extraction focus on the exudate lesions, avoiding the influence of the complicated backgrounds. While for healthy image, it guarantees selecting the most challenging region and then the non-DME training samples can be well represented. Thus the capability of the classification model in distinguishing the DME from the non-DME can be greatly enhanced by these hard mimics. Therefore, the integration of the two-stage framework enables the classification network to extract more representative and specific features based on the segmentation results instead of directly infer the presence of DME on the original images, further alleviating the issue of insufficient training data.

2.3.2. Network architecture

As shown in Fig. 5, the architecture of the classification network is similar to the downsampling part of the segmentation network. It consists of one 7 × 7 convolutional layer, one max pooling layer (both with stride of 2) and 5 residual blocks. In addition, we append a 7 × 7 average pooling layer at the residual block 5a to extract the global abstract features. Finally, the softmax classifier is added to recognize the presence of DME (i.e., discriminate whether a image contains exudates). *2.3.3 Details in training the classification network*

To train the classification network, we first select the region no more than 480 × 480 pixels and centered on the position with the maximal probability, and then resize the cropped region into a fixed size (224 × 224 in our experiments). To improve the robustness and reduce the overfitting, we use various transformations to augment the training set, including rotation, flipping, and mirroring. During testing, we also crop the region with the maximal segmentation probability and take the resized region into the classification network.

3. Experiments and results

3.1. Materials

We evaluated our method on two publicly available retinal image databases: HEI-MED and e-ophtha EX.

The HEI-MED (Hamilton Eye Institute Macular Edema Dataset) database [1] contains 169 fundus images with the resolution of 2196 × 1958. Of the 169 images in the database, 115 are healthy and 54 contain exudates. For each image, manual segmentation of exudate is provided. The e-ophtha EX database [3] provides pixel level annotation for exudate segmentation, which contains 82 retinal images with four different resolutions, ranging from 1440 × 960 pixels to 2544 × 1696 pixels. Among these images, there are 47 images containing exudates and 35 normal images. Because the training set and testing set are not explicitly specified in the original databases, we adopt the 5-fold cross validation to evaluate the performance of our method on the HEI-MED and e-ophtha EX databases.

3.2. Evaluation metrics

In the exudate segmentation process, any pixel is classified as either exudate or non-exudate. For evaluation purposes, we employ the following evaluation metrics: sensitivity (*Se*), positive predictive value (*PPV*) and *F-score*. They are respectively defined as:

$$Se = \frac{TP}{TP + FN}, PPV = \frac{TP}{TP + FP}, F\text{-}score = \frac{2 \times Se \times PPV}{Se + PPV}.$$

It is noting that *F-score* is the most important performance metric in our case, since it is high if and only if both *Se* and *PPV* are large. To calculate these parameters, we use the evaluation criterion proposed by Zhang et al. [3]. In this method, instead of counting the pixels detected correctly, they used the connected component level validation. Based on this criterion, pixels of each connected component are considered as true positive (*TP*) if they partially or totally overlap with the ground truth. The true negatives (*TN*), false positives (*FP*) and false negatives (*FN*) are calculated in the similar way. Readers can refer to Zhang et al. [3] for more details.

As for the classification, an image is classified as one of the two classes: “healthy” or “presence of DME”. The performance metrics include sensitivity (*Se*), Specificity (*Sp*), and accuracy (*Acc*). Because the quantitative measures are dependent on the threshold of the probability, a receiver operating characteristic (ROC) curve is a plot of *TP* fractions versus *FP* fractions given by varying the threshold of the predictive score on all images. The area under the ROC curve

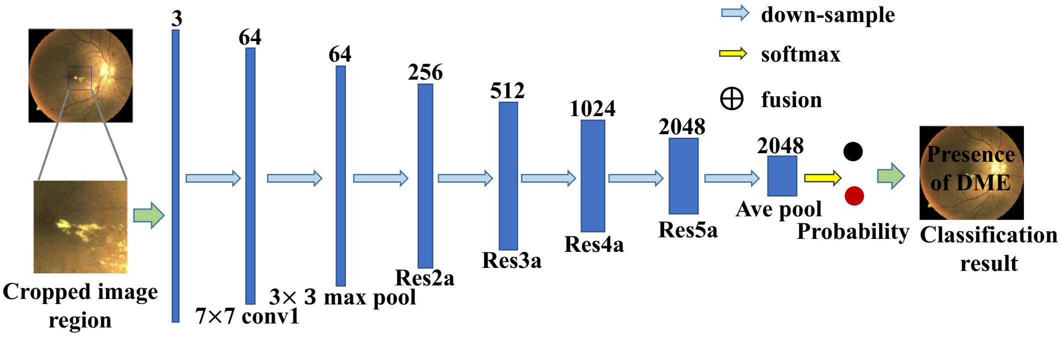


Fig. 5. The architecture of the proposed classification network in the second stage. Each box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The arrows denote the different operations.

(*AUC*) is also employed as a performance metric. *AUC* is the most important indicator that abstracts from the proportions of the decision classes and characterizes the entire ROC curve with a single number. A perfect classifier will have an *AUC* of 1. *Se*, *Sp*, and *Acc* are respectively defined as:

$$Se = \frac{TP}{TP + FN}, \; Sp = \frac{TN}{TN + FP}, \; Acc = \frac{TP + TN}{TP + FN + TN + FP}.$$

It is noting that for classification, *TP, TN, FP*, and *FN* are measured at image level instead of pixel level.

3.3. System implementation

The proposed method is implemented using the public available Caffe framework [22] in a NVIDIA GTX Titan GPU. The two-stage networks are trained using the stochastic gradient descent method (momentum = 0.9, weight decay = 0.0005). For the segmentation network, the learning rate is initialized as 0.001 and divided by 10 every 10,000 iterations. While for classification network, it is set as 0.01 initially and decreased by a factor of 10 every 1000 iterations. Thanks to the cascaded residual network, our method is very efficient, it takes about 1.04 seconds to process an image of size 2196 × 1958 (1 s for segmentation and 0.04 s for classification), which makes our algorithm suitable for real-world clinical applications.

Training a deep neural network requires a large amount of labeled training data, which remains a challenge for medical images because of the expense of expert annotation and scarcity of disease (e.g., lesions). Previous studies have indicated that filters trained on large scale well-annotated ImageNet could be transferred to different recognition tasks in other domains [23,24]. Therefore, the parameters of the classification network and the downsampling part in the segmentation network are both initialized by the pre-trained filters in ResNet [16]. This process can be considered as the pre-training phase in the neural network with good initialization, which accelerates the convergence of the network. Besides, the deconvolutional layers in the segmentation model were initialized by bilinear interpolation.

3.4. The performance of our method in segmentation

3.4.1. Qualitative evaluation

The segmentation result is a exudate probability map, in which each value represents the probability of each pixel belonging to the exudate class. To obtain the binary exudate segmentation, a thresholding scheme is applied to the probability map to determine whether a particular pixel belongs to the exudate class or not. Fig. 6 shows some qualitative results on HEI-MED and e-ophtha EX databases, respectively, which illustrate that the proposed method can segment the exudate accurately in various conditions without any complicated problem-specific preprocessing or postprocessing. For the problem of exudate segmentation, there are some challenging regions, including the exudate region with low contrast to optic disc, regions with optic nerve fibers and vessel reflections, and some artefacts regions. It is difficult to distinguish these regions from exudates. Fig. 7 shows some segmentation results of the propose method on these challenging regions. As shown in Fig. 7(a), although the intensity of exudate is close to that of the optic disk, the exudate can be classified with high accuracy by our method, while the whole optic disc can be recognized as non-exudate and its probability value being exudate is almost zero. It indicates the ability of the proposed method for exudate segmentation in the presence of low-contrast exudates. In Fig. 7(b), both optic nerve fibers and vessel reflections are present. Compared with true exudates, their probability values being exudate produced by the proposed method are significantly low, which suggests that our method can recognize exudate effectively in the presence of vessel reflections and optic artefacts. In addition, Fig. 7(c) demonstrates that our method achieves satisfactory results in the presence of the artefacts. All these challenging examples illustrate the effectiveness of the proposed method.

Table 1
Performance comparison of exudate segmentation methods on HEI-MED and e-ophtha EX databases.

Database	Method	Se	PPV	F-score
HEI-MED	Imani [25]	0.8126	0.6357	0.7133
	Pereira [26]	0.8082	0.7301	0.7632
	Prentašić [13]	0.8582	0.7335	0.7910
	Proposed method	**0.9255**	**0.8212**	**0.8499**
e-ophtha EX	Imani [25]	0.8032	0.7728	0.7877
	Zhang [3]	0.74	0.72	0.73
	Liu [27]	0.76	0.75	0.76
	Das [28]	0.8580	0.5793	0.6916
	Proposed method	**0.9227**	**0.9100**	**0.9053**

3.4.2. Comparison with other methods

We compared the segmentation performance of our method with that of state-of-the-art algorithms on HEI-MED, and e-ophtha EX databases. As illustrated in Table 1, our deep segmentation network outperforms the state-of-the-art methods by a large margin on both of the databases. In general, all performance measures, including *Se*, *PPV*, and *F-score* produced by the proposed algorithm are greater than those of other methods for both of the databases, which suggests the algorithm has a strong ability to identify exudate with a low FP rate compared to other methods. Moreover,

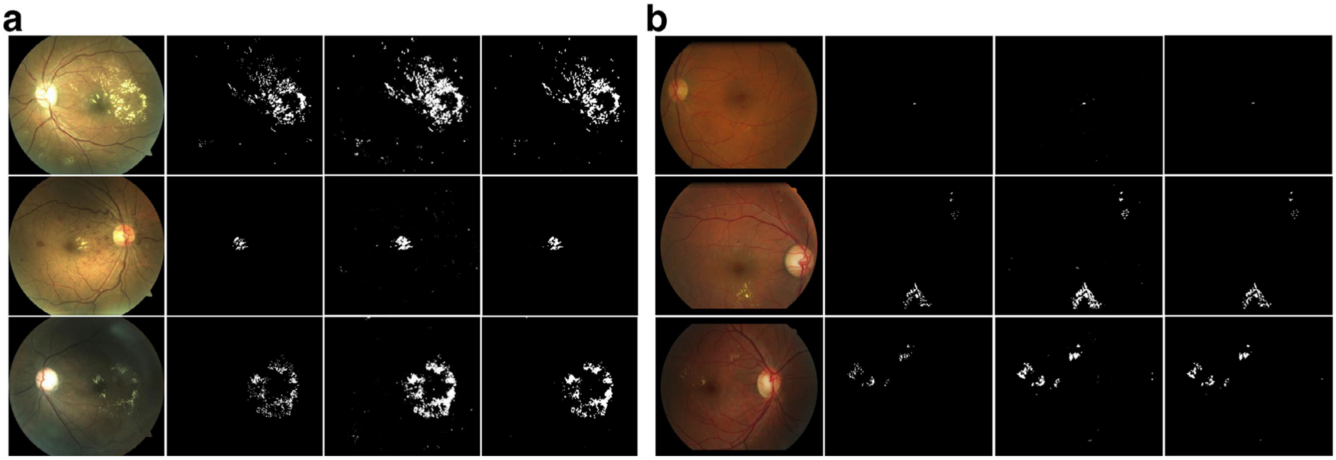
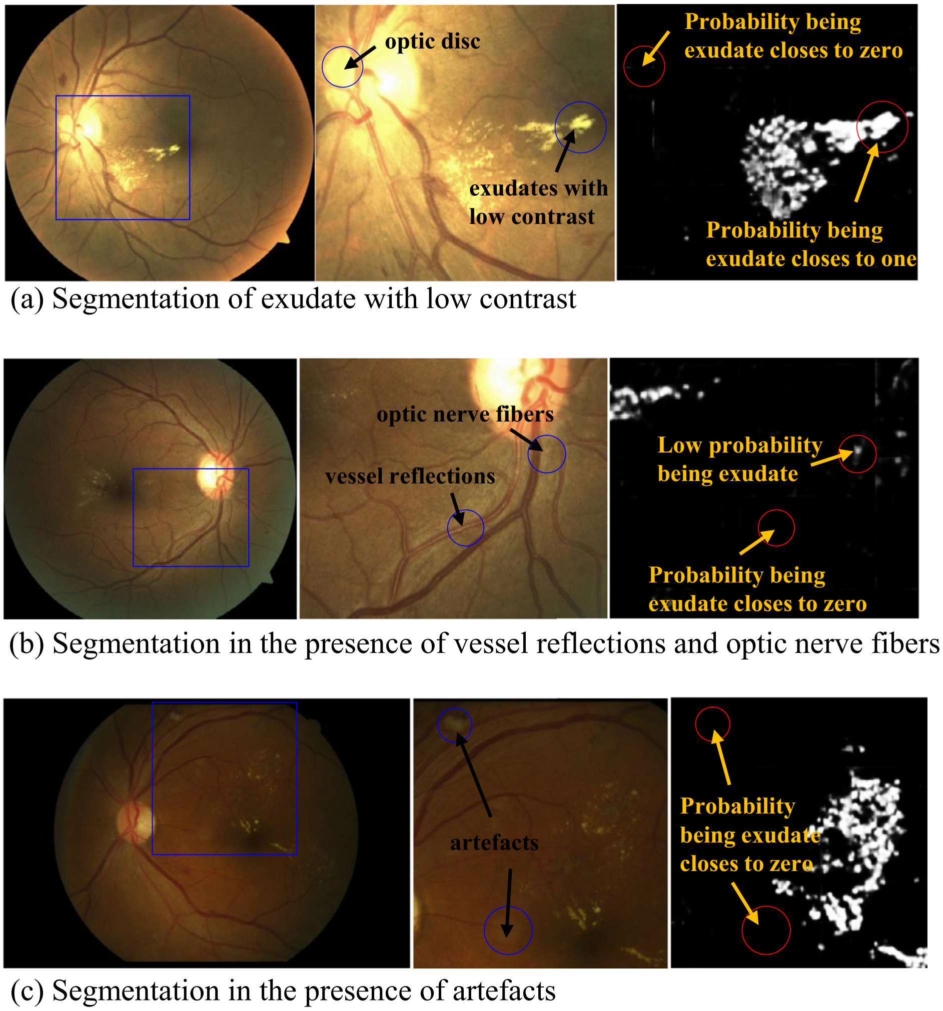


Fig. 6. (a) and (b) show some segmentation results on both of the HEI-MED and e-ophtha EX databases, respectively. The first through fourth columns are the retinal image, manual segmentation, probability map, and binary segmentation results (white and black pixels denote the exudate and non-exudate regions, respectively).


optic disc
exudates with low contrast
Probability being exudate closes to zero
Probability being exudate closes to one
optic nerve fibers
vessel reflections
Low probability being exudate
Probability being exudate closes to zero
artefacts
Probability being exudate closes to zero

(a) Segmentation of exudate with low contrast

(b) Segmentation in the presence of vessel reflections and optic nerve fibers

(c) Segmentation in the presence of artefacts

Fig. 7. Segmentation results in some challenging regions. (a) Segmentation of exudate with low contrast. (b) Segmentation in the presence of vessel reflection and optic nerve fibers. (c) Segmentation in the presence of artefacts.

Table 2
Computation time for processing one image.

Method	Computation time
Prentašić [13]	25.4 min
Proposed method	1 s

Table 3
Performance of recognition with and without segmentation network on the HEI-MED database.

Method	Se	Sp	Acc	AUC
Without segmentation	0.7963	0.9304	0.8876	0.9229
With segmentation	0.9630	0.9304	0.9408	0.9709

Table 4
Performance of recognition with and without classification network on the HEI-MED database.

Method	Se	Sp	Acc	AUC
Without classification	0.8333	0.9130	0.8876	0.9396
With classification	0.9630	0.9304	0.9408	0.9709

Table 5
Performance comparison of DME recognition methods in terms of AUC.

Database	Method	AUC
HEI-MED	Giancardo [1]	0.94
	Pereira [26]	0.67
	Zhang [3]	0.94
	Akram [29]	0.94
	Proposed method	**0.9709**
e-ophtha EX	Giancardo [1]	0.87
	Zhang [3]	0.95
	Proposed method	**0.9647**

in contrast to other methods which require complicated preprocessing steps to remove some similar anatomical structures, our method adopts an end-to-end fashion to train the segmentation network and needs no problem-specific preprocessing or postprocessing, which reduces the impact of subjective factors.

3.4.3. *Computation time*

Training the network takes about 8 hours on a single NVIDIA GTX Titan GPU. After the training process is completed, it is much faster to test an image. Generally, it takes about 1 second to segment an image of size 2196 × 1958 pixels using the trained network. We compared the computation time of our algorithm with the latest deep learning-based methods published by Prentašić and Lončarić [13]. As shown in Table 2, the computation time of our approach is much less than that in [13], which uses the sliding window way to obtain the segmentation results, leading to redundant computations on neighboring pixels. Moreover, our approach achieves better performance than existing methods. This makes our algorithm suitable for real-world clinical applications.

3.5. *The performance of our method in classification*

3.5.1. *Recognition DME with and without two-stage framework*

We adopt a two-stage framework to automatically identify DME, in order to investigate the necessity of the two-stage scheme, we carry out a series of comparison experiments on the HEI-MED database.

(*1) Recognize DME with and without segmentation network*

In order to validate the importance of segmentation network, we compare the classification performance with and without segmentation network. For the sake of fairness, both of the experiments adopt the same data augmentation strategy, use the same network architecture, and initialize the network by the pre-trained parameters in ResNet. As shown in the Table 3, the two-stage network achieves higher *AUC* value than directly takeing the original image into the deep residual classification network without segmentation stage. This is because the identify of exudate is very challenging (see Fig. 1), especially in the case of insufficient training data. Training the classification network based on the segmentation results can extract more representative features to distinguish DME from its hard mimics. Fig. 8 shows some automatically cropped regions from the segmentation results. Compared with directly performing classification on the original images, these regions focus on the exudate (for DME images) and its mimics (for non-DME images), which help to building a more discriminative training set for the classification network and avoid being disturbed by the complicated background in the original image.

(*2) Recognize DME with and without classification network*

We further carry out the experiments with and without classification network, the results are listed in Table 4, which suggest that the performance of the two-stage network is better than that of directly using the maximal probability value in segmentation map as the criterion of recognizing DME. This is because some challenging mimics in healthy images also have high probability values. Thus, directly using the maximal segmentation probability as the prediction score for recognizing DME will increase false positive rate and influence the final performance.

3.5.2. *Comparison with other methods*

We compared the recognition performance of our method with that of state-of-the-art algorithms on HEI-MED, and e-ophtha EX databases. Different from *Se, Sp*, and *Acc, AUC* abstracts from the proportions of decision classes and is independent of the threshold. Therefore, in Table 5, we compare the *AUC* value of our approach and that of other methods. As illustrated in Table 5, our two-stage network achieves better results than all of the state-of-the-art methods on both of the databases, which demonstrates the advantage of the proposed method in addressing the challenges of the DME recognition. Moreover, our method outperforms the previous classification method [1,3] by a large margin, which illustrates that employing deep residual network with effective learning mechanism can enhance the discriminative capability of CNNs for challenging medical image analysis tasks, even in the circumstance of limited training data.

3.6. *Ablation studies of our method*

To investigate the effect of depth in the network, the impact of multi-level feature fusion, and the role of transfer learning (pre-train), we perform extensive ablation studies. In view of the time spent, the following ablation analysis is carried out on the first cross validation of the HEI-MED database.

3.6.1. *Experiments on segmentation network depth*

To explore whether the increase in network depth can bring performance gains, we compare the performance of the fully convolutional VGG-16 network [19,30] and the proposed FCRN at different depths (29, 38, and 50 layers, respectively). To ensure a fair comparison, all of these networks adopt the same multi-level feature fusion scheme and knowledge transfer strategy (initialize the network by the pre-trained model trained on ImageNet database). The only difference is the depth of the network in the downsample module is different (see Table 6). As shown in Table 7, we draw the following conclusions. Firstly, the performances produced by FCRN with different depth all outperform that of FCN-VGG16, which suggest that increasing network depth can enhance the discriminative

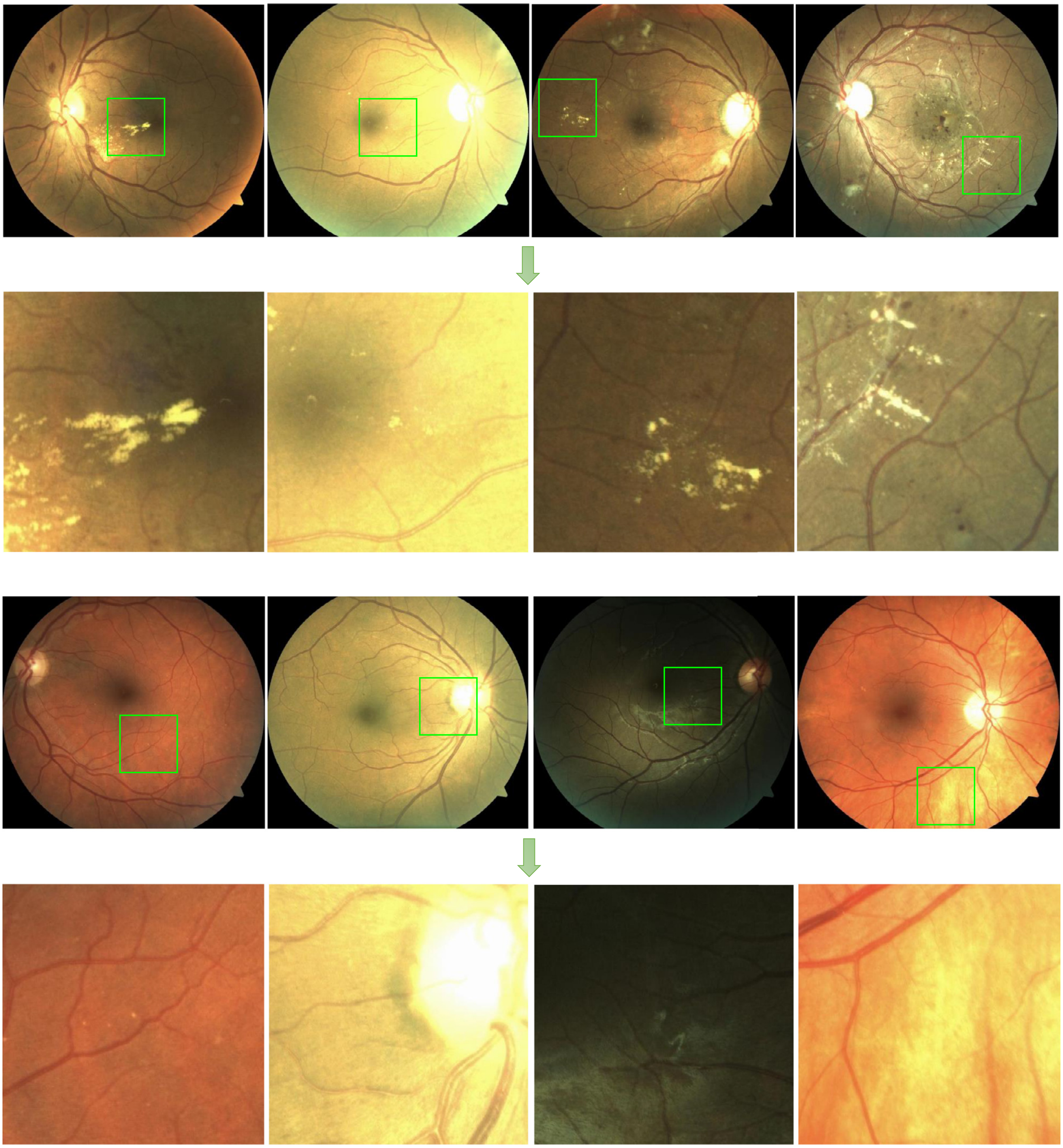


Fig. 8. Some automatically cropped regions based on the segmentation results. The first two rows are the DME images and the corresponding cropped regions, respectively. The last two rows are the non-DME images and the corresponding cropped regions, respectively.

ability of the network and thus improve the performance. Moreover, compared with simple stacking layers to increase depth, the mechanism of residual learning can help to propagate the gradient flow faster and result in better prediction accuracy. In addition, as the depth increase to 50 layers, the network performance is worse than that of 38 layer-FCRN, which partially attributes to limited training data in our task. Compared with 38 layer-FCRN, 50 layer-FCRN is deeper and has more parameters, thus it is difficult to train such deep network with insufficient data that we can obtain in our task. This illustrates there is a trade-off between the network depth and network performance with the insufficient dataset in medical image analysis. Here the 38-layer FCRN is selected as the segmentation network because it results in the good balance between the network depth and performance.

3.6.2. Experiments on the multi-level feature fusion strategy

To further investigate the role of multi-level feature fusion, we design three types of architecture in the upsample module of the proposed FCRN-38. The first one only upsamples the 32-pixel stride prediction map to get the final segmentation map, we refer to it as FCRN-32s. The second one fuses the information of the 32-pixel stride prediction map and the 16-pixel stride prediction

Table 6
Architecture of downsampling module in FCRN-29, −38, and −50.

29-layer	38-layer	50-layer
7 × 7, 64,stride 2		
3 × 3, max pool,stride 2		
ResBlock2a-2c		
ResBlock3a-3b	ResBlock3a-3d	ResBlock3a-3d
ResBlock4a-4b	ResBlock4a-4b	ResBlock4a-4f
ResBlock5a-5b	ResBlock5a-5c	ResBlock5a-5c

Table 7
Performance comparison of architecture with different depth.

Network	Se	PPV	F-score
FCN-VGG-16	0.7444	0.6766	0.6744
FCRN-29	0.8776	0.8524	0.8542
FCRN-38	**0.9088**	**0.8756**	**0.8884**
FCRN-50	0.8258	0.8747	0.8408

Table 8
Performance comparison of FCRN with different multi-level fusion scheme.

Network	Se	PPV	F-score
FCRN-32s	0.8413	0.7905	0.8011
FCRN-16s	0.8784	0.8385	0.8504
FCRN-8s	**0.9088**	**0.8756**	**0.8884**

Table 9
Performance comparison of FCRN with and without transfer learning.

	Se	PPV	F-score
FCRN-with pretrain	**0.9088**	**0.8756**	**0.8884**
FCRN-without pretrain	0.7861	0.8406	0.7924

map by upsampling operation, we refer to it as FCRN-16s. The third one integrates the information of the 32-, 16-, and 8-pixel stride prediction map, the details of fusion scheme is described in Section 2.2 and Fig. 4, we refer to it as FCRN-8s. As demonstrated in Table 8, the FCRN-8s produces the best performance among three types of fusion fashion in all evaluation metrics. It illustrates the effectiveness of the proposed multi-level feature fusion strategy and the significance of the combination of global abstract features from higher layers and local refine spatial information from shallower layers.

3.6.3. *Experiments on the importance of transfer learning*

To verify the effect of transfer learning, we compare the performance of the network with and without knowledge transferring from large scale ImageNet database. It is observed from Table 9 that the network initialized by the pre-trained parameters on ImageNet outperforms the network with random initialization by a large margin in all evaluation measurements. This suggests that good initialization is very important in deep networks, which can accelerate the optimization convergence rate and boost the network performance.

4. Discussion

Table 10
Performance comparison of the classification network with different cropped size.

Size	Se	Sp	Acc	AUC
240 × 240	0.8519	0.9217	0.8994	0.9238
480 × 480	0.8704	0.9565	0.9290	0.9712
720 × 720	0.8889	0.9565	0.9349	0.9620
960 × 960	0.8333	0.9565	0.9172	0.9523
1200 × 1200	0.8889	0.9304	0.9172	0.9523
2196 × 1958 (original image)	0.7778	0.9304	0.8817	0.9229

In this paper, we propose a two-stage framework to recognize the DME in retinal images. The first stage is to segment exudates from the retinal images, and then we crop a region centered on the position with the maximum probability and constrain its size to no more than 480 × 480 pixels. This size reflects the contextual information provided to the classification network. In order to investigate the influence of cropped size to the classification performance, for each image, based on the segmentation results, varied-scale regions containing the maximum probability pixel are cropped, and their size range from 240 × 240 to 1200 × 1200 pixels, and even the entire image. For the sake of fairness, the comparison experiments of different cropped size are conducted under the same parameter configuration and the same threshold (0.5) is set to obtain the corresponding evaluation results. As shown in Table 10, it suggests sizes that are too small (insufficient contextual information) or too large (disturbed by a complicated background) can degrade classification performance, while others within the modest range have little effect on classification performance.

5. Conclusion

In this paper, we propose an accurate and efficient method based on cascaded residual network to recognize DME, which is composed of two parts: segmentation and classification. Firstly, we adopt an end-to-end fashion to train the segmentation network which needs no problem-specific preprocessing or postprocessing. Moreover, the multi-level information fusion strategy is developed to further enhance the segmentation performance. Then based on the segmentation results, the region centered on the pixel with the maximal probability is cropped and fed into the classification network to infer the presence the DME. This allows the network to focus on the distinctions between exudate and its challenging mimics, avoiding the influence of complicated backgrounds and alleviating the issue of insufficient training data. Extensive experiments on two benchmark databases demonstrate the effectiveness of the our method. Compared with start-of-the-art methods, the proposed approach yields better performance with a fast processing speed. Furthermore, the proposed cascaded residual network framework can be easily applied to other similar biomedical segmentation or recognition tasks. In future work, on the one hand, we plan to adopt ensemble method to fuse the information of multiple networks and employ voting strategy to boost the classification performance. In addition, attention mechanism can be utilized to obtain more accurate segmentation results. On the other hand, we try to expand the proposed algorithm to recognize microaneurysms and haemorrhages, which are important for the diagnosis of DR.

Conflict of interest

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

This work was supported by Fok Ying Tung Education Foundation (grant 151068); National Natural Science Foundation of China (grant 61332002); and Foundation for Youth Science and Technology Innovation Research Team of Sichuan Pro vince (grant 2016TD0018).

References

[1] L. Giancardo, F. Meriaudeau, T.P. Karnowski, Y. Li, S. Garg, K.W. Tobin, E. Chaum, Exudate-based diabetic macular edema detection in fundus images using publicly available datasets, Med. Image Anal. 16 (1) (2012) 216–226.

[2] R. Bernardes, J. Cunha-Vaz, Optical Coherence Tomography: A Clinical and Technical Update, Springer Science & Business Media, 2012.

[3] X. Zhang, G. Thibault, E. Decencière, B. Marcotegui, B. Laÿ, R. Danno, G. Cazuguel, G. Quellec, M. Lamard, P. Massin, et al., Exudate detection in color retinal images for mass screening of diabetic retinopathy, Med. Image Anal. 18 (7) (2014) 1026–1043.

[4] C.I. Sánchez, M. García, A. Mayo, M.I. López, R. Hornero, Retinal image analysis based on mixture models to detect hard exudates, Med. Image Anal. 13 (4) (2009) 650–658.

[5] C. Sinthanayothin, J.F. Boyce, T.H. Williamson, H.L. Cook, E. Mensah, S. Lal, D. Usher, Automated detection of diabetic retinopathy on digital fundus images, Diabet. Med. 19 (2) (2002) 105–112.

[6] A. Sopharak, B. Uyyanonvara, S. Barman, Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy c-means clustering, Sensors 9 (3) (2009) 2148–2161.

[7] T. Walter, J.-C. Klein, P. Massin, A. Erginay, A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina, IEEE Trans. Med. Imaging 21 (10) (2002) 1236–1243.

[8] S. Ali, D. Sidibé, K.M. Adal, L. Giancardo, E. Chaum, T.P. Karnowski, F. Mériaudeau, Statistical atlas based exudate segmentation, Comput. Med. Imaging Gr. 37 (5) (2013) 358–368.

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

[10] L. Yu, H. Chen, Q. Dou, J. Qin, P.-A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imaging 36 (4) (2017) 994–1004.

[11] J. Mo, L. Zhang, Multi-level deep supervised networks for retinal vessel segmentation, Int. J. Comput. Assist. Radiol. Surg. (2017), doi:10.1007/s11548-017-1619-0.

[12] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the Computer Vision – ECCV 2014, Springer, 2014, pp. 818–833.

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

[14] O. Perdomo, S. Otalora, F. Rodrguez, J. Arevalo, F.A. Gonzlez, A novel machine learning model based on exudate localization to detect diabetic macular edema, in: Proceedings of the Ophthalmic Medical Image Analysis Third International Workshop, 2016, pp. 137–144.

[15] G.F. Montufar, R. Pascanu, K. Cho, Y. Bengio, On the number of linear regions of deep neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2924–2932.

[16] 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, 2016a, pp. 770–778.

[17] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016b, pp. 630–645.

[18] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.

[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[20] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift[C], in: International conference on machine learning, 2015, pp. 448–456.

[21] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.

[23] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 3320–3328.

[24] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.

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

[26] C. Pereira, L. Gonçalves, M. Ferreira, Exudate segmentation in fundus images using an ant colony optimization approach, Inf. Sci. 296 (2015) 14–24.

[27] Q. Liu, B. Zou, J. Chen, W. Ke, K. Yue, Z. Chen, G. Zhao, A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images, Comput. Med. Imaging Gr. 55 (2017) 78–86.

[28] V. Das, N.B. Puhan, Tsallis entropy and sparse reconstructive dictionary learning for exudate detection in diabetic retinopathy, J. Med. Imaging 4 (2) (2017) 024002.

[29] M.U. Akram, A. Tariq, S.A. Khan, M.Y. Javed, Automated detection of exudates and macula for grading of diabetic macular edema, Comput. Methods Progr. Biomed. 114 (2) (2014) 141–152.

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

Juan Mo received the B.S. and Masters degrees in mathematics in 2006 and 2009, respectively, from the Inner Mongolia University, Hohhot, China. She is currently pursuing the Ph.D. degree at the Machine Intelligence Laboratory, School of Computer Science and Engineering, Sichuan University, Chengdu, China. Her research interests include deep learning and medical image analysis.

Lei Zhang received the B.S. and Master's degrees in mathematics from the University of Electronic Science and Technology of China, Chengdu, China, in 2002 and 2005, respectively. She received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2008. Currently, she is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her research interests include the theory and applications of Neural Networks.

Yangqin Feng received the bachelor and master degrees from Department of Computer Science, Southwest University of Science and Technology, Mianyang, China, in 2011 and 2014, respectively. Currently, she is a third-year Ph.D. candidate at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her current research interests include machine intelligence and medical image processing.