MSLF-Net: A Multi-Scale and Multi-Level Feature Fusion Net for Diabetic Retinopathy Segmentation

Haitao Yan 1, Jiexin Xie 1, Deliang Zhu 2, Lukuan Jia 2 and Shijie Guo 1,2,*

1 Academy for Engineering & Technology, Fudan University, Shanghai 200433, China
2 School of Mechanical Engineering, Hebei University of Technology, Tianjin 300132, China
* Correspondence: guoshijie@fudan.edu.cn

Abstract: Diabetic Retinopathy (DR) is a diabetic complication that predisposes patients to visual impairments that could lead to blindness. Lesion segmentation using deep learning algorithms is an effective measure to screen and prevent early DR. However, there are several types of DR with varying sizes and high inter-class similarity, making segmentation difficult. In this paper, we propose a supervised segmentation method (MSLF-Net) based on multi-scale–multi-level feature fusion to achieve accurate end-to-end DR lesion segmentation. MSLF-Net builds a Multi-Scale Feature Extraction (MSFE) module to extract multi-scale information and provide more comprehensive features for segmentation. This paper further introduces the Multi-Level Feature Fusion (MLFF) module to improve feature fusion using a cross-layer structure. This structure only fuses low- and high-level features of the same class based on category supervision, avoiding feature contamination. Moreover, this paper produces additional masked images for the dataset and performs image enhancement operations to ensure that the proposed method is trainable and functional on small datasets. The extensive experiments are conducted on public datasets IDRID and e_ophtha. The results showed that our proposed feature enhancement method can perform feature fusion more effectively. Therefore, In the end-to-end DR segmentation neural network model, MSLF Net is superior to other similar models in segmentation, and can effectively improve the DR lesion segmentation performance.

Keywords: diabetic retinopathy; image segmentation; multi-scale; multi-level; feature fusion

1. Introduction

Diabetic Retinopathy (DR) is one of the most common retinal vascular complications of diabetes mellitus, mainly manifested by high blood sugar that causes insidious and uninterrupted damage to the retinal blood vessels [1]. DR could eventually lead to vision impairments that could potentially reach blindness in patients. However, DR has no obvious symptoms in its early stage and is not easily diagnosed. Once it progresses to an advanced stage, the condition becomes irreversible [2]. Many patients have already missed the best time for treatment by the time they discovered their impaired vision. In developing countries, the incidence of diabetic retinopathy has become the second leading cause of low vision or blindness in the adult working population after hereditary [3].

Ophthalmologists usually look for DR lesions in fundus images to diagnose the condition. In clinical practice, there are four types of abnormal lesions in DR: Micoraneurysms (MA), Hemorrhage (HE), Soft Exudate (SE), and Hard Exudate (EX) [4]. Doctors can view the type, size, distribution, and number of the four lesions via manual segmentation to determine the disease’s progression [5] (see Figure 1). However, the high cost and limited physician resources have made it difficult for the manually segmented DR screening technology to become widespread on a large scale [6]. With technology development, automatic lesion segmentation technology has gradually matured and been applied in clinical diagnosis.
The mainstream automatic DR segmentation techniques are divided into two types: non-deep and deep learning methods [7]. The prior methods rely on manual feature extraction, mainly morphological processing, classification, and region-growing methods [8–10]. However, these segmentation methods are not highly accurate; they also rely on manual feature extraction by professional doctors [11]. Therefore, it is difficult to employ them in medical diagnoses. Deep learning methods use automatic feature extraction and skip the traditional feature engineering steps [12]. This enables an end-to-end working approach and reduces the need for domain expertise. Fully convolutional neural networks (FCN) were the first deep learning methods applied for this purpose [13]. Subsequently, the emergence of U-Net pushed deep learning methods to their pinnacle [14].

Particularly, Noushin Eftekhari et al. [15] proposed a convolutional neural network-based method for MA single lesion segmentation. They used a two-stage training strategy to first select lesion candidates. They then used Convolutional Neural Networks (CNN) to classify MA and non-MA. However, their approach was time-consuming and labor-intensive. Moreover, the segmentation accuracy was not adequately high. Mo et al. [16] designed a Fully Convolutional Residual Network (FCRN) incorporating multi-level information for HE lesion segmentation. The FCRN could quickly and accurately segment lesions, avoiding extensive pre- or post-processing steps. However, the FCRN is only effective for EX lesions and not for the other three types of lesion segmentation. Li et al. [17] produced the first private dataset DDR for multi-lesion segmentation and attempted multi-lesion segmentation tasks using PSPNet [18] and DeeplabV3+ [19]. Guo et al. [20] developed an end-to-end multi-lesion segmentation network called L-seg, which is FCN-based. They also introduced a weighted fusion module and achieved promising results.

Although the neural network-based DR segmentation methods have achieved certain results, DR segmentation is still a challenging task due to the following reasons. First, four types of lesions exist in DR; they have high similarity between lesion classes and large intra-class distinctions. Moreover, the lesion sizes are extremely small and their shapes greatly vary; this makes effective feature extraction difficult. Second, the retinal structure is complex; it has anatomical structures such as blood vessels and optic discs, which are easily confused with lesions. Moreover, the retina’s size is much larger than that of DR lesions. It is a great challenge to segment small lesions with complex structures in retinal images. Third, medical images often involve patient privacy and high-cost manual annotation, resulting in a great lack of available public datasets. The quality of the available datasets is not ideal for adequate model training. Overall, it is necessary to further investigate the four lesion features of DR and propose a DR multi-lesion segmentation network with a wider application.

In this paper, we propose a Multi-Scale–Multi-Level feature (MSLF) fusion model to conduct the problem of DR segmentation. The proposed MSLF-Net contains a Multi-Scale Feature Extraction (MSFE) module and a Multi-Level Feature Fusion (MLFF) module. The former builds a pyramidal feature extraction block, which is used to extract multi-scale information and provide more comprehensive features for segmentation networks, and
the latter facilitates multi-level feature learning of the same category by a cross-layer structure that allows lower-level features of the same category to participate and enhance the semantic classification of higher-level features. MSLF-Net effectively improves the model’s segmentation accuracy for DR using the above improvements and outperforms other comparable end-to-end models in segmentation. This paper’s innovations and research contributions are summarized as:

1. This paper established effective pyramid feature extraction blocks, which can generate multi-scale features efficiently and enhance the segmentation of small lesions significantly.
2. This paper proposed the cross-layer and multi-scale class response feature fusion method, which makes the multi-scale feature fusion more effective, can adequately meet the challenges of small or complex features, avoids the complicated preprocessing work, achieves a high-performance end-to-end segmentation network, and reaches an advanced level in DR lesion segmentation.
3. This paper improved the image quality by preprocessing the field of view (FOV) mask images for the IDRID and e_ophtha dataset, so that the model can be trained and worked on small datasets, and to facilitate the future work of DR lesion segmentation.

The rest of this article is organized as follows. Section 2 elaborates the data and methodological concepts employed in this work. Section 3 records and describes the experimental results. Then, Section 4 discusses the experimental results. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Convolutional Neural Networks in DR Lesion Segmentations

CNN is widely used in medical image segmentation because it can automatically extract the most appropriate features, which improves CNN’s segmentation performance and reduces the need for professional knowledge in traditional feature extraction. The FCN network is the first convolution neural network applied to medical image segmentation (realizing the pixel-by-pixel image segmentation task for the first time). Tan et al. [21] designed a network with 10 convolution layers to segment MA, HE, and EX, but the network had too few convolution layers to extract the advanced features of lesions, and the overall segmentation accuracy was not high. Kou et al. [22] proposed a deep recursive convolution neural network, which combines residual structure with recursive convolution operation and applies it to the convolution layer. It can accumulate more effective features than typical CNNs and achieve better performance in MA segmentation tasks, but the method has a high miss detection rate. Huang et al. [23] designed a Transformer-based Relation Transformer Network (RTNet) for DR lesion segmentation. RTNet first uses a convolution network to extract DR lesion information and retinal vessel information, and then uses the cross-attention mechanism to obtain the distribution relationship between DR lesions and retinal vessels. By pathological dependence of lesions and blood vessels, the possibility of network mis-segmentation is reduced, and the highest performance is achieved in DR lesion segmentation tasks. However, RTNet needs to be divided into two phases during training, which is complex and too dependent on external information.

CNN has achieved great success in DR segmentation, however, it also faces new difficulties. On the one hand, the size of DR lesions is very small, some lesion areas are only a few pixels in size, and the CNN down-sampling operation can easily result in the loss of information about minor lesions. On the other hand, DR lesions often have extreme shapes (such as slender, narrow, and high), which make it difficult to extract features from CNN. This also limits the application of the pure CNN network in DR segmentation.

2.2. Feature Fusion in Medical Image

To solve the problems faced by the pure CNN network in extracting medical images such as DR lesions, researchers have proposed many improved methods, among which feature fusion effectively improves the segmentation performance of CNN on medical images. According to the operation mode, common feature fusion methods include element-
wise product, feature concatenation, parallel fusion of two models, and improved feature concatenation based on the attention mechanism or deconvolution, etc. U-Net is the improved feature concatenation fusion network based on skip, which combines the shallow features in the encoder with the deep features in the decoder through the skip structure. This enables U-Net to utilize both the high-resolution of the shallow features and the high-semantic information of the deep features, and has a good segmentation effect for organs or lesions with multiple scales and large morphological differences. Zhang et al. [24] uses the attention mechanism to regulate feature concatenation fusion at different semantic levels, which is superior to U-Net in detecting MA lesions. However, there is a large semantic level difference between the shallow features and the deep features, which makes it difficult for them to merge effectively, and feature pollution may even occur [25]. This makes the feature fusion method not effective for the segmentation of complex lesions. DeepFusion, proposed by Song et al. [26], provides a fuller and more comprehensive feature fusion. DeepFusion has two data input ports, one for the main data and the other for the associated data. Therefore, DeepFusion not only extracts the global and local features of information, but also fuses the features of the two datasets. Finally, DeepFusion fuses the global and local features through element-wise product, and greatly improves the accuracy and efficiency of feature extraction through two feature fusions. However, the data types used by DeepFusion differ from medical image segmentation data, making it more difficult to obtain the related information of medical image segmentation datasets. Especially for the DR lesion segmentation, there is only a single lesion segmentation dataset, and there is no corresponding associated information for the research. Ramamurthy et al. [27] designed a two-model parallel structure. One model extracts the prominent features of lesions, and the other model calibrates the features effectively. By fusing the two parallel features of models, we obtain more accurate high-level spatial information. This method has some advantages in identifying morphologically similar lesions, but it does not solve the problem (loss of micro features) caused by pooling and strided convolution.

3. Materials and Methods

In this section, the datasets and data processing methods are first introduced. Then, we give a brief overview of our proposed model, and elaborate on the components of the model. Finally, The methods for evaluating the performance of models are described.

3.1. Datasets

The proposed model MSLF-Net is evaluated on two public datasets, including IDRID [28] and e_ophtha [29]. The IDRID dataset comes from the 2018 Diabetic Retinopathy Segmentation and Grading Challenge, which is the only dataset with four kinds of DR lesions. IDRID consists of 81 cooler fundus images and lesion annotations with resolution of \( 4288 \times 2848 \). In the experiment, data were divided into training and testing sets, 54 images for training and 27 images for testing. e_ophtha provides only two types of lesion annotations, including EX lesions and MA lesions. Its images have various resolutions such as \( 2544 \times 1696 \), \( 1440 \times 960 \), etc. e_ophtha_EX consists of 47 color fundus images, and we set 32 images for training and 15 for testing. e_ophtha_MA has 148 images. The training set and testing set are 20 and 8, respectively.

3.2. Image Preprocessing and Augmentation

Image preprocessing and augmentation aim to resize the dataset images and increase image count. The images of two datasets used in the paper had high resolution and several pixel sizes, which makes it difficult to use them as input in our model. Accordingly, we resized the images to \( 1440 \times 960 \) pixels without deformation. To conduct model training on a big dataset, images were additionally augmented to enhance the dataset. We performed five random geometric transformations with a 50% probability: random vertical flip, random horizontal flip, random affine, random rotation, and random adjustments of brightness, contrast, saturation, hue in the range of 0.07.
In addition, only the circular area of the eyeball in the fundus image contains valid information, and the rest are filled with black pixels. In order to highlight the circular area of the eyeball and avoid irrelevant black pixels from affecting the network training effect, we imitated the fundus retinal datasets such as DRIVE [30] to make mask images for each image of IDRID and e_ophtha. The image of IDRID, lesion annotation, and mask are shown in Figure 2.

![Figure 2. Illustration of fundus image, DR lesions, and FOV mask image.](image)

3.3. Model Architecture

3.3.1. Model Overall

The structure of MSLF-Net was inspired by U-Net and adopts a symmetric encoder–decoder structure. The encoder part uses the pre-trained Vgg16 [31] to enhance the feature learning capability. The decoder uses transposed convolutional up-sampling. Such structure allows for effective feature integration. The model adds MFEF and MLFF to U-Net. The former fully extracts multi-scale information to enhance the model’s feature extraction capability. The latter enhances network segmentation accuracy through a cross-layer architecture that allows both lower- and higher-level features to participate in semantic classification. These two modules complement each other to achieve multi-scale feature fusion on the feature map of the same activation class. This ensures that the model has a stronger discriminative ability for lesions with variable scales and complex features. Moreover, it could obtain a finer segmentation capability. Additionally, we adopted the idea of deep supervision strategy [32] to retain the original output layer in the U-Net decoder, which is used to suppress background pixels and provide detailed lesion localization and structural information to the model. The overall structure of MSLF-Net is shown in Figure 3.

![Figure 3. The ovaerall architecture of MSLF-Net.](image)
3.3.2. Multiscale Feature Module (MFEF)

The MFEF module’s purpose is to leverage the physical contour feature information from the network’s shallow layer and the deep semantic category information from its deep layer. In general, the network’s shallow layer captures low-level information, such as points, lines, edges, and textures, while the deep layer extracts semantic category information. Additionally, it has a small down-sampling multiple and high resolution, which is suitable for segmenting small lesions. On the other hand, the deep layer does the opposite. There are lesions of different sizes in the fundus images of DR. Predictions using a single resolution image are ineffective. We borrowed the FPN [33] idea and designed a pyramidal multi-scale feature extraction module to collect the feature map information with different resolutions. To retain the extracted multi-scale feature information as much as possible, the MFEF module was designed with activation and convolution operations based on the idea of ResNet [34], and the “add” method was used to fuse the convolved feature maps before and after the operations.

The structure is shown in Figure 4. We passed each of the decoder module’s five convolution blocks through the MSFE module to obtain five new convolution blocks. The first four convolution blocks need to be up-sampled to the original map’s size via bilinear interpolation. The last convolutional block does not need to be up-sampled because its size is identical to that of the original one.

![Figure 4. The structure of the MSFE module.](image)

3.3.3. Feature Fusion Module (MLFF)

The specific operation of the MLFF module is to improve feature concatenate through feature channel rearrangement. The MLFF module was designed to avoid a brutal approach to directly fusing the multi-scale features extracted by the MSFE module, which can fully utilize the multi-scale information in a manner ensuring that the same-class activation features are fused.

In convolutional neural networks, there are two common methods of feature fusion. One operation is to add the feature maps element-by-element, such as Resnet, SSD [35], etc. The other operation is to add the feature maps by dimension and then adjust the number of channels with convolution, such as Densenet [36], U-Net, etc; this is shown in Figure 5a,b.

The above two fusion methods are only suitable for shallow feature fusion but not for deep feature fusion with existing category attributes. For image segmentation tasks, the label map is processed by One-hot coding, and the category information is sequentially mapped to different channels using a serial number. Then, the supervised output layer’s channels are also sequentially activated by category. If these feature maps are directly summed or processed with splicing operations, the activated channel category information will be destroyed. To protect each channel’s category information, the MLFF module performs a slice, splicing across the convolutional layers, instead of the corresponding regular splicing in the convolution. We wish to constrain each channel of these multi-scale feature maps implicitly through the output layer to carry the most relevant information to
the corresponding category. This guides the feature maps extracted by the MFEF module to converge to a category identity.

![Figure 5](image)

**Figure 5.** Detailed description of simple feature fusion and multi-level feature fusion.

As shown in Figure 5c, the same color images represent feature maps containing information of the same semantic category. The MLFF module classifies the channels of the convolution blocks by semantic information. The number of channels in each convolution block is identical to the number of categories, while each channel corresponds to the category activation of a class. In the feature fusion operation, the MLFF module stitches the channels of different convolutional blocks with the same category, according to the “Concatenate” connection method to highlight the information of different scales in the same category. The specific operations are as follows:

As shown in Figure 3, the MSFE module generates five convolutional blocks, each with \( c \) channels:

\[
A_o = (A_1^{(c)}, A_2^{(c)}, A_3^{(c)}, A_4^{(c)}, A_5^{(c)})
\]  

(1)

Then, MLFF spliced the channels of the same category through a slice operation across convolutional blocks to generate \( A_f \):

\[
A_f = (A_1^{(c)}, A_2^{(c)}, A_3^{(c)}, A_4^{(c)}, A_5^{(c)}, \ldots, A_1^{(c)}, A_2^{(c)}, A_3^{(c)}, A_4^{(c)}, A_5^{(c)})
\]  

(2)

Finally, \( c \)-group \( 1^*1 \) convolution is performed on \( A_f \) to get the output layer as shown in Figure 5c:

\[
Output = (A_1^{(c)}, A_2^{(c)}, \ldots, A_n^{(c)})
\]  

(3)

### 3.4. Implementation Details and Experiment Settings

In MSLF-Net, the encoder consists of five convolutional blocks, while the decoder is composed of one transition block and five up-sampling blocks. The five convolution blocks in the encoder have the same structure as Vgg16 and use its pre-training weights on the ImageNet dataset [37] to initialize. The structure of the decoder is the same as that of U-Net. The MFEF module compresses the channel of the decoder convolutional blocks to five. While the MLFF module rearranges and compresses the convolutional blocks processed by MFEF module according to the channel order to obtain the output. The detailed configuration of MSLF-Net is shown in Table 1.
Table 1. The detailed configuration of MSLF-Net.

| Stage | Layers | Input | Input Size | Operations        | Output Size |
|-------|--------|-------|------------|-------------------|-------------|
|       | Encoder| Block1| Input image| 1440 × 960 × 3   | Conv × Maxpool × 720 × 480 × 64 |
|       |        | Block2| Block1     | 720 × 480 × 64   | Conv × Maxpool × 360 × 240 × 128 |
|       |        | Block3| Block2     | 360 × 240 × 128  | Conv × Maxpool × 180 × 120 × 256 |
|       |        | Block4| Block3     | 180 × 120 × 256  | Conv × Maxpool × 90 × 60 × 512 |
|       |        | Block5| Block4     | 90 × 60 × 512    | Conv × Maxpool × 90 × 60 × 512 |
|       | Decoder| UpBlock5| Block5 | 90 × 60 × 512   | Conv × UpSample × Concat × 180 × 120 × 256 |
|       |        | UpBlock4| Block5 Block4| 180 × 120 × 512 | Concat × Conv × UpSample × 360 × 240 × 128 |
|       |        | UpBlock3| Block4 Block3| 360 × 240 × 128 | Concat × Conv × UpSample × 720 × 480 × 64 |
|       |        | UpBlock2| Block3 Block2| 720 × 480 × 128  | Concat × Conv × UpSample × 1440 × 960 × 64 |
|       |        | UpBlock1| Block2 Block1| 1440 × 960 × 64 | Concat × Conv × UpSample × 1440 × 960 × 5 |
|       | MSFE   | Layer5| UpBlock5   | 90 × 60 × 512    | Conv × UpSample × 1440 × 960 × 5 |
|       |        | Layer4| UpBlock4   | 180 × 120 × 512  | Conv × UpSample × 1440 × 960 × 5 |
|       |        | Layer3| UpBlock3   | 360 × 240 × 128  | Conv × UpSample × 1440 × 960 × 5 |
|       |        | Layer2| UpBlock2   | 720 × 480 × 128  | Conv × UpSample × 1440 × 960 × 5 |
|       |        | Layer1| UpBlock1   | 1440 × 960 × 64  | Conv × 1440 × 960 × 5 |
|       | MLFF   | Fusion| Layer 1-5  | 1440 × 960 × 25  | Fusion × 1440 × 960 × 5 |

The loss function is a hybrid loss function, including the weighted cross-entropy and dice loss functions. The loss function formula is as follows:

The weighted cross-entropy loss function formula is:

$$L_{cross} = - \sum_{c=1}^{M} \omega_c y_c \log(p_c)$$  \hspace{1cm} (4)

where, $M$ ($M = 5$) and $c$ are the numbers of class and class, respectively. $\omega$ stands for class-weight, and the value is set as 100, 100, 50, 30, 1 for MA, HE, SE, EX, and background, respectively, which is calculated by the ratio of the number of labeled pixels to the number of all pixels in the dataset. $y$ denotes the label value, and $p$ indicates the predicted probability.

The dice loss function formula is:

$$L_{dice} = c - \sum_{c=1}^{M} \frac{2 \times TP_p(c)}{2 \times TP_p(c) + FN_p(c) + FP_p(c)}$$  \hspace{1cm} (5)

where, $TP_p(c)$, $FN_p(c)$, and $FP_p(c)$ are the predicted probabilities of true positive, false negative, and false positive for each class, respectively; $c$ and $M$ have the same meaning as the above formula.

The hybrid loss function is:

$$L_{loss} = L_{cross} + \lambda \times L_{dice}$$  \hspace{1cm} (6)

where $\lambda$ is the weight ratio of two loss functions, which is set to 1.0. The influence of $\lambda$ on the segmentation performance of the model will be discussed in the ablation experiment.

We implemented the experimental architecture using PyTorch, and executed it on NVIDIA 3090 GPU. In the experiment, we trained models for 200 epochs using the SGD optimizer with momentum 0.9 and weight decay 0.0005. We adjusted batchsize to 2 and Initial learning rate to 0.01. In addition, we adopted the Poly strategy to dynamically adjust the learning rate during training. The calculation formula of the Poly strategy is as follows:

$$lr = Initial lr \times (1 - \frac{epoch}{total\_epoch})^{0.9}$$  \hspace{1cm} (7)

The epoch-wise learning rate curve and loss curve during training on the IDRID dataset are presented in Figure 6.
3.5. Model Performance Evaluation

The number of background pixels in DR images is much larger than the number of lesion pixels. For high class-imbalance, the PR Curve is concerned with True-Positive predictions. Therefore, the PR curve is generally considered to be better than other indicators in this situation. To compare the segmentation performance of different models visually, AUPR is also introduced. AUPR is the value of the area under the PR curve. The closer the AUPR value is to 1, the better the model’s predictive performance. In addition, we calculated the mean value from 5 experiments as the final experimental result to evaluate the overall model performance for multi-lesion segmentation of DR.

4. Results and Discussion

In this sub-section, we present the results of comparative experiments and ablation experiments, the former is used to evaluate the segmentation performance of MSLF-Net, and the latter revealed distinct functions associated with different modules in MSLF-Net.

4.1. Comparative Experiments

We compared our method’s segmentation performance with publicly available previous works on the IDRID dataset. The methods involved in our experiments included FRCN, UNet++ [38], MANet [39], PSPNet, DeeplabV3+, L-Seg, and RTNet. Note that RTNet has additionally introduced the retinal blood vessel dataset. We need to remove its blood vessel information and compare it with the results of RTNet-base. In addition, we introduce the feature fusion method proposed by Ramamurthy et al. [27]. The following experimental results are taken from the average of five experiments.

As shown in Table 2, our method ranked first in the AUC_PR of SE, outperforming other compared methods. It ranked second in the AUC_PR of MA, EX and HE, showing a slight disadvantage. It also achieved the highest score in the mAUC_PR metric. To sum up, Our method has achieved an advanced performance compared to other methods in segmentation of four lesions. The method of L-Seg outperformed our method in MA segmentation. This is probably because our proposed multi-scale feature extraction strategy loses some features, rendering our boosting effect to be not as good as the multi-level side supervision approach used by the L-Seg. RTNet-base propose a transformer-based attention block, which obtained detailed lesion information. RTNet-base has achieved the best results in the segmentation of HE and EX lesions. However, the transformer block pays more attention to the global relationship and ignores the local relationship, which seriously affects the segmentation effect of SE lesions. The feature fusion method proposed by Ramamurthy et al. still has problems of feature loss caused by pooling and strided convolution, so it does not perform well in DR segmentation.
Table 2. Comparison experiments based on IDRID.

| Method       | MA   | HE   | SE   | EX   | mAUC_PR |
|--------------|------|------|------|------|---------|
| FRCN         | 0.3383 | 0.4200 | 0.5152 | 0.5472 | 0.4552  |
| UNet++       | 0.3535 | 0.5472 | 0.6052 | 0.7415 | 0.5619  |
| MANet        | 0.3220 | 0.4607 | 0.6466 | 0.8294 | 0.5647  |
| PSPNet       | 0.0024 | 0.0340 | 0.1825 | 0.2483 | 0.1168  |
| DeeplabV3+   | 0.3500 | 0.6142 | 0.7416 | 0.8042 | 0.6275  |
| Paper [27]   | 0.3499 | 0.5273 | 0.6449 | 0.7229 | 0.5613  |
| L-Seg        | 0.4627 | 0.6374 | 0.7113 | 0.7945 | 0.6515  |
| RTNet-base   | 0.4279 | 0.6570 | 0.5968 | 0.8659 | 0.6369  |
| MSLF-Net     | 0.4393 | 0.6411 | 0.7597 | 0.8644 | 0.6761  |

The PR curve obtained from the experiment is shown in Figure 7. The higher the PR curve, the larger the area included, indicating a better performance of the model.

![PR Curves](image)

Figure 7. PR curves of MSLF-Net on IDRID.

We also validated our proposed method’s effectiveness and reliability on the e_ophtha dataset. Compared with the IDRID dataset, the e_ophtha had only two DR lesion annotations, MA and EX, while it had lower image resolutions and a large number of low-quality images with issues of uneven lighting, underexposure, overexposure, etc. Performing segmentation experiments on the e_ophtha dataset was more challenging and placed higher demands on the model. As shown in Table 3, compared with other methods, our method achieved the best results for the segmentation of both lesions. We could observe that compared with the segmentation results on the IDRID dataset, the segmentation results of all the methods on the e_ophtha dataset showed different degradation degrees. Especially in the segmentation results of MA lesions, the segmentation performance of other methods decreased by 54% to 89%, respectively; however, our method’s segmentation performance decreased by 49%, showing the strongest robustness among all methods. Generally, our method performed poorly in segmentation for MA because, on one hand, the MA lesion
size was too small and had only a few pixels in size, which made feature extraction difficult. On the other, the retinal vessels were similar to the MA lesion features, causing a high occurrence of missegmentation phenomena. The PR curve obtained from the experiment is shown in Figure 8.

Table 3. Comparison experiments based on e_ophtha.

| Method      | MA    | HE    | SE    | EX    | mAUC_PR |
|-------------|-------|-------|-------|-------|---------|
| FRCN        | 0.0269| -     | -     | -     | 0.1675  | 0.0972  |
| UNet++      | 0.1636| -     | -     | -     | 0.7649  | 0.4643  |
| MANet       | 0.1098| -     | -     | -     | 0.7216  | 0.4157  |
| PSPNet      | 0.0237| -     | -     | -     | 0.6161  | 0.3199  |
| DeeplabV3+  | 0.1281| -     | -     | -     | 0.7105  | 0.4193  |
| Paper [27]  | 0.1367| -     | -     | -     | 0.6578  | 0.3973  |
| L-Seg       | 0.1687| -     | -     | -     | 0.4171  | 0.2929  |
| MSLF-Net    | 0.2179| -     | -     | -     | 0.8001  | 0.5090  |

Figure 8. PR curves of MSLF-Net on e_ophtha.

To sum up, the results presented show that Our 3M-Net outperforms the other five methods.

4.2. Ablation Studies on IDRID Dataset
4.2.1. Analysis of the Model Components

In this sub-section, we also performed ablation experiments to further understand the impact of each component on the model. First, we analyzed the effect of multi-scale information with the baseline model. Then, we discussed how the general fusion and cross-layer category fusion approaches helped the model. From Table 4, it is clear that these two modules effectively improved its segmentation performance. We introduced the MFEF module to provide the model with complete multi-scale information. The introduced MLFF module integrated the corresponding category-specific activations from different scales as the final fusion through implicit supervision of the final output. This fusion method effectively avoids the mutual interference of category information and feature contamination during feature fusion.

Table 4. Performance comparison of different model components.

| Method      | MA    | HE    | SE    | EX    | mAUC_PR |
|-------------|-------|-------|-------|-------|---------|
| Baseline    | 0.4059| 0.6348| 0.7017| 0.8460| 0.6471  |
| +MSFE       | 0.3959| 0.6446| 0.7160| 0.8464| 0.6507  |
| +MLFF       | 0.4393| 0.6411| 0.7597| 0.8644| 0.6761  |
4.2.2. Analysis of the Hybrid Loss Function

The loss function used in this paper consists of two parts. In order to verify the influence of loss functions with different weights on segmentation performance, we choose the appropriate $\lambda$ in Formula (6) through ablation experiments. Where, $\lambda = -1$ means only DICE loss function works, and $\lambda = 0$ means only cross-entropy loss function works. With the increase of $\lambda$, the weight of the Dice loss function becomes larger. Table 5 shows that the segmentation performance of the model is poor when using a separate loss function. When using the mixed loss function, the segmentation performance of the model is significantly improved. With the increase of $\lambda$, when $\lambda = 1.0$, the segmentation performance of the model was the best. When $\lambda$ continued to increase, the segmentation performance of the model started to decline. Therefore, the optimal $\lambda$ is 1.0.

Table 5. Performance comparison of weightage of hybrid loss function.

| $\lambda$ | MA   | HE   | SE   | EX   | mAUC_PR |
|----------|------|------|------|------|---------|
| $-1$     | 0.3826 | 0.6227 | 0.6635 | 0.8576 | 0.6316  |
| 0        | 0.4017 | 0.6174 | 0.7533 | 0.8457 | 0.6545  |
| 0.1      | 0.4234 | 0.6154 | 0.7490 | 0.8561 | 0.6604  |
| 0.2      | 0.4187 | 0.6162 | 0.7116 | 0.8091 | 0.6389  |
| 0.3      | 0.4325 | 0.6360 | 0.7300 | 0.8410 | 0.6599  |
| 0.4      | 0.4094 | 0.6179 | 0.7623 | 0.8469 | 0.6591  |
| 0.5      | 0.4198 | 0.6489 | 0.7324 | 0.8540 | 0.6638  |
| 0.6      | 0.4339 | 0.6290 | 0.7451 | 0.8517 | 0.6649  |
| 0.7      | 0.4227 | 0.6311 | 0.7508 | 0.8278 | 0.6581  |
| 0.8      | 0.3889 | 0.6431 | 0.7498 | 0.8419 | 0.6559  |
| 0.9      | 0.4459 | 0.6472 | 0.7394 | 0.8513 | 0.6709  |
| 1.0      | 0.4393 | 0.6411 | 0.7597 | 0.8644 | 0.6761  |
| 1.1      | 0.4469 | 0.6354 | 0.7211 | 0.8511 | 0.6636  |
| 1.2      | 0.4328 | 0.6390 | 0.7327 | 0.8479 | 0.6631  |

4.2.3. Analysis of the Image Augmentation

In view of the shortcomings of DR in segmenting datasets, We conducted image preprocessing and Augmentation for the IDRID and e_ophtha datasets. Table 6 shows the ablation experiments of FOV mask image and geometric transformations. The experiments show that both can significantly promote the segmentation performance of the model, especially with the increase of FOV mask image, the mAUPR index of the model increases by 9.7%.

Table 6. Performance comparison of weightage of hybrid loss function.

| Operation | MA   | HE   | SE   | EX   | mAUC_PR |
|-----------|------|------|------|------|---------|
| baseline  | 0.3326 | 0.5568 | 0.6236 | 0.7594 | 0.5681  |
| +Augmentation | 0.3742 | 0.5754 | 0.6756 | 0.8394 | 0.6162  |
| +FOV     | 0.4393 | 0.6411 | 0.7597 | 0.8644 | 0.6761  |

4.3. Discussion

There are several methods of examining fundus images, including ophthalmoscopy, optical coherence tomography (OCT), fluorescein fundus angiography (FFA), and fundus color photography [40]. The first three methods either require high-cost equipment or some clinically restricted agents, such as pupil diluents and contrast agents, making it difficult to popularize their use. In contrast, fundus images are not harmful and can be digitally stored, making them suitable for large-scale lesion screening. This study’s main contribution is the lesion segmentation of fundus images by the neural network to assist doctors in diagnosing DR conditions. Presently, there is still much room for improvement in DR segmentation techniques due to the complexity of fundus images and DR lesions [41]. These issues motivate us to further investigate this problem.
The purpose of DR segmentation is to accurately mark all lesion regions. The main problems faced in this process are the difficulties of tiny feature extraction and noise interference [42]. To address these issues, we proposed a multi-scale cross-layer feature fusion strategy and conducted comparative experiments with common feature enhancement methods, such as pyramid pooling, dilated convolution, feature superposition, attention mechanisms, and depth supervision. The experimental results demonstrated that our proposed method adequately performed the DR segmentation tasks, and our depth-supervised approach also had merits in tiny features extraction.

The main problem we faced on the e_ophtha dataset was noise interference. The large number of low-quality images made it difficult for our model to extract features. Thus, its segmentation performance for MA substantially deteriorated. As shown in Figure 9, in the segmentation result graph of MA, each method showed missed segmentation. Particularly speaking, UNet++ had the most serious problems, while DeepLabV3+ and our proposed model maintained correct segmentation for larger lesions. These results suggested that null convolution is a method that can be learned from or adopted in future studies.

For the IDRID dataset, the segmentation performance of HE, SE, and EX was satisfactory, but the segmentation performance for MA lesions with smaller sizes was not. This dataset’s image quality was high, such that feature extraction was relatively simple. The main problem we faced was with tiny feature extraction. Figure 10 shows the four results of lesion segmentation plotted for the IDRID dataset.
With a comparison between the segmentation result plot of each method in Figure 10 and GroundTruth, it is apparent that all methods did not have any missed segmentation of lesions. However, missegmentation occurred frequently; a large number of normal retinal tissues were missegmented as lesions. From this phenomenon, we conclude that distinguishing between normal retinal tissues and lesions is the key to improving segmentation performance in the presence of high-quality fundus images. The occurrence of DR is associated with structural changes in the retinal vasculature. For example, MA is an abnormal proliferation of retinal capillaries following swelling, and HE is a progressive lesion resulting from the rupture of a retinal vessel or microaneurysm. Therefore, the color, contour, and other appearance features of many early lesions are extremely similar to those of blood vessels. This fact is the main reason for missegmentation. However, we can obtain useful information from this pathological connection. We carefully analyzed the pathological relationship and combined it with the observation of the multi-lesion segmentation results in Figure 10. The distribution of DR lesions is related to the distribution of retinal vessels. The existing methods have ignored this pathological distribution relationship, resulting in a large number of missegmentations that could have been avoided. This topic is what we need to focus our attention on in future studies.

Overall, as shown in the segmentation results in Figures 9 and 10, our proposed method significantly improves the DR segmentation accuracy. This result is valuable both
for enhancing physicians’ confidence in automatic DR segmentation techniques and for screening the conditions of general diabetic patients. Our method accelerates the application of deep-learning algorithms to DR clinical diagnosis to a certain extent, partially overcoming the problems of physician shortages and high work intensity. These achievements are prerequisites for moving an approach from theory to clinical practice. However, neural networks are a data-driven science. Available DR segmentation datasets are not very abundant, in which there are insufficient data samples and solidified characteristics by the devices and patient distribution areas where the samples are collected. We are currently inconclusive about the existence of data bias in geographical distributions. However, the inadequate training of models, a limited generalization ability, and the lack of reproducibility arising from data bias and insufficient datasets are obvious issues. They are evidenced by the fact that the accuracy and generalization of the DR segmentation techniques are much lower than those of other medical image segmentation approaches. This phenomenon severely hinders clinical applications and is currently a pressing problem.

5. Conclusions

In this paper, we proposed a new automatic DR lesion segmentation model, MSLF-Net. This model could fully extract multi-scale information and perform effective class hierarchical feature fusion to improve the segmentation accuracy of DR lesions via feature enrichment and feature contamination avoidance. It also has an enhanced anti-noise capability. Moreover, we optimized the DR segmentation dataset to improve the model’s training efficiency and quality. To verify the validity and reliability of MSLF-Net, we conducted comparative experiments on the IDRID and e_ophtha datasets. Then, we conducted ablation experiments for each component of the model to understand its impact. The experimental results showed that our model has a satisfactory performance in DR segmentation compared with the existing DR segmentation methods. This study can facilitate large-scale screening and condition diagnosis of DR. In the future, we plan to focus on the segmentation of small lesions to further improve the performance of, and ultimately promote, the DR segmentation task to be widely used in clinical diagnosis.

Author Contributions: Writing—original draft preparation, H.Y.; review and editing, J.X.; data collection, D.Z. and L.J.; supervision and project administration, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used two public datasets to evaluate our proposed segmentation model. They are the IDRID dataset and e_ophtha dataset. The two datasets can be accessed from the following URLs: https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid and https://www.adcis.net/en/third-party/e-ophtha (accessed on 28 August 2022). The mask images that we added to the dataset can be accessed from the following URLs: https://drive.google.com/file/d/1Gt5MHvo_0qVboXV155c9t0ZdBwHwpT1c/view?usp=sharing (accessed on 21 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Danaei, G.; Finucane, M.M.; Lu, Y.; Singh, G.M.; Cowan, M.J.; Paciorek, C.J.; Lin, J.K.; Farzadfar, F.; Khang, Y.H.; Stevens, G.A.; et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet* 2011, 378, 31–40. [CrossRef] [PubMed]
2. Zhang, G.; Chen, H.; Chen, W.; Zhang, M. Prevalence and risk factors for diabetic retinopathy in China: A multi-hospital-based cross-sectional study. *Br. J. Ophthalmol.* 2017, 101, 1591–1595. [CrossRef] [PubMed]
3. Atlas, I.D. A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018. Thomas RL, Halim S, Gurudas S, Sivaprasad S, Owens DR. *Diabetes Res. Clin. Pract.* 2019, 157, 107840.
4. Wilkinson, C.P.; Ferris, F.L., III; Klein, R.E.; Lee, P.P.; Agardh, C.D.; Davis, M.; Dills, D.; Kumpik, A.; Pararajasegaram, R.; Verdaguer, J.T.; et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* **2003**, *110*, 1677–1682. [CrossRef]

5. Bellemo, V.; Lim, Z.W.; Lim, G.; Nguyen, Q.D.; Xie, Y.; Yip, M.Y.; Hamzah, H.; Ho, J.; Lee, X.Q.; Hsu, W.; et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit. Health* **2019**, *1*, e35–e44. [CrossRef]

6. Voets, M.; Møllersten, K.; Bongo, L.A. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *arXiv* **2018**, arXiv:1803.04337

7. Patton, N.; Aslam, T.M.; MacGillivray, T.; Deary, I.J.; Dhillon, B.; Eikelboom, R.H.; Yogesan, K.; Constable, I.J. Retinal image analysis: concepts, applications and potential. *Prog. Retin. Eye Res.* **2006**, *25*, 99–127. [CrossRef]

8. Shivaram, J.M.; Patil, R.; Aravind, H. Automated detection and quantification of haemorrhages in diabetic retinopathy images using image arithmetic and mathematical morphology methods. *Int. J. Recent Trends Eng.* **2009**, *2*, 174.

9. Weiwei, G.; Jiaxin, S.; Yuliang, W.; Chun, L.; Jing, Z. Application of local adaptive region growth method based on multi-template matching in automatic detection of retinal hemorrhage. *Spectrosc. Spectr. Anal.* **2013**, *33*, 448–453.

10. Zivkovic, Z. Improved adaptive Gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 23–26 August 2004.

11. Huang, Y.; Hu, J. Residual Neural Network Based Classification of Macular Edema in OCT. In *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence*, Portland, OR, USA, 4–6 November 2019.

12. Xue, J.; Yan, S.; Qu, J.; Qi, F.; Qu, C.; Zhang, H.; Chen, M.; Liu, T.; Li, D.; Liu, X. Deep membrane systems for multitask segmentation in diabetic retinopathy. *Knowl.-Based Syst.* **2019**, *183*, 104887. [CrossRef]

13. Long, J.; Shellhammer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 7–12 June 2015.

14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 5–9 October 2015.

15. Eftekhar, N.; Pourreza, H.R.; Masoudi, M.; Ghiasi-Shirazi, K.; Saeedi, E. Microaneurysm detection in fundus images using a two-step convolutional neural network. *Biomed. Eng. Online* **2019**, *18*, 67. [CrossRef][PubMed]

16. Mo, J.; Zhang, L.; Feng, Y. Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks. *Neurocomputing* **2018**, *290*, 161–171.

17. Li, T.; Gao, Y.; Wang, K.; Guo, S.; Liu, H.; Kang, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* **2019**, *501*, 511–522. [CrossRef]

18. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017.

19. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8–14 September 2018.

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

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

22. Kou, C.; Li, W.; Liang, W.; Yu, Z.; Hao, J. Microaneurysms segmentation with a U-Net based on recurrent residual convolutional neural network. *J. Med. Imaging* **2019**, *6*, 025008. [CrossRef]

23. Huang, S.; Li, J.; Yao, X.; Shen, N.; Xu, T. RTNet: Relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Trans. Med. Imaging* **2022**, *41*, 1596–1607. [CrossRef]

24. Zhang, L.; Feng, S.; Duan, G.; Li, Y.; Liu, G. Detection of microaneurysms in fundus images based on an attention mechanism. *Genes* **2019**, *10*, 817. [CrossRef]

25. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8–14 September 2018.

26. Song, T.; Zhang, X.; Ding, M.; Rodriguez-Paton, A.; Wang, S.; Wang, G. DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions. *Methods 2022*, *204*, 269–277. [CrossRef]

27. Ramamurthy, K.; George, T.T.; Shah, Y.; Sasidhar, P. A Novel Multi-Feature Fusion Method for Classification of Gastrointestinal Malignancies Using Endoscopy Images. *Diagnostics* **2022**, *12*, 2316. [CrossRef]

28. Porwal, P.; Pachade, S.; Kokare, M.; Deshmukh, G.; Mériaudeau, F. IDRID: Diabetic Retinopathy—Segmentation and Grading Challenge. *Med. Image Anal.* **2019**, *59*, 101561. [CrossRef]

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

30. Staal, J.; Abramoff, M.D.; Niemeijer, M.; Viergever, M.A.; Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **2004**, *23*, 501–509. [CrossRef][PubMed]

31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

32. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In *Proceedings of the Artificial Intelligence and Statistics*, San Diego, CA, USA, 9–12 May 2015.
33. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

35. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.

36. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2017.

37. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–21 June 2009.

38. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 1856–1867. [CrossRef]

39. Fan, T.; Wang, G.; Li, Y.; Wang, H. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access* **2020**, *8*, 179656–179665. [CrossRef]

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

41. Furtado, P.; Baptista, C.; Paiva, I. Segmentation of Diabetic Retinopathy Lesions by Deep Learning: Achievements and Limitations. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, Valletta, Malta, 24–26 February 2020.

42. Antal, B.; Hajdu, A. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowl.-Based Syst.* **2014**, *60*, 20–27. [CrossRef]