WSRFNet: Wavelet-Based Scale-Specific Recurrent Feedback Network for Diabetic Retinopathy Lesion Segmentation

Xuan Li and **Xiangqian Wu***
Faculty of Computing, Harbin Institute of Technology, Harbin, China
xuanli@stu.hit.edu.cn, xqwu@hit.edu.cn

Abstract

Diabetic retinopathy lesion segmentation (DRLS) faces a challenge of significant variation in the size of different lesions. An effective method to address this challenge is to fuse multi-scale features. To boost the performance of this kind of method, most existing DRLS methods work on devising sophisticated multi-scale feature fusion modules. Differently, we focus on improving the quality of the multi-scale features to enhance the fused multi-scale feature representation. To this end, we design a Wavelet-based Scale-specific Recurrent Feedback Network (WSRFNet), which refines multi-scale features using recurrent feedback mechanism. Specifically, to avoid information loss when introducing feedback to multi-scale features, we propose a wavelet-based feedback pyramid module (WFPM), which is based on a reversible downsampling operation, i.e., Haar wavelet transform. Unlike scale-agnostic feedback used in previous feedback methods, we develop a scale-specific refinement module (SRM), which utilizes scale-specific feedback to pointedly refine features of different scales. Experimental results on IDRiD and DDR datasets show that our approach outperforms state-of-the-art models. The code is available at https://github.com/xuanli01/WSRFNet.

1 Introduction

Diabetic retinopathy (DR), which is the main ocular complication of diabetes mellitus, has been the major cause of visual impairment and blindness, especially in the working-age individuals today [Ting *et al.*, 2016; Thomas *et al.*, 2019]. It is estimated that about 103 million adults suffered from DR all over the world in 2020. The numbers are predicted to increase to 161 million in 2045 [Teo *et al.*, 2021]. The common manifestations of DR lesions include hard exudates (EX), hemorrhages (HE), soft exudates (SE) and microaneurysms (MA). Ophthalmologists normally diagnose DR based on the presence of these lesions in colorful fundus images. Due to the small number of ophthalmologists, one ophthalmologist needs to screen a great deal of fundus images, which is a heavy workload. Therefore, automatic DR lesion segmentation (DRLS) is of vital importance to assist the diagnosis of DR and reduce the workload of ophthalmologists.

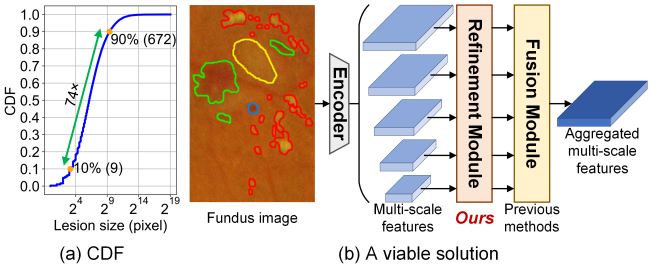

Figure 1: Challenge of DRLS and the solution. (a) shows CDF of the lesion size in DDR training set. In the fundus image in (b), the regions marked with red, green, yellow and blue edges are EX, HE, SE and MA, respectively. (b) illustrates a viable solution to the large size variation challenge.

A big challenge in DRLS is that the diverse lesions are largely different in size. For clear explanation of this challenge, we count the lesion size in DDR training set [Li *et al.*, 2019a] and draw its cumulative distribution function (CDF). As shown in Fig. 1(a), the largest 10% lesions are at least 74 times larger than the smallest 10% lesions. Visual lesion annotation in the fundus image (see Fig. 1(b)) also illustrates the large variation in lesion size. An effective method to tackle this challenge is to aggregate multi-scale features extracted from a feature encoder, which is shown in Fig. 1(b). The performance of the method depends on two factors, i.e., the effectiveness of multi-scale feature fusion module and the quality of the input features of different scales. Most existing methods [Guo *et al.*, 2019; Bo *et al.*, 2022] focus on the former factor to develop sophisticated multi-scale feature fusion modules and alleviate the challenge to some extent. However, the latter factor has not yet been investigated in DRLS. Observing this, our work will explore the latter factor in DRLS for the first time by designing a refinement module to rectify the multi-scale features to improve the quality of them for better aggregated multi-scale feature representation (see Fig. 1(b)).

The refinement of the multi-scale features has recently been investigated in other dense prediction tasks such as

*Corresponding author.

salient object detection (SOD) and camouflaged object detection (COD) tasks. One kind of method [Wei *et al.*, 2020a; Hu *et al.*, 2023], which has been proved to be effective, is to leverage the aggregated multi-scale features from previous iteration as feedback to refine the features of different scales via designed iterative feedback methods. We deeply analyze the works of such kind of method and find two main issues: 1) *Spatial information loss in refinement*. Current methods use the irreversible downsampling operations such as strided convolution and bilinear interpolation to downsample the feedback features [Wei *et al.*, 2020a] or the refined multi-scale features [Hu *et al.*, 2023], which may lead to the loss of spatial information that is important to pixel-wise predictions. 2) *Scale-agnostic feedback for features of different scales*. Large-scale features from low-level layers of the encoder contain rich details but inaccurate semantic information. Small-scale features from high-level layers contain precise semantics but lack enough details. Therefore, large-scale features may need more semantic refinement, while small-scale features may require more detail refinement. However, previous approaches apply the same feedback (i.e., scale-agnostic feedback) for features of different scales without considering their characteristics, which is difficult to meet their different requirements for the refinement.

To address above-mentioned issues, in this paper, we propose a Wavelet-based Scale-specific Recurrent Feedback Network, named WSRFNet, which recursively utilizes feedback to refine the multi-scale features to acquire better aggregated multi-scale feature representation for accurate DRLS. Specifically, for features of each scale, we generate feedback from the fused multi-scale features in previous loop to refine them. To reduce computation, we employ downsampling operation to obtain feedback of the same scale as the features needed to be corrected instead of upsampling the multi-scale features to the scale of high-resolution feedback [Hu *et al.*, 2023]. To avoid the loss of feedback information when downsampling, we devise a wavelet-based feedback pyramid module (WFPM), which produces initial feedback features for features of each scale based on Haar wavelet transform (HWT) that is a lossless information transformation method [Stankovic and Falkowski, 2003]. Considering diverse characteristics of different scale features, we design a scale-specific refinement module (SRM), which conducts refinement with scale-specific feedback obtained by merging initial feedback features with features requiring correction.

In summary, this work makes the following contributions:

- For the DRLS task, we design a Wavelet-based Scale-specific Recurrent Feedback Network (WSRFNet), which aims to obtain exquisite aggregated multi-scale feature representation by refining the multi-scale features in a recurrent feedback manner.
- We devise a wavelet-based feedback pyramid module (WFPM) to keep as much information as possible via Haar wavelet transform, when the feedback information flows into each level of the network.
- We propose a scale-specific refinement module (SRM), which generates scale-specific feedback to pointedly refine the multi-scale features.
- Comprehensive experiments on IDRiD [Porwal *et al.*, 2020] and DDR [Li *et al.*, 2019a] datasets demonstrate the superiority of our method over the state-of-the-art (SOTA) approaches.

2 Related Work

2.1 Diabetic Retinopathy Lesion Segmentation

Existing DRLS works can be categorized into two main classes: one-type and multi-type lesion segmentation. For the former one, one-type lesion segmentation result is predicted with one forward propagation [Dai *et al.*, 2018; Guo *et al.*, 2020; Liu *et al.*, 2021a; Zhang *et al.*, 2022]. For instance, SS-MAF [Zhang *et al.*, 2022] designs an auxiliary super-resolution task, which brings in helpful detailed features for tiny EX lesion and boundaries detection.

For the latter one, a unified model is applied to segment multi-type of lesions simultaneously. For example, L-seg [Guo *et al.*, 2019] designs a multi-scale feature weighted fusion method to handle the issue that small lesion regions could not response at high level of network. RTNet [Huang *et al.*, 2022] proposes to capture the intra-class dependencies among multi-lesion and inter-class relations between multi-lesion and vessels. SAA [Bo *et al.*, 2022] devises a scale-aware attention block to re-weight importance of multi-scale features dynamically to alleviate the inconsistent scale problem. M2MRF [Liu *et al.*, 2023] proposes a novel feature reassembly operator to preserve subtle activations about small lesions as much as possible. [Guo *et al.*, 2024] utilize a set of parallel fully convolutional neural networks (FCN) with different input scales that can better extract the features of lesions with different sizes and introduce the weight CNN that can better fuse the advantages of these FCN. Different from previous methods, we focus on the refinement of multi-scale features to address the large scale variation problem for better performance of multi-type lesion segmentation.

2.2 Feedback Mechanisms

Feedback mechanisms have been successfully applied in various computer vision tasks such as super-resolution [Han *et al.*, 2018; Li *et al.*, 2019b], SOD [Wei *et al.*, 2020a] and COD [Hu *et al.*, 2023]. Most of the feedback mechanisms are implemented using iterative strategy, allowing the network to utilize output information to rectify previous states. For example, [Li *et al.*, 2019b] design an image super-resolution feedback network, which refines low-level representation with high-level information under a recurrent structure. F^3Net [Wei *et al.*, 2020a] devises a cascaded feedback decoder for SOD, which can feedback features of both high semantics and high resolutions to previous ones to refine them. HitNet [Hu *et al.*, 2023] embeds a global feedback connection into the multi-scale framework via a feedback fusion block, which integrates the information from multi-scale outputs. However, previous feedback mechanisms for dense prediction tasks adopt irreversible downsampling operations and produce the same feedback for multi-scale features, which may limit the performance of the models as mentioned above (Section 1). To maintain as much information for correction

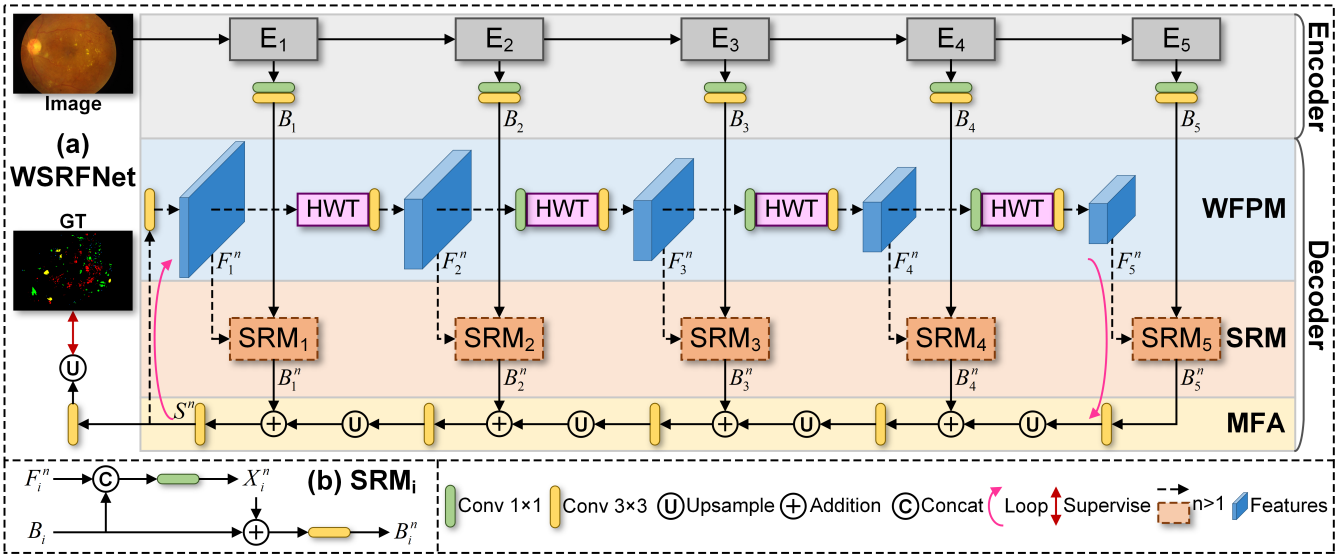

Figure 2: Overall architecture of our proposed method (better viewed in color).

as possible when downsampling, we apply HWT in our feedback generation module. To tailor the feedback to the features of each scale, we integrate the feedback with features requiring refinement to obtain scale-specific feedback.

3 Method

Fig. 2 illustrates the overall architecture of our proposed WSRFNet, which follows an encoder-decoder structure. The encoder is used to extract initial multi-scale features $\mathcal{B} = \{B_i\}_{i=1}^{5}$ from the input images. Specifically, the encoder utilizes the backbone network $\{E_i\}_{i=1}^{5}$ to extract the multi-level features from the input images and passes the extracted features through two convolutional layers to adjust the number of their channels to $C = 64$ for subsequent processing [Wei *et al.*, 2020b]. For ease of statement, $i \in \{1, 2, 3, 4, 5\}$ indicates the i-th level from bottom to top. The output features B_i from different levels of the encoder correspond to different scales, i.e., $\frac{1}{2^i}$ resolution $H \times W$ of the input images.

The decoder, which consists of three components: WFPM, SRM and multi-scale feature aggregator (MFA), is designed to gain a high-quality fused multi-scale feature representation by refining features of different scales in a recurrent way. The details of recurrent refinement, WFPM, SRM, MFA and loss function are introduced in Section 3.1, Section 3.2, Section 3.3, Section 3.4 and Section 3.5, respectively.

3.1 Recurrent Refinement of WSRFNet

Most existing remedies to large size variation problem in DRLS is to develop delicate multi-scale feature fusion modules to strengthen integrated multi-scale feature representation. Differently, we achieve this goal from a novel perspective, i.e., the refinement of multi-scale features. We perform refinement using the recursive feedback mechanism, which is inspired by the effectiveness of the iterative feedback method used for the correction of multi-scale features in other dense prediction tasks [Wei *et al.*, 2020a; Hu *et al.*, 2023].

To be specific, the loop number n=1 assumes the first loop and no feedback features from previous loop. In the first loop, the initial multi-scale features are fed into MFA to fuse the multi-scale features. For loop number (n>1), WFPM starts by producing initial feedback pyramid $\mathcal{F}^n = \{F_i^n\}_{i=1}^{5}$ from the fused multi-scale features S^{n-1} of previous loop for initial multi-scale features. Then, SRM utilizes $\mathcal{F}^n$ to generate scale-specific feedback to refine the initial multi-scale features. After that, the refined multi-scale features $\mathcal{B}^n = \{B_i^n\}_{i=1}^{5}$ are passed through the MFA same as the first loop. In each loop, the decoder shares weights. The above recurrent refinement can be formulated as

$$\begin{cases} S^1 = \text{MFA}(\mathcal{B}), & n = 1 \\ \mathcal{F}^n = \text{WFPM}(S^{n-1}) & \\ \mathcal{B}^n = \text{SRM}(\mathcal{F}^n, \mathcal{B}) & \\ S^n = \text{MFA}(\mathcal{B}^n), & n > 1 \end{cases} \tag{1}$$

3.2 Wavelet-Based Feedback Pyramid Module

Previous feedback methods employ irreversible downsampling operations such as bilinear interpolation to obtain feedback for initial multi-scale features, which consequentially leads to information loss. The lost information may contain spatial details of the objects which are important to high performance DRLS. To avoid information dropping, we design a WFPM to produce initial feedback pyramid from aggregated multi-scale features of previous loop with the assist of HWT (see Fig. 2(a)). HWT is a reversible downsampling operation because of its biorthogonal property. Therefore, incorporated with HWT, WFPM enables less information loss when generating initial multi-scale feedback.

Concretely, given that the resolution of the lowest-level initial features B_1 is the same as S^{n-1}, a 3×3 convolutional layer is used to enhance S^{n-1} to produce the initial feedback features $F_1^n \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ of the lowest level (i=1). For level number (i>1), initial low-level feedback features F_{i-1}^n are

passed through a 1×1 convolutional layer $Conv_{1\times1}$ to adjust their channel number to C for reduction of computation, if their channel number is $4C$.

Then, HWT is applied to downsample the channel-compressed feedback features, nondestructively. Formally, take 2D feature map $Z \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times C}$ (i>1) as the input of HWT for the illustration of the implementation of HWT. HWT employs four filters W_{LL}, W_{LH}, W_{HL}, W_{HH} to decompose Z into four Haar wavelet subbands, which are corresponding to $Z_{LL} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, $Z_{LH} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, $Z_{HL} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, $Z_{HH} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, respectively. Z_{LL} contains all low-frequency information of Z. Z_{LH}, Z_{HL} and Z_{HH} include all high-frequency information of Z. Next, to retain all the information of the input feature map, we utilize a cross-channel concatenation operation to merge all the wavelet subbands to obtain the output of the HWT, which is $[Z_{LL}, Z_{LH}, Z_{HL}, Z_{HH}] \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 4C}$.

At last, a 3×3 convolutional layer $Conv_{3\times3}$ is used to enlarge receptive field to adapt the increasement of the level to get the initial feedback features $F_i^n \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 4C}$ for the i-th level. The WFPM can be formulated as

$$\begin{cases} F_1^n = Conv_{3\times3}(S^{n-1}), & i=1 \\ F_2^n = Conv_{3\times3}(\text{HWT}(F_1^n)), & i=2 \\ F_i^n = Conv_{3\times3}(\text{HWT}(Conv_{1\times1}(F_{i-1}^n))), & i>2 \end{cases} \quad (2)$$

3.3 Scale-Specific Refinement Module

Large-scale initial features lack accurate semantics, which is caused by the small receptive field of the low-level layer. Small-scale initial features are coarse in details, due to multiple downsampling in the backbone network. The same feedback used by current works [Wei *et al.*, 2020a; Hu *et al.*, 2023] is hard to satisfy diverse needs of different scale features for the correction. Considering the unique characteristics of initial features of each scale, SRM aims to pointedly perform refinement via scale-specific feedback features. Fig. 2(b) shows the structure of our SRM.

To be specific, SRM contains five blocks, which are denoted by $\{\text{SRM}_i\}_{i=1}^5$. Each block is used for the refinement of features of each scale. For the implementation of SRM_i, firstly, we combine initial feedback features F_i^n with i-th level initial features B_i via a cross-channel concatenation operation $Concat$. Then, the combined features are fed into a 1×1 convolutional layer to acquire scale-specific feedback features $X_i^n \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, which is calculated by

$$X_i^n = Conv_{1\times1}(Concat(B_i, F_i^n)) \quad (3)$$

The 1×1 convolutional layer has three main purposes: 1) to compress the channel number of the combined features from $5C$ to C to align with the i-th level initial features B_i for subsequent refinement, 2) to adequately merge cross-channel information of the combined features so that characteristics of the i-th level initial features can be injected into the initial feedback to generate more targeted feedback for i-th level initial features, and 3) to filter redundant information to obtain more representative feedback.

After that, we rectify the initial features with scale-specific feedback features via an element-wise addition operation ($\oplus$) followed by a 3×3 convolutional layer for enhancement to get the refined features B_i^n, which is denoted by

$$B_i^n = Conv_{3\times3}(B_i \oplus X_i^n) \quad (4)$$

3.4 Multi-Scale Feature Aggregator

MFA aims to obtain multi-scale information by fusing the features of different scales via an FPN-like structure [Lin *et al.*, 2017], which is shown in Fig. 2(a). In practice, MFA firstly utilizes a bilinear interpolation operation to upsample the high-level features D_i^n by a factor of 2. Then, the output features are integrated with the low-level features B_{i-1}^n via an element-wise addition operation. After that, a 3×3 convolutional layer is applied to enhance the merged features. Taking $\mathcal{B}^n$ as input, MFA can be formulated as

$$D_{i-1}^n = Conv_{3\times3}(\text{Upsample}(D_i^n) \oplus B_{i-1}^n),\ (i>1) \quad (5)$$

where $D_5^n = Conv_{3\times3}(B_5^n)$, D_1^n is equal to S^n.

3.5 Loss Function

To guarantee better fusion of the multi-scale features and get the satisfactory feedback, following [Hu *et al.*, 2023], we impose constraint from the ground truth G on the segmentation maps predicted by the aggregated multi-scale features in each loop. In practice, $\{S^n\}_{n=1}^N$ are fed into a 3×3 convolutional layer and an upsampling layer to produce the segmentation map $\{P^n\}_{n=1}^N$, where N is the total loop number. The final loss function $\mathcal{L}$ is calculated by

$$\mathcal{L} = \sum_{n=1}^{N} \frac{n}{N} \mathcal{L}_d(P^n, G) \quad (6)$$

where $\mathcal{L}_d$ is Dice loss [Milletari *et al.*, 2016] that is commonly used in DRLS, $\frac{n}{N}$ represents the weight of each loop constraint, which is set similar to [Hu *et al.*, 2023].

As the fused multi-scale features are supervised in each loop, they may contain critical information for better prediction. Noting this, we combine the aggregated multi-scale features of all loops for prediction via a weighted average operation in testing, which is expressed as

$$S_f = \frac{1}{\sum_{n=1}^{N} n} \sum_{n=1}^{N} nS^n \quad (7)$$

where n is the weight of the fused multi-scale features of each loop, which is set according to the weight of each loop constraint in training (Eq. 6), S_f are the final features for prediction. In testing, we pass S_f through the trained 3×3 convolutional layer and an upsampling layer to get the ultimate segmentation map P_f.

4 Experiments

4.1 Datasets and Evaluation Metrics

The performance of our method is evaluated on two popular benchmark datasets, i.e., IDRiD [Porwal *et al.*, 2020] and DDR [Li *et al.*, 2019a]. Both datasets contain four types of annotation, including EX, HE, SE and MA.

Method	AUPR					Dice					IoU				
	mAUPR	EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU	EX	HE	SE	MA
L-seg	65.15	79.45	63.74	71.13	46.27	-	-	-	-	-	-	-	-	-	-
RTNet	70.76	**90.24**	68.80	75.02	*48.97*	-	-	-	-	-	-	-	-	-	-
SAA	67.38	88.12	67.04	72.81	41.52	-	-	-	-	-	-	-	-	-	-
Guo et al.	*70.96*	*89.61*	68.84	*75.63*	**49.75**	*67.02*	79.47	63.59	*72.33*	**52.68**	-	-	-	-	-
M2MRF-C	67.24	82.16	68.69	69.32	48.80	65.71	*79.85*	66.42	67.92	*48.63*	*49.94*	*66.46*	49.72	51.43	**32.13**
Deeplabv3+	66.63	84.95	68.29	65.64	47.65	63.07	78.21	64.29	62.53	47.23	47.00	64.22	47.38	45.49	30.92
HRNet48	68.90	85.65	*72.09*	72.31	45.56	65.17	79.28	*66.49*	68.42	46.48	49.44	65.67	*49.80*	*52.00*	30.28
Swin-base	69.03	85.29	**72.60**	72.25	45.97	65.61	79.00	**67.92**	68.42	47.11	49.88	65.28	**51.42**	*52.00*	30.82
Ours	**71.05**	87.15	69.65	**79.43**	47.98	**67.53**	**81.03**	65.09	**75.40**	48.62	**52.25**	**68.12**	48.24	**60.51**	*32.11*

Table 1: Quantitative comparison on IDRiD dataset. The bold and italic fonts indicate the best and second-best performance, respectively.

Method	AUPR					Dice					IoU				
	mAUPR	EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU	EX	HE	SE	MA
L-seg	32.08	55.46	35.86	26.48	10.52	-	-	-	-	-	-	-	-	-	-
RTNet	33.62	56.71	36.56	29.43	11.76	-	-	-	-	-	-	-	-	-	-
SAA	41.02	62.69	44.56	37.47	19.33	-	-	-	-	-	-	-	-	-	-
Guo et al.	35.30	60.14	36.52	32.60	11.94	34.80	54.06	33.17	38.76	15.22	-	-	-	-	-
M2MRF-C	*48.94*	**63.59**	*54.43*	49.35	*28.38*	45.40	**60.62**	45.16	47.78	**28.04**	30.09	**43.49**	29.17	31.39	**16.31**
Deeplabv3+	45.20	61.46	52.13	38.89	28.33	40.66	57.26	39.82	39.33	26.25	26.14	40.12	24.86	24.48	15.11
HRNet48	48.27	61.89	53.81	48.93	**28.45**	45.21	*59.07*	*48.29*	47.46	26.03	29.95	*41.92*	*31.83*	31.11	14.96
Swin-base	47.98	*63.10*	51.35	*52.99*	24.46	*45.48*	58.33	47.08	*52.74*	23.76	*30.31*	41.17	30.79	*35.81*	13.48
Ours	**50.07**	62.54	**57.54**	**53.79**	26.42	**47.66**	58.59	**51.81**	**53.56**	*26.68*	**32.09**	41.43	**34.96**	**36.57**	*15.39*

Table 2: Quantitative comparison on DDR test set. The bold and italic fonts indicate the best and second-best performance, respectively.

IDRiD Dataset

This dataset is available for DRLS in ISBI-2018 grand challenge [Porwal *et al.*, 2018; Porwal *et al.*, 2020], which has 81 colour fundus images with 54/27 images for training and testing. Each image contains 4288×2848 pixels.

DDR Dataset

This dataset is provided by Ocular Disease Intelligent Recognition (ODIR-2019) [Li *et al.*, 2019a], which contains 757 colour fundus images with 383/149/225 images for training, validation and testing. The sizes of images range from 1088×1920 to 3456×5184 pixels.

Evaluation Metrics

To quantitatively measure the proposed model, we follow the suggestion proposed by [Liu *et al.*, 2023], and adopt standard metrics including class-wise Area Under Precision-Recall curve (AUPR), mean class-wise AUPR (mAUPR), class-wise Dice coefficient (Dice), mean class-wise Dice coefficient (mDice), class-wise intersection-over-union (IoU) and mean class-wise IoU (mIoU).

4.2 Implementation Details

EfficientNet-B1 [Tan and Le, 2019], pre-trained on ImageNet [Russakovsky *et al.*, 2015], is used as the backbone network. The number of feature channels for the five levels of EfficientNet-B1 are 16, 24, 40, 112 and 1280. Rotation (90°, 180° and 270°), flipping (horizontal and vertical) and random cropping are applied to augment the images. Following [Liu *et al.*, 2023; Guo *et al.*, 2019; Bo *et al.*, 2022], each image is resized to 1440×960 pixels in IDRiD and 1024×1024 pixels in DDR during training. Following [Liu *et al.*, 2023], we adopt poly strategy with the initial learning rate of 0.01 and power of 0.9. Our network is trained end-to-end using stochastic gradient descent (SGD) optimizer. Momentum and weight decay are set to 0.9 and 0.0005, respectively. We train the model for 80k iterations on IDRiD and 100k iterations on DDR with a mini-batch size of 4. All the experiments of our method are performed using one NVIDIA GeForce RTX 3090 GPU with 24GB of memory.

4.3 Comparison with State-of-the-art Methods

We quantitatively compare the proposed WSRFNet with eight SOTA methods, including L-seg [Guo *et al.*, 2019], RTNet [Huang *et al.*, 2022], SAA [Bo *et al.*, 2022], Guo et al. [Guo *et al.*, 2024], M2MRF-C [Liu *et al.*, 2023], Deeplabv3+ [Chen *et al.*, 2018], HRNet48 [Wang *et al.*, 2020] and Swin-base [Liu *et al.*, 2021b]. The last three approaches, which are for general semantic segmentation (GSS), are implemented using the same training setting as ours for a fair comparison. HRNet48 utilizes FCN [Long *et al.*, 2015] as the decoder. Deeplabv3+ takes ResNet-50 [He *et al.*, 2016] as the backbone network. Swin-base uses UPerNet [Xiao *et al.*, 2018] as the decoder. The results of rest methods, which are specially designed for DRLS, are directly taken from original papers. A few details of some DRLS methods should be noted that: 1) RTNet utilizes two auxiliary datasets, which are DRIVE [Staal *et al.*, 2004] and STARE [Hoover *et al.*, 2000] with pixel-level vessel labels. 2) The backbone network of SAA is pretrained on the DDR training set.

Results on IDRiD Dataset

We show the quantitative comparison on IDRiD dataset in Table 1. Compared with the both GSS and DRLS methods,

	mAUPR	mDice	mIoU
N=1	58.10	56.59	40.37
N=2	60.34	58.84	42.75
N=3	61.95	60.24	44.16
N=4	62.46	60.91	44.91
N=5	60.93	59.47	43.34

Table 3: Ablation studies of the total loop number N on DDR validation set using three evaluation metrics.

	mAUPR	mDice	mIoU
P^1	45.51	44.30	29.27
P^2	49.63	46.79	31.30
P^3	49.93	47.53	31.98
P^4	49.74	47.59	32.04
P_f	50.07	47.66	32.09

Table 4: Quantitative results of exploration of the recurrent refinement and prediction method.

our WSRFNet achieves the best mAUPR, mDice and mIoU values. In the individual evaluation, we gain the best result on SE segmentation under all metrics.

Results on DDR Test Set

The comparative results of each method on DDR test set are listed in Table 2. This dataset is more challenging, for the reason that its test set contains more images. Although some previous methods can perform well on IDRiD dataset, most of them achieve unsatisfactory results on this dataset. It can be seen from the quantitative results that our WSRFNet also achieves the best performance among all other methods under mAUPR, mDice and mIoU metrics. Specifically, compared with the second-best method (i.e., [Guo *et al.*, 2024]) on IDRiD dataset, we outperform their model on DDR test set by a large margin, i.e., 14.77% improvement on mAUPR and 12.86% improvement on mDice. It is worth noting that we obtain the best performance on HE and SE segmentation under all metrics in the individual evaluation.

4.4 Ablation Study

Before exploring the influence of each module, a hyper-parameter (i.e., N) needs to be determined. N represents the total loop number of the decoder. Considering that the IDRiD dataset does not contain a validation set, following [Liu *et al.*, 2023], hyperparameter tuning is performed on the validation set of DDR. We gradually increase N from 1 to 5 and measure the performance of corresponding models in Table 3. With the increase of N, the performance of our model gradually rises first, and then falls at N=5. One possible reason is that our model is locally optimal at N=4. As N increases, we may find a value of N for the globally optimal model. However, the computational cost will rise alongside the increase in N. To balance the computational cost and performance, we choose N=4 as default in our WSRFNet. Our baseline model (1st row) is the model with N set to 1. All the following investigations are performed on the DDR test set using three evaluation metrics, i.e., mAUPR, mDice and mIoU.

Method	mAUPR	mDice	mIoU
Bilinear Interpolation	48.32	45.90	30.50
Stride Convolution	47.97	45.38	30.03
Ours	50.07	47.66	32.09

Table 5: Ablation studies of the WFPM.

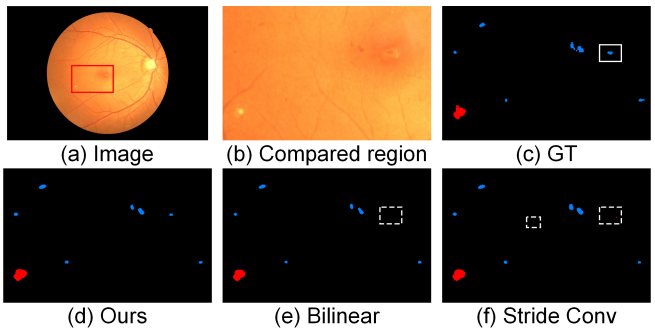

Figure 3: Qualitative results of exploration of the WFPM on DDR test set. The compared region (b) is the area marked with red box in the image (a). Regions filled in blue and red are MA and EX produced by ophthalmologists or segmentation methods. The challenging lesion is marked with white solid boxes in the ground truth (GT) map (c). Wrong and miss identifications are marked with white dotted boxes in the visual results (d-f).

Exploration of the Recurrent Refinement

We probe into the influence of the recurrent refinement on the aggregated multi-scale feature representation in this section. Concretely, we conduct the exploration by comparing the performance of prediction P^n produced by the fused multi-scale features of each loop. The results are listed in the first 4 rows of Table 4. The performance under mDice and mIoU metrics gradually rises with the increase of the loop number n, which illustrates that aggregated representation related to precise region coverage iteratively corrects itself under the recurrent refinement. The best mIoU and mDice values are in the 4th loop while the best mAUPR value is in the 3rd loop, which demonstrates that the fused multi-scale features may contain diverse crucial information for high performance DRLS in the different loops. Therefore, in testing, the weighted combination of the fused multi-scale features in all loops (Eq. 7) can enhance the corrected feature representation in the last loop, which is quantitatively proved by comparing the performance of P^4 (4th row) and P_f (5th row).

Exploration of the WFPM

We study the impact of the WFPM by directly replacing the WFPM in our full model with other downsampling operations used in other feedback networks, i.e., bilinear interpolation used by [Wei *et al.*, 2020a] and stride convolution used by [Hu *et al.*, 2023]. Specifically, the kernel size and stride of stride convolution are set to 3×3 and 2, respectively. It can be seen from the Table 5 that our WFPM achieves the best performance, which indicates the effectiveness of the WFPM.

Method	mAUPR	mDice	mIoU
Addition	48.60	45.13	29.89
Concatenation	48.23	45.52	30.25
Ours	50.07	47.66	32.09

Table 6: Ablation studies of the SRM.

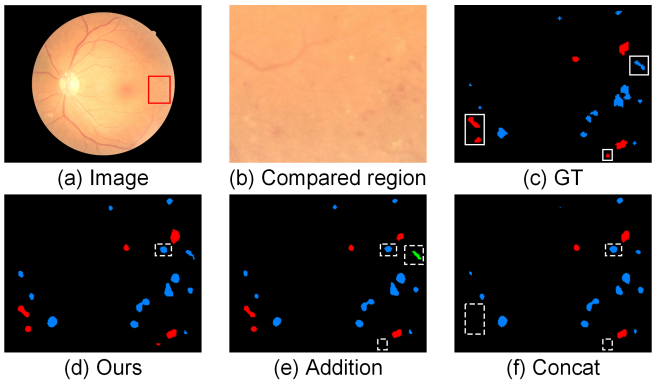

Figure 4: Visual results of exploration of the SRM on DDR test set. The compared region (b) is obtained by rotating the region marked with red box in the image (a) for better viewing. Regions filled in red, blue and green are EX, MA and HE by ophthalmologists or segmentation models. The challenging lesion is marked with white solid boxes in the GT map (c). Miss and wrong identifications are marked with white dotted boxes in qualitative results (d-f).

Fig. 3 shows visual comparison of other methods and our WFPM. It can be observed that our approach can produce better results for small lesions, which is benefited from the ability of WFPM to retain as much information as possible.

Exploration of the SRM

We investigate the impact of the SRM by replacing the SRM in our full model with other refinement modules employed in other feedback methods, i.e., element-wise addition used by [Wei *et al.*, 2020a] and cross-channel concatenation used by [Hu *et al.*, 2023]. In practice, the refinement module based on element-wise addition is implemented by replacing X_i^n in Eq. 4 with F_i^n. We conduct the cross-channel concatenation approach to refine the initial multi-scale features by removing B_i in Eq. 4. It can be observed from the Table 6 that our approach outperforms other methods, which demonstrates the effectiveness of the proposed SRM.

Visual results of other methods and ours are displayed in Fig. 4. We observe that our model can generate more accurate segmentation results for lesions of different scales. We provide an unsuccessful case marked with dotted boxes, which is shown in Fig. 4(d). This region is also incorrectly identified by other approaches. A possible reason is that the region is too similar to MA in appearance to properly segment it.

Extension to Other Backbones

To further demonstrate the effectiveness of the proposed refinement method, we combine it with other widely used backbones including VGG-16 [Simonyan and Zisserman, 2014], ResNet-50 [He *et al.*, 2016] and EfficientNet-B0 [Tan and Le, 2019]. We take backbones with MFA as the baseline models. As shown in Table 7, the baseline models are significantly improved by the proposed method with acceptable increasement of number of parameters, which indicates that the proposed methods are effective to other popular backbones. It is worth noting that compared with other models using the same backbone network including L-seg (VGG-16), Deeplabv3+ (ResNet-50), and SAA (EfficientNet-B0) in Table 2, our approaches outperform them by large margins.

Method	mAUPR	mDice	mIoU	Param
VGG-16	44.48	41.87	27.13	15.19M
+ Our method	49.46	47.39	31.92	17.92M
ResNet-50	42.87	40.15	25.94	24.13M
+ Our method	48.43	46.40	30.94	26.86M
EfficientNet-B0	44.16	40.68	26.17	4.47M
+ Our method	49.23	46.74	31.27	7.20M
EfficientNet-B1	45.45	42.13	27.34	6.98M
+ Our method	50.07	47.66	32.09	9.71M

Table 7: Quantitative results based on different backbones.

5 Conclusion

In this work, our goal is to deal with the challenge of large size variation in DRLS via improving the quality of multi-scale features to augment the aggregated multi-scale feature representation. To achieve this, we propose a WSRFNet, which adopts a recurrent feedback manner to rectify multi-scale features. In particular, to alleviate the issue of information loss in previous feedback models, we develop a WFPM to produce initial feedback for features of different scales via a lossless dowmsampling operation, i.e., HWT. To cater to the diverse requirements of different scale features for correction as much as possible, we devise the SRM. This module conducts refinement using scale-specific feedback, which is obtained by combing initial feedback with features requiring correction. Comprehensive experiments on two public benchmark datasets demonstrate the superiority of our WSRFNet over the state-of-the-art approaches.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62073105, U23A20389 and 62306095, by the Natural Science Foundation of Heilongjiang Province of China under Grant ZD2022F002, and by the Heilongjiang Touyan Innovation Team Program.

References

[Bo *et al.*, 2022] Wang Bo, Tao Li, Xinhui Liu, and Kai Wang. Saa: scale-aware attention block for multi-lesion segmentation of fundus images. In *International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam.

Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[Dai *et al.*, 2018] Ling Dai, Ruogu Fang, Huating Li, Xuhong Hou, Bin Sheng, Qiang Wu, and Weiping Jia. Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. *IEEE Transactions on Medical Imaging*, 37(5):1149–1161, 2018.

[Guo *et al.*, 2019] Song Guo, Tao Li, Hong Kang, Ning Li, Yujun Zhang, and Kai Wang. L-seg: An end-to-end unified framework for multi-lesion segmentation of fundus images. *Neurocomputing*, 349:52–63, 2019.

[Guo *et al.*, 2020] Song Guo, Kai Wang, Hong Kang, Teng Liu, Yingqi Gao, and Tao Li. Bin loss for hard exudates segmentation in fundus images. *Neurocomputing*, 392:314–324, 2020.

[Guo *et al.*, 2024] Tianjiao Guo, Jie Yang, and Qi Yu. Diabetic retinopathy lesion segmentation using deep multiscale framework. *Biomedical Signal Processing and Control*, 88:105050, 2024.

[Han *et al.*, 2018] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image super-resolution via dual-state recurrent networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 1654–1663, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[Hoover *et al.*, 2000] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.

[Hu *et al.*, 2023] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 881–889, 2023.

[Huang *et al.*, 2022] Shiqi Huang, Jianan Li, Yuze Xiao, Ning Shen, and Tingfa Xu. Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Transactions on Medical Imaging*, 41(6):1596–1607, 2022.

[Li *et al.*, 2019a] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

[Li *et al.*, 2019b] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3867–3876, 2019.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125, 2017.

[Liu *et al.*, 2021a] Qing Liu, Haotian Liu, Yang Zhao, and Yixiong Liang. Dual-branch network with dual-sampling modulated dice loss for hard exudate segmentation in color fundus images. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1091–1102, 2021.

[Liu *et al.*, 2021b] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Qing Liu, Haotian Liu, Wei Ke, and Yixiong Liang. Automated lesion segmentation in fundus images with many-to-many reassembly of features. *Pattern Recognition*, 136:109191, 2023.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440, 2015.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International conference on 3D vision (3DV)*, pages 565–571, 2016.

[Porwal *et al.*, 2018] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[Porwal *et al.*, 2020] Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, et al. Idrid: Diabetic retinopathy–segmentation and grading challenge. *Medical image analysis*, 59:101561, 2020.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Staal *et al.*, 2004] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[Stankovic and Falkowski, 2003] Radomir S Stankovic and Bogdan J Falkowski. The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44, 2003.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning (ICML)*, pages 6105–6114, 2019.

[Teo *et al.*, 2021] Zhen Ling Teo, Yih-Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyungtaek Rim, Ning Cheung, Mukharram M Bikbov, Ya Xing Wang, Yating Tang, Yi Lu, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*, 128(11):1580–1591, 2021.

[Thomas *et al.*, 2019] RL Thomas, S Halim, S Gurudas, S Sivaprasad, and DR Owens. Idf diabetes atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018. *Diabetes research and clinical practice*, 157:107840, 2019.

[Ting *et al.*, 2016] Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & experimental ophthalmology*, 44(4):260–277, 2016.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[Wei *et al.*, 2020a] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12321–12328, 2020.

[Wei *et al.*, 2020b] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 13025–13034, 2020.

[Xiao *et al.*, 2018] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[Zhang *et al.*, 2022] Jiayi Zhang, Xiaoshan Chen, Zhongxi Qiu, Mingming Yang, Yan Hu, and Jiang Liu. Hard exudate segmentation supplemented by super-resolution with multi-scale attention fusion module. In *International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1375–1380, 2022.