Hard exudate segmentation in retinal image with attention mechanism

Ze Si | Dongmei Fu | Yang Liu | Zhicheng Huang

1 School of Automation and Electrical Engineering, University of Science and Technology Beijing, Haidian District, Beijing, China
2 Beijing Engineering Research Center of Industrial Spectrum Imaging, University of Science and Technology Beijing, Beijing, China

Correspondence
Dongmei Fu, School of Automation and Electrical Engineering, University of Science and Technology Beijing, No. 30 Xueyuan Road, Haidian District, Beijing, China.
Email: fdm_ustb@ustb.edu.cn

Funding information
Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Grant/Award Number: BG0150

Abstract
Diabetic retinopathy (DR) is the main reason that causes preventable blindness. Hard exudate is one of the earliest signs of diabetic retinopathy. Precise detection of hard exudate is helpful for the early diagnosis of diabetic retinopathy. Fully convolutional network (FCN) shows great performance on hard exudate segmentation task. However, there are limitations for fully convolutional network to build long-range dependencies in different regions of the image. Convolution operator extract features in local area, segmentation results based on local features are likely to be wrong in some cases. Another channel attention method was proposed, and two different attention modules are used in the segmentation model. In this way, long-range dependencies across different image regions are built efficiently in different stages of feature extraction. In addition, a new loss function is designed to deal with the data imbalance problem in hard exudate segmentation task. The proposed method was evaluated by two public datasets, and the comparative experiments show the effectiveness of the proposed method.

1 | INTRODUCTION

Diabetic patients often have long periods of high blood glucose. Symptoms of metabolic imbalance and microcirculatory disorders will cause retinal damage and lead to ocular microvascular complications [1]. That is what we called diabetic retinopathy (DR). If the necessary treatment is not carried out in time, patients may experience decreased vision or even blindness. Symptoms of DR in the early stage are imperceptible, patients often lose the best treatment opportunity. Due to the growing number of diabetic patients and lacking trained specialists, the screening of DR has become a worldwide public health problem. It is necessary to do research on the automated methods to detect DR.

Hard exudate caused by chronic retinal neuroepithelial edema is characterized by clear and irregular yellow-white dots on the retina. Hard exudat occurs in the early stage of DR. It is an important signal that the patient is at risk of being blind. Therefore, it is a key link for achieving automatic detection of DR. Figure 1 shows retinal image and hard exudate lesion’s pixel-level label in the E-Ophtha dataset.

The FCN is a deep learning–based method which has great performance in segmentation tasks. But there are some limitations in the convolutional neural network. As convolution operator has a local receptive field, global information will not be concerned in this process, which leads to the fact that long-range dependencies among different regions in the image can only be processed after passing through several layers. Small and shallow neural network will not able to learn and represent long-range dependencies across different image regions. Even for the deep model, there are still some problems that prevent the model from learning long-range dependencies. For example, optimization algorithms may have trouble discovering parameter values that carefully coordinate multiple layers to capture these dependencies, and these parameterizations may be statistically brittle and prone to failure when applied to previously unseen inputs [2]. Compared with the segmentation tasks in complicated scenes, there is less semantic information in retinal images. It is difficult to distinguish exudate area and some highlight fundus tissues and isolated highlight area by local features solely. It is necessary to establish long-range...
dependencies between the detected area and the surrounding area. We try to retain the advantages of the FCN and remedy the weakness of convolution layers to improve the precision of the hard exudate segmentation result. The basic model structure we proposed is an encoder–decoder model with attention mechanism. Encoder–decoder model with skip connection is one of the most common segmentation models. Encoder extract features from the images. Decoder maps the low-resolution features extracted by the encoder to the high-resolution pixel space. Skip connection structure combines the high-resolution feature maps with the corresponding size feature maps from the decoder to enhance the segmentation details. Based on the improved encoder–decoder model, two different attention mechanism modules are used. The position attention module is used in the encode stage to build dependencies among different positions by global feature relationships. Channel attention module is proposed, and it is used in the decode stage to build dependencies among different channels based on the correlations of all the channels. The improved encoder–decoder model with attention mechanism proposed is an end-to-end method. Compared with the traditional method, it avoids complicated feature design. Compared with previous deep learning methods, the proposed method established position-wise and channel-wise long-range dependencies. Based on the global information, the proposed method has better performance than the state-of-the-art methods.

2 RELATED WORK

2.1 Segmentation model

The early application of neural networks in image segmentation task is implemented by fully connected neural networks. Features are extracted in the neighbourhood of each pixel. Then, fully connected neural network was trained as a classifier to determine the category of each pixel based on the features. This method is inefficient and the segmentation result is highly dependent on the quality of the artificial designed features. With the development of neural networks, the convolutional neural network replaced the fully connected neural network in the segmentation task. For a given pixel, patch centred on the pixel is input into the model to extract features, a fully connected layer is trained as a classifier. Application of convolutional neural network avoids the complicated artificial feature extraction process. But in practical applications, for a given image, overlapping image patches causes repeated storage and redundant convolution computation [3]. Long et al. [4] proposed the fully convolutional network (FCN), which was more precise and efficient. From then on, almost all of semantic segmentation studies adopted this basic structure. Deeplab series model [5–7] showed impressive results in the semantic segmentation task of complex scenes by the use of dilated convolution, conditional random field, atrous spatial pyramid pooling, and other techniques. Skip-connection and feature fusion are common ways to improve segmentation results. For example, U-net [8] is widely used in clinical image segmentation. With dense encoder–decoder FCN structure and skip-connection, U-net has a great advantage in tiny object segmentation tasks.

2.2 Attention mechanism

Attention mechanism was first proposed by Benigio et al. [9]. Since then, the application of attention mechanism has attracted researchers in natural language processing. Similar to human attention, attention mechanism intends to screen the high-value information which is most useful for the current tasks in the overall information received. Vaswani et al. [10] proposed a method for establishing global dependencies of input information using self-attention mechanism and applied it in machine translation. The attention mechanism was widely used in the field of natural language processing in the early years. In recent years, the attention mechanism has attracted researchers in the computer vision field. Wang et al. [11] used the attention module named non-local block to establish the temporal and spatial dependencies of video sequences. This method greatly improved the video classification performance compared with previous methods. Zhang et al. [2] introduced self-attention mechanism in a generative adversarial network (GAN) to generate consistent scenes or objects by using long-distance complementary features of images.

2.3 Hard exudate segmentation

Automatic hard exudate segmentation has received significant attention during the past few years. Sopharak et al. [12] employed histogram equalization and morphological reconstruction for exudate segmentation. Sinthanayothin et al. [13] proposed a recursive region growing algorithm combined with adaptive intensity thresholding to detect the candidates of lesion areas. Both of the methods above represent simple image processing method for this task. The advantage of the methods like these is easy to grasp and most of them have less requirement of computing resources. But at the same time, the scope of its applicability is narrow and it has limited ability to adapt to the change of images. Sanchez et al. [14] used a mixture model
combined with dynamic thresholding algorithm for hard exudate segmentation. This kind of method combined both image process method and pattern recognition method and formed a more robust method, which has great adaptive capacity for different images. In general, they have better performance than simple image processing method in most of the scenes. For these methods, the most critical part is feature extraction, the quality of the features affects the performance of the model. Deep learning methods raised these years provided great solutions for segmentation problems. Compared with traditional pattern recognition method, it solves feature extraction problem which is one of the key challenges of the related methods. Mo et al. [15] used FCN with skip connection to improve the segmentation result. Chudzik et al. [16] proposed automatic an hard exudate segmentation method based on a deep fully convolutional neural network with PCA-dimensional reduction which performed better than all other methods before. Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. Zheng et al. [17] used an FCN-based method with GAN to improve the performance of segmentation model. The FCN was used as a generate model for exudate segmentation, and another convolutional network was used as an adversarial to distinguish the segmentation result from FCN and the ground truth. These two different networks working against each other and promote each other, and the proposed model will have better performance on segmentation task. As illustrated above, convolutional neural network’s local receptive field make it ineffective to build long-range dependencies and it limits deep model's performance on segmentation task, Zheng et al. [17] use the adversarial network to promote model performance, but in practice, GAN is hard to train, and it is still unsolved for establishing long-range dependencies. Most of the hard exudate segmentation methods in the published works rely on multi-layer feature fusion to improve the segmentation details. But during the feature extraction process and mapping process, the relationship between different locations is not considered. Therefore, a hard exudate segmentation method with attention mechanism is proposed in this paper.

3 | METHOD

3.1 | Model structure

The basic segmentation network is a fully convolution network like U-net. For a given three-channel retinal image, the model outputs a probability map which has the same size of the given image. Each element of the probability map represents the probability that the corresponding pixel of the retina image belongs to the lesion area. The most striking characteristic of the model is dense skip connection. Skip connection is an essential part for the hard exudate segmentation, the shallow layers provide low-level features like boundary information, and feature fusion significantly improves segmentation results. Different kinds of attention modules are added to the encoder–decoder network. Figure 2 shows the structure of the proposed model.

Basic block1 consists of residual block and downsampling unit, basic block2 consists of residual block and upsampling unit. Compared with standard U-net structure, max-pooling layers are replaced by stride convolutional layers as the downsampling unit to reduce the loss of information. Batch normalization layers [10] are added after each convolutional layer to reduce the complexity of model training and the risk of overfitting. To reduce computing cost, concat in skip connection is replaced by element-wise summation. In order to prevent the network from degrading, residual blocks [18] are used as the basic block in the encoder and decoder parts.

3.2 | Position attention mechanism

In ordinary FCN, the output of a layer is the convolution result on the feature map from the upper layer. Each point of the output feature map depends only on the local area of the upper level feature. It is inefficient for the convolutional neural network to establish the long-range dependencies [2]. So, the position attention mechanism is used in the shallow layer to build
long-range dependencies across different regions of the feature map. Position attention mechanism is defined as [11, 19] and the input–output relations of position attention module is expressed as (1).

$$y_j = \frac{1}{C(x)} \sum_{i} f(x_i, x_j) g(x_j).$$ (1)

Here, $x$ is the input feature map of position attention module, $x \in R^{T\times H\times W}$, $x_j$ is a vector composed of each channel in $j$th position of the input feature map, $y_j$ is a vector of the corresponding position in the output feature map. According to the formula, each output of the position attention block is related to all the positions of the input feature map. For convolution layers in the deep model, only deep layers have a large receptive field. If the position attention block is used before superficial convolution layers, then all the layers will have global receptive field. 

$\{x_i, y_j\} \in R^T$, $f(x_i, x_j)$ is a similarity measure function, it is used to measure the similarity of two vectors, $C(x)$ is the normalization coefficient, and we choose $C(x)$ same as it in [11]. $g(x_j)$ is a linear function which computes a representation of the input feature at the position $j$.

$$f(x_i, x_j) = \theta(x_i) ^ T \phi(x_j),$$ (2)

$$C(x) = \sum_{j} f(x_i, x_j),$$ (3)

$$g(x_j) = W^g x_j.$$ (4)

$\theta(x_i), \phi(x_j)$ in $f(x_i, x_j)$ are linear functions as $g(x_j)$, where $\theta(x_i) = W^\theta x_i, \phi(x_j) = W^\phi x_j$. To reduce the amount of calculation, for the input feature map with $T$ channels, it is compressed into $T/2$ channels. For position $x_i$ in input feature map, $x_j \in R^T$ and $g(x_j) \in R^{T/2}$, $T/2$ is expressed as $\tilde{T}$ and we have parameters $\{W^\theta, W^\phi, W^g\} \in R^{T \times T/2}$ obtained by training process.

For position attention mechanism, each position of the output feature map is determined by all the positions in the input feature map as shown in Figure 3.

The pseudocode of position attention mechanism algorithm is shown in following chart.

It is time consuming and inefficient to traverse all the positions on the input feature map in a loop. To embed position attention module into the network, position attention is implemented through parallel matrix operation. The implementation of the operation process of position attention mechanism is shown in Figure 4.

**Algorithm 1** The pseudocode of position attention mechanism

1: function "position attention"
2: for $i \in R^{T\times H\times W}$ in each point of $X$ do
3: $X_{i} = W_{g} \cdot X_{i}$
4: $X_{i} = W_{c} \cdot X_{i}$
5: for $i \in R^{T\times H\times W}$ in $X$ do
6: $X_{i} = W_{g} \cdot s_{i}$
7: $M_{i} = c_{i}^{T} \cdot X_{i}$
8: end for
9: $M_{i} = \left( 1 / \sum e^{M_{i} \cdot g_{i}} \right.$
10: end for
11: $X_{C} = \left( X_{C1}, \ldots, X_{C12} \right)$
12: $N = H \times W$
13: for $i \in range N$ do
14: $Y_{i} = M_{i} \cdot X_{C}$
15: end for
16: return $Y + X$
17: end function

As illustrated in Figure 4, $X \in R^{T\times H\times W}$, $x_i$ is an input feature map. All the feature points in the input feature map share weights $W^\theta, W^\phi, W^g$, and $1 \times 1$ convolutional layers with $T$ input channels and $\tilde{T}$ output channels are used to implement $\theta, \phi,$ and $g$. Given a feature map, first, it is fed into a $1 \times 1$ convolutional layer. $\{A, B, C\} \in R^{T \times H \times W}$ corresponds to the outputs of $\theta, \phi, \phi$, and $g$, respectively. Then, $A, B, C$ are reshaped.
and \( \{A, B, C\} \in R^{T \times N} \), here \( N \) is the size of the feature map, \( N = H \times W \). After that, \( A, C \) are transposed to \( \{A, C\} \in R^{N \times T} \). Then, a matrix multiplication between \( A \) and \( B \) is applied and SoftMax operation is performed on each row to generate attention map \( AM \in R^{N \times N} \), each element of attention map \( AM_{ij} \) can be expressed as (5).

\[
AM_{ij} = \frac{\sum_{k=1}^{T} A_{ik} B_{kj}}{\sum_{k=1}^{N} \sum_{l=1}^{T} A_{il} B_{kl}}.
\]

Here, \( AM_{ij} \) represents the \( ij \)th position that impacts on the \( i \)th position. Matrix multiplication between \( AM \) and \( C \) is applied, the result is reshaped and fed into 1 \( \times \) 1 convolutional layer for channel transformation and then added back to the input feature map to obtain \( Y \) as the output of the position attention module.

### 3.3 Channel attention mechanism

For image semantic segmentation tasks, each channel of the feature map can be regarded as the response of certain semantic features. Establishing dependencies among different semantic features is meaningful as interdependent features can be used to improve the feature’s representation of specific semantics. Based on this motivation, the channel attention mechanism is proposed.

Similar to the position attention mechanism, each channel of the channel attention module’s output is a weighted sum of different channels in the input feature map as shown in Figure 5.

The input–output relations of channel attention module proposed can be expressed as (6):

\[
y_i = \frac{1}{C(x)} \sum_{j} f(x_i, x_j) x_j.
\]

Here, \( x_i \) is one of the channels of input feature map, \( y_i \) is the corresponding channel of the output feature map. \( x_i \) and \( y_i \) are expanded into one-dimensional column vectors for reducing calculation amount, \( \{x_i, y_i\} \in R^N \), \( N \) is the size of feature map, \( f(x_i, x_j) \) is the correlation measure function (7):

\[
f(x_i, x_j) = e^{E((x_i - E(x_i))(x_j - E(x_j)))}.
\]

The pseudocode of the channel attention mechanism algorithm is shown in the following chart (Algorithm 2).

The implementation process of the channel attention mechanism with matrix operations is shown in Figure 6.

Finally, the input feature map \( X \in R^{T \times H \times W} \) is reshaped to \( A \in R^{T \times N} \). The \( i \)th row of \( A \) corresponds to the vector expansion from the \( i \)th channel of the original feature map. The expected value of the \( i \)th channel is approximated by the average of the channel. Original feature map \( X \) is fed into the global pooling layer, the result is expanded to \( B \in R^{T \times N} \), each row of \( B \) is the average of the corresponding channel in the original feature map. The result of element-wise subtraction between \( A \) and \( B \) performs matrix multiplication with the transpose of itself and then each element is multiplied by the factor \( 1/N \) to get the covariance matrix for each channel. Finally, a SoftMax layer is applied to get the channel attention map \( CM \in R^{T \times T} \),

**FIGURE 5** Channel attention mechanism schematic diagram

---

**Algorithm 2** The pseudocode of channel attention mechanism

**Input:** Input feature map \( X \) with \( T \) channels

**Output:** Attention feature map \( Y \) integrates various features.

1. function: ”channel attention”
2. for \( X_i \in R^{T \times H} \) in each channel of \( X \) do
3. \( N = W \times H \)
4. \( X_{il} = X_i \) reshape to \( R^{N \times 1} \)
5. for \( X_{il} \in R^{T \times W} \) in each channel of \( X \) do
6. \( X_{il} = X_{il} \) reshape to \( R^{N \times 1} \)
7. \( M_{ij} = e^{N[X_{il} - \bar{x}_i]T(X_{il} - \bar{x}_i)} \)
8. end for
9. \( M_i = \left( \frac{M_{i1}}{\sum_{j} M_{ij}}, \ldots, \frac{M_{iT}}{\sum_{j} M_{ij}} \right) \)
10. end for
11. \( X_4 = (X_{11}, \ldots, X_{4T}) \)
12. for \( k \) in range \( T \) do
13. \( Y_k = M_k \cdot X_4^T \)
14. end for
15. return \( Y + X \)

The covariance of \( x_i \) and \( x_j \), \( E((x_i - E(x_i))(x_j - E(x_j))) \) is approximated by the mean of \( x_i \), \( f(x_i, x_j) \) is expressed as (8) in this way, here \( \bar{x}_i \) is the mean vector of \( x_i \):

\[
f(x_i, x_j) = \frac{e^{\frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{N}}}{N}.
\]

**C(x)** is a normalization coefficient like in position attention mechanism.

\[
C(x) = \sum_{j} f(x_i, x_j).
\]
Each element $C_{Mi}$ in $CM$ can be expressed as (10).

$$CM_{ij} = \frac{\sum_{k=1}^{N}(A - B)_{ik}(A - B)_{kj}'}{\sum_{k=1}^{N}(A - B)_{ik}(A - B)_{kj}}. \quad (10)$$

Here, $CM_{ij}$ represents the correlation between the $j$th channel and the $i$th channel of the input feature map. Matrix multiplication between $CM$ and $A$ is applied, the result is reshaped and then added back to the input feature map to obtain $Y$ as the channel attention module’s output.

### 3.4 Data processing method

The training process of the deep model requires a large number of training samples. For hard exudate segmentation tasks, the number of available samples with the expert labels is limited so data augmentation is necessary.

Medical images generally have large size. It is a common practice in similar tasks to cut the original image into patches. In this way, each patch is treated as a training sample. During the past years, $32 \times 32$ is the commonly used patch size in hard exudate segmentation task. Due to the fact that attention mechanism is used to build position-wise and channel-wise long-range dependencies, the training sample should be large enough to provide sufficient information. So, $256 \times 256$ patches are used to train the model.

First of all, the image is augmented by flip and rotate, then a sliding window is used to cut the images into patches. There is a large number of patches without lesion. In order to control the number of samples and increase the proportion of valid samples, 50% of the non-lesion patches are randomly discarded.

### 3.5 Dice cross entropy loss function

For most of the image samples, hard exudate areas only account for a small portion of the whole image. Although 50% of the non-lesion patches are discarded, there is still a large number of patches without hard exudate. As shown in Figure 7, the number of pixels in the lesion area is much smaller than those in the non-lesion area.

Cross entropy loss is the most common loss function during the training process of the deep model. Based on the data processing mentioned above, cross entropy loss for each patch is calculated as (12).

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i). \quad (11)$$

$N$ is the size of current patch, $y_i$ is the label of the $i$th pixel, $y_i = 1$ when the pixel belongs to the lesion area, otherwise $y_i = 0$, $p_i$ is the model’s predictive value of the pixel (similarly hereinafter). In general, $\sum_{i=1}^{N} (1 - y_i)$ is much larger than $\sum_{i=1}^{N} y_i$, the pixels in lesion area have little contribution to the overall loss function. In this way, the training process focuses more on learning features of the non-lesion area, which make the network’s predictions be biased towards the non-lesion areas. Dice cross entropy loss function is proposed to solve this problem.

The dice coefficient is a similarity measurement of sets. Dice is used as an evaluation metric for segmentation result in some segmentation tasks. For binary segmentation, the dice coefficient is defined as (11).

$$Dice = \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|}. \quad (12)$$
Here, y is the label and \( \hat{y} \) are binary matrices with 0 or 1. \( \hat{y} \) is obtained by threshold operation on the probability map generated by the model. The threshold function is non-differentiable. In order to optimize the network’s weights with a gradient descent method, the loss function must be derivable. Milletari et al. [20] used soft dice as their loss function.

\[
S_{\text{SoftDice}} = \frac{2 \sum_{i=1}^{N} \hat{p}_i y_i}{\sum_{i=1}^{N} \hat{p}_i^2 + \sum_{i=1}^{N} y_i^2},
\]

(13)

As mentioned above, there are a large number of patches without lesion pixels. These patches are important for the training process too, but dice cannot be calculated in these patches. We propose dice cross entropy loss function (DCEL).

\[
DCEL = -y \log(D_1) - (1 - y) \log(D_2),
\]

(14)

\[
D_1 = \frac{2 \sum_{i=1}^{N} \hat{p}_i y_i}{\sum_{i=1}^{N} \hat{p}_i^2 + \sum_{i=1}^{N} y_i^2},
\]

(15)

\[
D_2 = \frac{2 \sum_{i=1}^{N} (1 - \hat{p}_i)(1 - y_i)}{\sum_{i=1}^{N} (1 - \hat{p}_i)^2 + \sum_{i=1}^{N} (1 - y_i)^2}.
\]

(16)

Dice cross entropy loss function has the similar form with normal cross entropy loss. In DCEL. We set extra label \( Y \) to each patch. Patches with hard exudate have label \( Y = 1 \), patches without hard exudate have label \( Y = 0 \). As shown in (16), the dice of patches without lesion are calculated by inverse processing.

During the training process, the sum of dice entropy loss function and normal cross entropy loss function is used as an objective function. This method not only makes the training process pay more attention to the lesion area but also considers the non-lesion area. Dice cross entropy loss function alleviates the problem caused by imbalanced sample distribution. Moreover, experiments show that the loss function proposed have greatly improved the efficiency of the model training process.

4 | EXPERIMENTS

4.1 | Materials

To verify the effectiveness of the proposed method, the proposed method is evaluated on two public datasets with pixel-level labels: e-ophtha [21] and HEI-MED [22]. This two datasets can be downloaded free from the internet.

E-ophtha is a suitable dataset to evaluate our model because experienced professionals have labelled all the images in this dataset in pixel level. E-ophtha provides 82 retinal images pixels with 45° field of view. With four different resolutions, ranging from 1440 \( \times \) 960 pixels to 2544 \( \times \) 1696, among which contain 47 images with exudate and 35 normal images. All the labels in the e-ophtha are provided as a lossless compression black and white picture.

HEI-MED consisting of 169 retinal images is a publicly available dataset developed by Giancardo, which is captured at a resolution of 2196 \( \times \) 1958 pixels with 45° field of view from patients of different age groups. All the labels in HEI-MED are provided as mat files, all the labels can be viewed and saved by officially provided matlab script [23]. The diversity of the images makes HEI-MED a good dataset for model training.

4.2 | Evaluation metrics

Current evaluation metrics for hard exudate segmentation task are sensitivity (SE), specificity (SP), positive predictive value (PPV), and F1.

\[
\text{SE} = \frac{TP}{TP + FN},
\]

(17)

\[
\text{SP} = \frac{TN}{TN + FP}.
\]

(18)

\[
\text{PPV} = \frac{TP}{TP + FP}.
\]

(19)

\[
F1 = \frac{2 \times \text{SE} \times \text{PPV}}{\text{SE} + \text{PPV}}.
\]

(20)

It should be noted that TP is the number of pixels that belong to the lesion area and are predicted correctly; TN is the number of pixels that belong to the healthy area and are predicted correctly; FP is the number of pixels that belong to the healthy area and are predicted mistakenly; FN is the number of pixels that belong to the lesion area and are predicted mistakenly.

Some previous studies also used accuracy (ACC) to evaluate their methods. Most of the methods can achieve high ACC and SP because of the imbalanced data. These two metrics are not practical for comparing the models’ performance on hard exudate segmentation task. SE, PPV, and F1 are instructive evaluation metrics. Neither SE nor PPV can be used to evaluate the performance of the models independently. F1 is calculated by both SE and PPV, it strikes a balance between sensitivity and positive predictive value. In summary, F1 is the most important metric to evaluate the performance of the hard exudate segmentation method.

We use the evaluation method proposed by Zhang et al. [24] to calculate TP, FP, TN, and FN.

\[
D = \{D_1, D_2, D_3, ..., D_N\}
\]

is a collection of single connected regions of the segmentation result. \( G = \{G_1, G_2, G_3, ..., G_N\} \) is a collection of single connected regions of the label.

A pixel is considered as a TP if and only if it belongs to set \( D \n G \)

\[
\{D \cap G\} \cup \left\{D_i \left| \frac{|D_i \cap G|}{|D_i|} > \sigma \right. \right\} \cup \left\{G_j \left| \frac{|G_j \cap D|}{|G_j|} > \sigma \right. \right\}.
\]

(21)
TABLE 1 Performance comparison on e-ophtha with pixel-based metric

| Method           | Year | SE   | SP   | ACC  |
|------------------|------|------|------|------|
| Toth [25]        | 2015 | 0.434| –    | –    |
| Ardiyanto [26]   | 2016 | 0.169| 0.996| 0.994|
| Kusakunniran [27]| 2018 | 0.564| 0.999| 0.999|
| Proposed         | 2019 | 0.656| 0.999| 0.999|

Bold terms marked the highest score.

A pixel is considered as a FP if and only if it belongs to set (22).

\[ \{D_j\} \cup \left\{ \frac{|D_j \cap G|}{|D_j|} \leq \sigma \right\} \quad (22) \]

A pixel is considered as a FN if and only if it belongs to set (23).

\[ \{G_j\} \cup \left\{ \frac{|G_j \cap D|}{|G_j|} \leq \sigma \right\} \quad (23) \]

All other pixels are considered as TN.

4.3 System implementation

The proposed method is implemented on the MXNet framework. The model is trained on two NVIDIA GTX 1080Ti. During the training process, Adam optimizer with parameters 0.9 and 0.999 is used. The learning rate is set to be 0.001. During the training process, weight decay with \( \lambda = 0.001 \) is applied to prevent the model from over-fitting.

In the test stage, each image in the test set is cut into patches. All the predicted probability patches are stitched together. These patches are not isolated, they overlap with each other. So about the predicted value on overlapping parts, they can be calculated by the average of overlapping pixels.

4.4 Result comparison

Figure 8 shows several segmentation results generated by proposed method.

To prove the validity of the proposed method, comparison of the proposed method with several methods on e-ophtha dataset with two different evaluation metrics is shown in Table 1 and Table 2, respectively. Methods compared in each table used totally the same dataset and the same evaluation metric, in Table 1 and Table 2, fivefold cross validation is adopted to evaluate our method’s performance.

For HEI-MED dataset, most of the exiting methods used region-based method to evaluate proposed methods. Table 3 shows the comparison of the proposed method and other methods on HEI-MED dataset.

Another deep learning–based method proposed by Zheng et al. [17] called Mu-net used 60 images to train the model and 22 images to test rather than use fivefold cross validation. We compared the proposed method with it in Table 4.

The proposed method performs better than most of the methods. Deep learning–based methods are more robust than simple image processing method, and deep model generate better features compared with traditional pattern recognition method. Compared with other deep learning–based methods, the proposed method used bigger patches, to make sure the model can get enough information around each pixel to distinguish lesion area and some highlighted area. More importantly, attention modules build relationships of different features on large scale. In this way, different kinds of features can be
connected to enhance the ability of features expression. For hard exudate segmentation, clinical experience indicates that exudate often appear as many discrete yellow dots, and a single point which has the same appearances is not exudate in common situations. Position attention mechanism used can work out this problem, because position attention module can build dependency across features in different positions, the output feature of the module contains information of surrounding area of each point to describe exudate more accurately.

Although the proposed method shows great performance, there is an inevitable problem, the proposed method is time consuming, which is a common problem in deep learning method. It costs 5 s to predict an image with the proposed method. That is a long time compared with traditional methods. There are two reasons for this. Firstly, attention model is time consuming because of the great amount of calculation. For a feature map with $n$ positions, position attention map have $n^2$ elements, for a feature map with $c$ channels, channel attention map have $c^2$ elements. Position attention module is placed in the shallow layer in general condition, so $n$ is much larger than $c$, although we use position attention module in shallow layer, it still cost so much time. Secondly, in the predict stage, internal storage of the device cannot hold that large amount of calculation, so we cut original image into $256 \times 256$ patches to predict, it cost about 0.005 s to predict a patch. But this process have to repeat hundreds of times to predict the whole image, as shown in Table 5, the larger the test image size, the longer the computation time. In general, it is acceptable for static images like retinal image.

In order to make clear the relationship of all the factors and their effect on hard exudate segmentation, the contrast experiment analyses are carried out. Table 6 shows the experimental results and Figure 9 compares the performance of different models in detail. Obviously, we got much higher PPV and better segmentation results with the proposed method.

To verify the advantage of the dice cross entropy loss function and improved Unet proposed, Unet and improved Unet are trained on the same onefold data with two different loss functions. Figure 10 shows different models with different loss functions training processes’ variation of the pixel-based F1 score with the number of training epochs.

For the commonly used methods, they have some obvious flaws. The problems of the cross entropy loss in hard exudates segmentation have been analyzed. As shown in Figure 10, training model with cross entropy loss function is inefficient, for hard exudates segmentation task, U-net cannot be trained effectively, F1 score of U-net training process with cross entropy loss function almost stagnates in the previous rounds of training. Improved Unet has a great contribution to reduce the difficulty of the training process, as it improves the convergence speed and the model’s performance on segmentation task.

### TABLE 3  Performance comparison on HEI-MED with region-based metric

| Method      | Year | SE   | PPV | F1  |
|-------------|------|------|-----|-----|
| Imani [22]  | 2016 | 0.812| 0.636| 0.713|
| Pereira [33]| 2015 | 0.808| 0.731| 0.763|
| Prentavisic [34]| 2016 | 0.858| 0.734| 0.791|
| Mo [15]     | 2018 | 0.926| 0.821| 0.849|
| Proposed    | 2019 | 0.871| 0.867| 0.862|

Bold terms marked the highest score.

### TABLE 4  Performance comparison with Mu-net on e-ophtha dataset

| Method      | Year | SE   | PPV  | F1  |
|-------------|------|------|------|-----|
| Mu-net [17] | 2018 | 0.909| 0.947| 0.928|
| Proposed    | 2019 | 0.928| 0.941| 0.932|

Bold terms marked the highest score.

### TABLE 5  Computation time on images with different sizes

| Dataset | Image size | Time (s) |
|---------|------------|----------|
| E-ophtha| $1440 \times 960$ | 0.422 |
| E-ophtha| $2196 \times 1958$ | 1.316 |
| HEI-MED | $2544 \times 1696$ | 1.316 |
| IDRID   | $4288 \times 2848$ | 3.727 |

### TABLE 6  Performance comparison on different models

| Model                      | SE   | PPV  | F1  |
|----------------------------|------|------|-----|
| Unet                       | 0.85 | 0.75 | 0.75|
| Improved Unet              | 0.88 | 0.76 | 0.78|
| Improved Unet + position attention module | 0.89 | 0.78 | 0.80|
| Improved Unet + channel attention module | 0.89 | 0.75 | 0.79|
| Improved Unet + two attention modules | 0.89 | 0.81 | 0.82|

Bold terms marked the highest score.
5 | CONCLUSION

As a worldwide public health problem, DR is the main reason which causes preventable blindness. Hard exudate is one of the earliest signs of DR. Precise detection of hard exudate is helpful for early diagnosis of DR. Automatic hard exudate segmentation is the key link for achieving automatic detection of DR. Deep learning-based method FCN has impressive performance on segmentation task, but there are some problems in FCN. We explained the necessity of building long-range dependencies among different positions and channels in hard exudate segmentation task. A hard exudates segmentation method based on FCN with attention mechanism is proposed to remedy the weakness of convolutional neural network. It is the first time that the related method is applied to hard exudates segmentation task, we proposed another channel attention module based on covariance matrix. Improved U-net with two attention modules have a great performance on hard exudate segmentation task.

For the data imbalanced problem, a novel loss function is proposed. Proposed loss function allowed the training process to focus more on the lesion area and will not ignore non-lesion area, so the model can learn more essential features, which make the training more efficient. We believe dice cross entropy loss function can be used on other tasks with imbalance data.

The experiment result shows the efficiency of the proposed method, proposed models have better performance than the other segmentation method and dice cross entropy loss function that make training process more efficient. However, there is a minor deficiency. The time consumed of the proposed method should not be overlooked, serial processing spends too much time waiting for the patch queue. Further research on parallel processing is indispensable.

ACKNOWLEDGEMENT

This work was supported by Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology, Beijing (No. BG0150).

REFERENCES

1. Yilmaz, S., et al.: Changes of normal appearing optic nerve head on diffusion-weighted imaging in patients with diabetic retinopathy. Clin. Imaging 42, 60–63 (2017)
2. Zhang, H., et al.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318, 2018
3. Liu, Y., et al.: Optic disc segmentation in fundus images using adversarial training. IET Image Process. 13(2), 375–381 (2018)
4. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. pp. 3431–3440 (2015)
5. Chen, L.C., et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062, 2014
6. Chen, L.C., et al.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017
7. Chen, L.C., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2018)
8. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Munich, Germany. pp. 234–241 (2015)
9. Bhodan, D., et al.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014
10. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015
11. Wang, X., et al.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, Utah. pp. 7794–7803 (2018)
12. Sopharak, A., et al.: Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods. Comput. Med. Imaging Graph. 32(8), 720–727 (2008)
13. Sinthanayothin, C., et al.: Automated detection of diabetic retinopathy on digital fundus images. Diabet. Med. 19(2), 105–112 (2002)
14. Chudzik, P., et al.: Exudates segmentation using fully convolutional neural networks and auxiliary codebook. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Honolulu, HI, USA. pp. 770–773 (2018)
15. Zheng, R., et al.: Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. Biomed. Opt. Express 9(10), 4863–4878 (2018)
16. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas. pp. 770–778 (2016)
17. Qu, Z., et al.: Learning recurrent structure-guided attention network for multi-person pose estimation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 418–423 (2019)
18. Milletari, F., et al.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571 (2016)
19. Decencière, E., et al.: Teleophta: Machine learning and image processing methods for teleophthalmology. Irbm 34(2), 196–203 (2013)
20. Imani, E., Pourreza, H.R.: A novel method for retinal exudate segmentation using signal separation algorithm. Comput. Methods Programs Biomed. 133, 195–205 (2016)
21. Giancardo, L., et al.: Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med. Image Anal. 16(1), 216–226 (2012)
24. Zhang, X., et al.: Exudate detection in color retinal images for mass screening of diabetic retinopathy. Med. Image Anal. 18(7), 1026–1043 (2014)
25. Toth, J., et al.: Improving the performance of an ensemble-based exudate detection system using stochastic parameter optimization. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Milan, Italy. pp. 5243–5246 (2015)
26. Ardiyanto, I., et al.: Maximum entropy principle for exudates segmentation in retinal fundus images. In: 2016 International Conference on Information & Communication Technology and Systems (ICTS). IEEE, Surabaya, Indonesia. pp. 119–123 (2016)
27. Kusakunniran, W., et al.: Hard exudates segmentation based on learned initial seeds and iterative graph cut. Comput. Methods Programs Biomed. 158, 173–183 (2018)
28. Walter, T., et al: 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), 1236–1243 (2002)
29. Welfer, D., et al.: A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images. Comput. Med. Imaging Graph. 34(3), 228–235 (2010)
30. Das, V., Puhan, N.B.: Tsallis entropy and sparse reconstructive dictionary learning for exudate detection in diabetic retinopathy. J. Med. Imaging 4(2), 024002 (2017)
31. Chudezik, P., et al.: Exudate segmentation using fully convolutional neural networks and inception modules. In: Proceedings SPIE 10574, Medical Imaging 2018: Image Processing, 1057430, 2 March 2018
32. Liu, Q., et al.: A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images. Comput. Med. Imaging Graph. 55, 78–86 (2017)
33. Pereira, C., et al.: Exudate segmentation in fundus images using an ant colony optimization approach. Inform. Sci. 296, 14–24 (2015)
34. Prentašić, P., Lončarić, S.: Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. Comput. Methods Programs Biomed. 137, 281–292 (2016)

How to cite this article: Si Z, Fu D, Liu Y, Huang Z. Hard exudate segmentation in retinal image with attention mechanism. IET Image Process. 2021;15:587–597. https://doi.org/10.1049/ipr2.12007