HRDecoder: High-Resolution Decoder Network for Fundus Image Lesion Segmentation

Ziyuan Ding[1], Yixiong Liang[1]✉[0000−0003−0407−5838], Shichao Kan[1], and Qing Liu[2]

[1] Central South University, Hunan, China
{ziyuanding, yxliang, kanshichao}@csu.edu.cn
[2] University of Oulu, Oulu, Finland

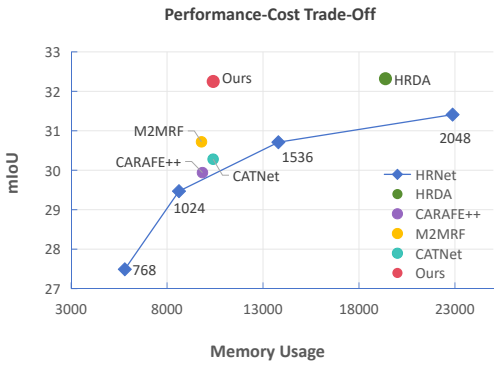
Abstract. High resolution is crucial for precise segmentation in fundus images, yet handling high-resolution inputs incurs considerable GPU memory costs, with diminishing performance gains as overhead increases. To address this issue while tackling the challenge of segmenting tiny objects, recent studies have explored local-global feature fusion methods. These methods preserve fine details using local regions and capture context information from downscaled global images. However, the necessity of multiple forward passes inevitably incurs significant computational overhead, greatly affecting inference speed. In this paper, we propose HRDecoder, a simple High-Resolution Decoder network for fundus image segmentation. It integrates a High-resolution Representation Learning (HRL) module to capture fine-grained local features and a High-resolution Feature Fusion (HFF) module to fuse multi-scale local-global feature maps. HRDecoder effectively improves the overall segmentation accuracy of fundus lesions while maintaining reasonable memory usage, computational overhead, and inference speed. Experimental results on the IDRID and DDR datasets demonstrate the effectiveness of our method. The code is available at https://github.com/CVIU-CSU/HRDecoder.

Keywords: Fundus Image · Lesion Segmentation · High-Resolution.

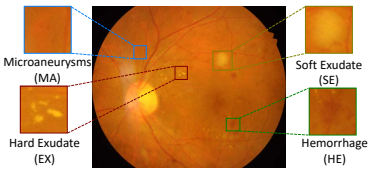
1 Introduction

Fundus image lesion segmentation poses a significant challenge in medical image analysis, which is crucial for the early detection and monitoring of various retinal diseases. The pixel-wise classification of tiny lesions , as shown in Fig. 1b, demands considerably higher resolution compared to other segmentation tasks. Simply increasing the input resolution does boost performance of tiny lesions. While this enhancement is accompanied by rising memory usage,increasing computational overhead, and slower inference speed. These issues significantly hinder practical application and further performance improvement of models.

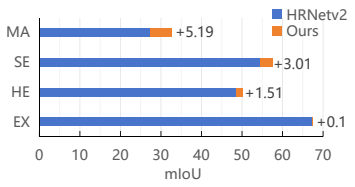
Numerous efforts have been made to address the segmentation of tiny objects efficiently, two primary approaches are explored. Some studies explore FPN-like [12] or UNet-like [20] multi-scale features fusion methods to compensate for



(a) Performance-cost Trade-off on DDR [10]



(b) An example from IDRiD [19]



(c) Improvement on IDRiD [19]

Fig. 1: (a) Our method reaches SOTA performance and is memory efficient, the numbers represent input resolutions. (b) Example of tiny lesions in fundus images. (c) Performance gains on each category.

fine-grained details and enhance performance on small objects [29,15,22,6,5,7,27]. However, these bottom-up global feature fusion methods are susceptible to distortion from low-resolution feature maps, resulting in suboptimal performance. There are also other methods that emphasize local feature fusion [13,14,25,2,8]. M2MRF [13] enhances performance on tiny lesions by designing specialized local feature fusion modules, while it excessively emphasizes local features and exhibits slow convergence on larger lesions. HRDA [8] and GLNet [2] design a dual-branch network composed of a High-Resolution(HR) branch to learn texture details from randomly cropped images and a Low-Resolution(LR) branch to extract contextual information through scaling operations, consequently achieving impressive results. While they need multiple forward passes, which significantly increase computational overhead and sharply decrease inference speed, thus restricting further applications.

Instead, we propose HRDecoder, a simple framework combining the idea of local-global high-resolution crops and multi-scale feature fusion at the decoder stage, to efficiently and effectively segment tiny lesions. HRDecoder consists of a HR Representation Learning (HRL) module to mine detailed features and a HR Feature Fusion (HFF) module for integrating multi-scale features, significantly enhancing performance on small lesions (see Fig. 1c). By simply using scaling and cropping and a light-weight decoder, our method does not introduce extra parameters, and can significantly alleviate issues of high memory usage, computational overhead, and slow inference speed, as shown in Fig. 1a.

We summarize the contributions as follows: (1) We propose HRDecoder, a simple framework to address the challenge of segmenting tiny lesions in fundus images. (2) Our method not only improves segmentation performance but

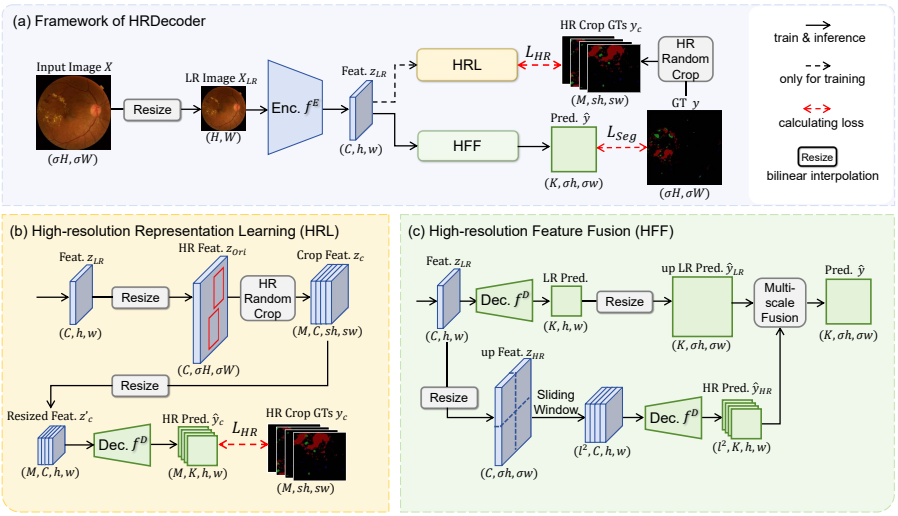


Fig. 2: Overview of our proposed HRDecoder. (a) Training and testing pipeline. (b) The HRL module aim to learn local detailed features from simulated HR feature maps. (c) The HFF module aggregates HR multi-scale features.

also effectively mitigates high memory consumption, high computational overhead and slow inference speed. (3) Our framework dose not introduce any extra trainable parameters and can be easily applied to existing methods.

2 Method

In this work, we propose HRDecoder for fundus image lesion segmentation. Our method consists of two modules: one that simulates high-resolution inputs to enhance the representation learning of small objects (Sec. 2.1), and another that integrates high-resolution multiscale feature maps to capture detailed information and local contextual cues for tiny targets (Sec. 2.2). Our approach can be easily applied to various networks, enhancing performance at reasonable costs in terms of memory usage, computational overhead, and inference speed.

As shown in Fig. 2a, given an input image $x \in \mathbb{R}^{3\times\sigma H\times\sigma W}$ with corresponding GT $y \in \mathbb{R}^{K\times\sigma H\times\sigma W}$, we first downsample x to $x_{LR} \in \mathbb{R}^{3\times H\times W}$. LR features $z_{LR}{=}f_E(x_{LR}) \in \mathbb{R}^{C\times h\times w}$ are then extracted by encoder f_E. Subsequently, the HRL module is adopted to enhance detail representation during training, and the HFF for generating prediction $\hat{y} \in \mathbb{R}^{K\times\sigma h\times\sigma w}$ during training and inference.

2.1 High-resolution Representation Learning

In [8], HR local patches are randomly cropped out to maintain details and simultaneously original image is resized into LR inputs to learn long-range context.

The main drawback is that images need to go through encoder f_E multiple times, which incurs considerable memory consumption and computational overhead. Moreover, sliding window is used for preserving details during inference, resulting in a notable reduction in inference speed. Therefore, we introduce the HRL module to address above issues (see Fig. 2b).

The HRL simply uses a lightweight decoder to mine details from HR feature maps. Specifically, we first upsample feature z_{LR} and obtain $z_{Ori} \in \mathbb{R}^{C\times\sigma H\times\sigma W}$, then random cropping is adopted to get M patches of features z_c and corresponding GT y_c. We follow [8] to randomly sample M bounding boxes $b_c^i, i \in \{0, 1...M-1\}$ from a discrete uniform distribution within the original size:

$$z_c^i = z_{Ori}[b_{c,1}^i : b_{c,2}^i, b_{c,3}^i : b_{c,4}^i], \quad y_c^i = y[b_{c,1}^i : b_{c,2}^i, b_{c,3}^i : b_{c,4}^i], \tag{1}$$

$$\begin{aligned} b_{c,1}^i &\sim \mathcal{U}\{0,\ (\sigma H - sh)/k\} \cdot k, \quad b_{c,2}^i = b_{c,1}^i + sh, \\ b_{c,3}^i &\sim \mathcal{U}\{0,\ (\sigma W - sw)/k\} \cdot k, \quad b_{c,4}^i = b_{c,3}^i + sw. \end{aligned} \tag{2}$$

Here, we adjust the coordinates to be divided by k in order to mitigate distortion of features and the scale ratio of random cropping s can be represented as (1-δ,1+δ). We set s to mimic random crop operation in preprocessing to enable the model to further learn lesion features at different scales. Later, z_c will be scaled to a uniform size of $h \times w$ and obtain z_c'. With resized features z_c', we calculate the HR loss L_{HR} for cropped GT masks y_c, which is formulated as:

$$L_{HR} = \Sigma_i^{|K|} L_{Dice}(f_D(z_c')^i, y_c^i), \tag{3}$$

where L_{Dice} represents the binary Dice loss and $|K|$ is the number of classes. With the supervision of L_{HR}, the segmentation model can progressively learn intricate local features from high-resolution GT, thereby mitigating distortions during upsampling and downsampling processes and enhancing segmentation performance, especially for tiny lesions.

2.2 High-resolution Feature Fusion

The HRL module empowers the extraction of intricate information from large-scale LR feature maps. Nevertheless, the HRL module tend to over emphasize localized tiny lesions and disregard larger ones. Thus, the small-scale LR feature map is essential to provide a holistic contextual understanding. To this end, we propose a simple HFF module to integrate multi-scale feature maps and generate a more comprehensive and refined prediction.

Specifically, we follow [8,2] and adopt a dual-branch network. As shown in Fig. 2c, the LR branch inputs the LR feature z_{LR} through the decoder f_D to obtain global prediction, which is then upsampled to $\hat{y}_{LR}$ by a factor of σ. This branch serves to provide rich global contextual information. Meanwhile, the HR branch takes in HR feature z_{HR}, also upscaled by a factor of σ, to capture detailed local information. We use a sliding window to get l^2 crops, where l^2 denotes the number of windows. Typically, we set both the stride and window

size as (h, w). HR prediction $\hat{y}_{HR}$ is later obtained by aggregating the results from sliding windows. With $\hat{y}_{LR}$ and $\hat{y}_{HR}$, we fuse them together to generate the final prediction $\hat{y}$:

$$\hat{y} = f_A(z) \odot \hat{y}_{HR} + (1 - f_A(z)) \odot \hat{y}_{LR}. \tag{4}$$

Here, we use $f_A(\cdot)$ to denote a general form of scale attention. In HRDA [8], it represents using an additional decoder head f_A to learn a scale attention on the LR branch. While in our method, we simply design f_A as a fixed value of 0.5. Given that most regions in fundus images are background areas, learning scale attention from LR branch may hinder the ability to capture a coherent scene layout, while weighted sum can better grasp lesion details and contextual information (further discussed in Tab. 3). With the fuse prediction $\hat{y}$, we can calculate the segmentation loss

$$L_{Seg} = \Sigma_i^{|K|} L_{Dice}(\hat{y}^i, y^i). \tag{5}$$

Eventually, the overall loss function is formulated as:

$$L = L_{Seg} + \lambda L_{HR}, \tag{6}$$

where λ is hyper-parameter and is empirically set to 0.1.

By employing the HRL and HFF modules, our HRDecoder learns detailed features and retains contextual information using small-scale inputs through the encoder and large-scale feature maps through the decoder. By simply using interpolation and cropping operations, our method does not introduce extra parameters and can be easily applied to existing methods. Furthermore, given that the resource overhead of the decoder is significantly lower than that of the encoder, HRDecoder can effectively alleviate issues such as high memory usage, computational overhead, and slow inference speeds.

3 Experiment

3.1 Datasets and Implementation Details

Datasets: We conduct experiments on two main public retinal lesion segmentation datasets, i.e. IDRiD [19] and DDR [10]. The IDRiD dataset contains 81 high-quality retinal lesion segmentation images with a unified resolution of 4288×2848, 54 for training and 27 for testing. The DDR dataset consists of 757 color fundus images, with 383 for training, 149 for validation and the rest 225 for testing. The image resolutions vary from 1088×1920 to 3456×5184. Both datasets provide pixel-level annotations for four different lesions, i.e. hard exudates (EX), hemorrhages (HE), soft exudates (SE), and microaneurysms (MA).

Implementation Details: HRDecoder is implemented based on the MMSegmentation [4] framework. We adopt HRNetv2 [23] as backbone and simple FCNHead [18] as decoder. Images from IDRiD and DDR datasets are resized to

Table 1: Comparison with previous SOTA methods on IDRiD [19] and DDR [10] datasets. † represents we reproduce it with the same experimental settings as ours. * means we implement our method with ConvNeXt [17] backbone. Results are averaged over three repetitions.

Methods	AUPR					F					IoU				
	EX	HE	SE	MA	mAUPR	EX	HE	SE	MA	mF	EX	HE	SE	MA	mIoU
IDRiD															
DNL [28]	75.12	64.04	64.73	32.48	59.09	73.15	61.87	63.96	32.78	57.94	57.67	44.80	47.03	19.61	42.28
HRNetv2† [23]	82.75	67.85	71.96	44.01	66.64	80.56	65.35	70.36	42.97	64.81	67.44	48.53	54.62	27.36	49.49
Swin† [16]	85.39	68.67	74.28	43.47	67.95	77.57	65.36	70.88	45.59	64.85	63.32	48.22	54.77	29.58	48.97
Segformer† [26]	82.47	69.23	73.69	37.63	65.75	76.10	64.55	69.49	40.65	62.70	61.43	47.66	53.24	25.52	46.96
Mask2Former† [3]	84.68	69.01	74.88	42.53	67.78	79.64	66.47	68.05	44.64	64.70	66.17	49.78	51.59	28.75	49.07
IFA [9]	81.92	69.01	70.47	46.35	66.94	79.80	**67.43**	69.12	46.35	65.68	66.39	**50.86**	52.82	30.17	50.06
PCAA [14]	81.63	66.74	75.49	43.33	66.80	79.58	64.59	**74.13**	43.17	65.37	66.09	47.70	**58.89**	27.53	50.05
TGANet [21]	82.16	65.60	68.86	42.19	64.70	80.01	63.46	67.89	41.29	63.16	66.67	46.48	51.39	26.01	47.64
LViT [11]	82.19	63.36	70.32	43.65	64.88	79.99	60.96	69.33	43.44	63.43	66.65	43.85	53.06	27.74	47.82
M2MRF† [13]	82.10	67.96	71.77	46.83	67.17	79.81	65.93	70.36	46.28	65.60	66.40	49.18	54.31	30.18	50.02
ConvNeXt† [17]	83.96	**72.64**	77.12	45.97	69.93	76.52	68.17	72.82	47.48	66.25	61.98	51.72	57.26	31.13	50.52
Bi-VLGM [24]	82.48	69.32	74.50	46.20	68.13	80.51	67.42	72.95	45.98	66.71	67.38	50.85	57.41	29.85	51.37
HRDecoder	**87.55**	70.80	**77.65**	**49.16**	**71.29**	**80.61**	66.68	72.99	**49.35**	**67.41**	**67.54**	50.04	57.63	**32.55**	**51.94**
HRDecoder*	84.65	71.53	76.23	50.28	70.67	79.34	68.45	72.89	50.58	67.82	66.10	52.26	57.45	33.96	52.44
DDR															
DNL [28]	56.05	47.81	42.01	14.71	40.15	53.36	42.71	40.40	15.60	38.02	36.39	27.15	25.33	8.46	24.33
HRNetv2† [23]	61.48	51.01	47.42	24.62	46.13	58.78	48.95	46.17	24.61	44.63	41.58	32.32	29.98	13.98	29.47
Swin† [16]	64.49	55.80	50.07	19.51	47.47	60.05	53.46	47.96	25.97	46.86	42.91	36.50	31.23	14.92	31.39
Segformer† [26]	61.43	52.53	33.88	19.88	41.93	57.49	46.47	32.16	24.73	40.21	40.34	30.30	19.18	14.11	25.98
Mask2Former† [3]	63.28	55.26	51.87	19.03	47.36	59.60	49.73	**53.36**	23.63	46.58	42.45	33.35	**36.28**	13.40	31.37
IFA [9]	61.51	46.19	48.90	12.98	42.40	56.76	46.28	48.25	0.55	37.96	39.62	30.11	31.80	0.28	25.45
PCAA [14]	60.57	57.46	41.49	18.58	44.53	56.89	54.47	36.68	20.57	42.15	39.76	37.43	22.46	11.47	27.78
TGANet [21]	60.49	52.63	43.55	26.81	45.87	58.92	42.19	41.27	26.92	42.33	41.76	26.73	26.00	15.55	27.51
LViT [11]	61.35	46.29	48.06	27.61	45.83	59.15	42.85	46.88	27.78	44.17	42.00	27.27	30.62	16.13	29.01
M2MRF† [13]	63.74	54.88	49.95	**27.91**	49.12	60.41	47.60	48.73	27.70	46.11	43.28	31.27	32.25	16.08	30.72
ConvNeXt† [17]	63.87	**58.03**	50.61	15.89	47.10	58.71	**55.18**	50.91	22.22	46.76	41.50	**38.05**	34.08	12.37	31.50
Bi-VLGM [24]	62.01	57.38	50.95	26.19	49.13	57.90	54.38	50.81	26.06	47.29	40.75	37.34	34.06	14.98	31.78
HRDecoder	**64.84**	55.69	**51.93**	24.60	**49.27**	**60.67**	52.70	51.57	**27.92**	**48.21**	**43.54**	35.03	34.22	**16.22**	**32.25**
HRDecoder*	63.73	55.40	52.70	21.08	48.20	59.81	53.67	53.56	26.44	48.37	42.67	36.67	36.57	15.17	32.77

1440×960 and 1024×1024 in previous protocols; and we set scale factor σ to 2, i.e. 2880×1920 for IDRiD and 2048×2048 for DDR, respectively. To reduce memory cost during training on IDRiD dataset, we randomly crop images to 1920×1920 and sliding window is used for inference. SGD with a learning rate of 0.01 is used for optimization. Total batch size is 4, iterations are set to 20k for IDRiD and 40k for DDR. For hyper-parameters, crop number M is set to 2 and 4, respectively. We set δ=0.25 to implicitly learn multi-scale features. For evaluation metrics, we follow [13,24] and utilize the commonly used IoU, F-score, AUPR and their mean values.

3.2 Comparison with State-of-the-Art (SOTA) Methods

SOTA Segmentation Methods: First, we present a comprehensive comparison between our method and previous SOTA approaches including different backbones [23,16,26,17], feature-enhanced methods [28,3,14,13] and multi-modal methods [9,21,11,24] on IDRiD [19] and DDR [10] datasets in Tab. 1. The best and second-best scores are denoted in bold and underlined, respectively. We

Table 2: Comparison with different feature fusion methods on IDRiD [19] and DDR [10] testing set. ‡ means the result is from [6] or [24]. Results are averaged over three repetitions.

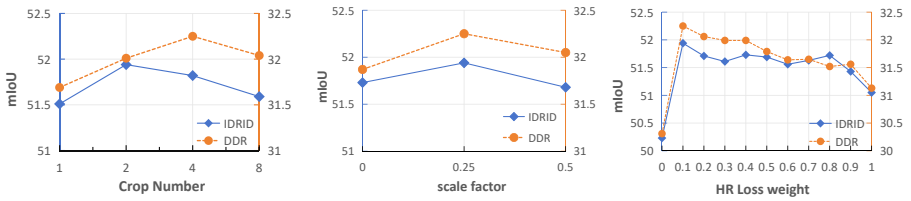
Methods	IDRiD			DDR			Params(M)	GFLOPs	Memory(GB)	FPS(img/s)
	mAUPR	mF	mIoU	mAUPR	mF	mIoU				
HRNet-1024 [23]	66.64	64.81	49.49	46.13	44.63	29.57	65.57	355.62	8.42	6.92
HRNet-2048 [23]	69.37	66.87	51.45	48.56	46.96	31.41	65.57	1422.48	22.34	2.36
CARAFE++ [22]	67.56	65.52	49.92	47.48	45.27	29.94	66.13	481.71	9.61	4.88
PMCNet‡ [6]	68.08	56.02	43.12	36.44	39.31	32.29	–	–	–	–
SenFormer [1]	66.36	63.04	47.35	43.80	41.97	27.36	68.81	413.43	9.74	5.45
CATNet [15]	67.36	65.82	50.34	45.74	45.57	30.28	67.34	383.05	10.18	7.23
PCAA‡ [14]	66.80	65.37	50.05	44.52	52.15	27.78	–	–	–	–
M2MRF [13]	67.17	65.60	50.02	49.12	46.11	30.72	70.30	353.65	9.56	7.56
HRDA [8]	71.17	67.45	52.09	49.13	47.95	32.32	66.06	2113.45	18.93	1.52
HRDecoder	71.29	67.41	51.94	49.27	47.81	32.25	65.57	440.69	10.26	4.86

primarily report results utilizing HRNet as backbone for fair comparison. For IDRiD dataset, HRDecoder achieves results of 71.29%, 67.41%, and 51.94% in mAUPR, mF, and mIoU, respectively. Our method outperforms previous SOTA by a large margin of 3% in mAUPR. In terms of each lesion, our method achieves the best or second-best in 3 out of 4 in AUPR, F and IoU. For DDR dataset, consistent conclusions can be drawn as on IDRiD.

We observe that ConvNeXt [17] and Swin [16,3] exhibit relatively good performance. We assert that CNN models or local window attention can effectively capture intricate features within localized regions. In contrast, Transformer-based methods [26,1] struggle to extract detailed information from fundus images due to long-range dependency mechanism. Therefore, a simple CNN decoder is helpful in capturing fine-grained features efficiently and effectively. Benefiting from the high-resolution design in [23] and our exploration of HR feature maps, HRDecoder significantly improves performance on tiny lesions within fundus images. Visual results on IDRiD and DDR are provided in supplementary material.

Multi-scale Feature Fusion Methods: To demonstrate the effectiveness and efficiency of HRDecoder, we compare various multi-scale feature fusion methods in terms of performance, parameter, computational overhead, GPU memory, and inference speed in Tab. 2. We report results of HRNet [23] at different scales in the first group. The second group [22,6,1,15] employs FPN-like [12] global feature fusion methods. The third group [14,13,8] utilizes local feature fusion methods.

Global feature fusion methods integrate features across multiple scales and improve segmentation performance. However, they are generally less effective than local feature fusion methods. The HRDA [8] can significantly improve segmentation performance. However, it requires 5×GFLOPs and 2×GPU memory compared with above methods due to multiple forwards of encoder. Additionally, the sliding window notably reduces its inference speed. In contrast, our method achieves comparable or even superior performance to HRDA without significantly increasing computational overhead or memory usage. Overall, HRDecoder can achieve a better performance-cost trade-off.



(a) (b) (c)

Fig. 3: Ablation on HRL module. (a) Influence of the number of HR crops M. (b) Influence of scale ratio δ of HR crops. (c) Influence of HR loss weigh λ.

Table 3: Comparison of different fusion method on IDRiD test set.

Fusion Methods $f_A(\cdot)$	mAUPR	mF	mIoU
LR Attn.	70.34	67.06	51.60
HR Attn.	71.02	67.35	51.83
Weighted Sum	71.29	67.41	51.94

Table 4: Influence of HR scale factor σ on IDRiD test set.

Scale Factor σ	EX	HE	SE	MA	mIoU
1×	67.44	48.53	54.62	27.36	49.49
2×	67.54	50.04	57.63	32.55	51.94
4×	70.06	45.63	0.00	34.49	37.55

3.3 Ablation Study

In Fig. 3, we discuss the ablation study on HRL module. Fig. 3a shows the impact of different numbers of HR crops. Training with larger M can enhance the ability to capture details, while too many HR crops may lead the model to focus excessively on tiny lesions e.g. MA and neglect larger lesions e.g. SE. Hence, we set M to 2 and 4 for IDRiD and DDR, respectively. Fig. 3b illustrates the influence of scale ratio (1-δ,1+δ). A larger δ results in a larger ratio range of HR crop size. It is similar to the random crop operation during preprocessing. Implicitly inputting multi-scale features enables the model learn features of various scales to address lesions of different sizes. We set δ=0.25, as overly large δ may introduce excessive uncertainty. Fig. 3c demonstrates the sensitivity to HR loss weight λ. A larger λ tends to emphasize local features and diminish contextual information. Experimental results show that 0.1 is a preferable choice.

We also conduct ablation studies on the HFF module. In Tab. 3 we compare three different fusion strategies: learning scale attention from LR branch, from HR branch and weighted fusion. In HRDA [8], attention is learned from LR branch using an attention head, which is heavily affected by the predominance of background regions in fundus images. While learning attention from HR detailed features or simple weighted fusion is favored and we adopt the simple parameter-free method. Tab. 4 demonstrates the effect of scaling ratios σ. σ=1 corresponds to the vanilla HRNet [23]. When σ is set too large, HR crops will excessively focus on tiny lesions e.g. EX and MA. However, for slightly larger lesions e.g. SE, the model may misclassify due to severe lack of local context. Further ablation studies are provided in supplementary material.

4 Conclusion

In this paper, we propose HRDecoder, a straightforward framework that combines an HRL module to mine local details and an HFF module to capture contextual information, for fundus image segmentation. Our approach strikes a balance between performance and memory usage, inference speed, and overheads without introducing extra parameters. Overall, HRDecoder can achieve SOTA performance while maintaining manageable overheads. There are certain limitations: the simple feature fusion is specifically designed for fundus images, whereas scale attention may have broader applicability in various tasks. Additionally, the simple CNN-based decoder may fall short in capturing contextual information. Despite these limitations, we believe that our approach provides a simple and versatile solution.

Acknowledgments. This manuscript was supported in part by the National Key Research and Development Program of China under Grant 2021YFF1201202, the Key Research and Development Program of Hunan Province under Grant 2023SK2029, and the Natural Science Foundation of Hunan Province under Grant 2024JJ5444 and 2023JJ30699. The authors wish to acknowledge High Performance Computing Center of Central South University for computational resources.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y.H., Song, X.: Efficient self-ensemble for semantic segmentation. In: 33rd British Machine Vision Conference (2022)
2. Chen, W., Jiang, Z., Wang, Z., Cui, K., Qian, X.: Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8924–8933 (2019)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
4. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020)
5. Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., Chen, X.: Cpfnet: Context pyramid fusion network for medical image segmentation. IEEE Trans. Med. Imaging **39**(10), 3008–3018 (2020)
6. He, A., Wang, K., Li, T., Bo, W., Kang, H., Fu, H.: Progressive multiscale consistent network for multiclass fundus lesion segmentation. IEEE Trans. Med. Imaging **41**(11), 3146–3157 (2022)
7. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6202–6212 (2023)

8. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 372–391. Springer (2022)
9. Hu, H., Chen, Y., Xu, J., Borse, S., Cai, H., Porikli, F., Wang, X.: Learning implicit feature alignment function for semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 487–505. Springer (2022)
10. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
11. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE Trans. Med. Imaging (2023)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
13. Liu, Q., Liu, H., Ke, W., Liang, Y.: Automated lesion segmentation in fundus images with many-to-many reassembly of features. Pattern Recognition **136**, 109191 (2023)
14. Liu, S.A., Xie, H., Xu, H., Zhang, Y., Tian, Q.: Partial class activation attention for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16836–16845 (2022)
15. Liu, Y., Li, H., Hu, C., Luo, S., Luo, Y., Chen, C.W.: Learning to aggregate multi-scale context for instance segmentation in remote sensing images. IEEE Transactions on Neural Networks and Learning Systems (2024)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
19. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Med. Image Anal. **59**, 101561 (2020)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. pp. 234–241. Springer, Cham (2015)
21. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: Tganet: Text-guided attention for improved polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. pp. 151–160. Springer, Cham (2022)
22. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe++: Unified content-aware reassembly of features. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 4674–4687 (2022)
23. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2020)

24. Wenting, C., Jie, L., Yixuan, Y.: Bi-vlgm: Bi-level class-severity-aware vision-language graph matching for text guided medical image segmentation. arXiv preprint arXiv:2305.12231 (2023)
25. Wu, H., Wang, W., Zhong, J., Lei, B., Wen, Z., Qin, J.: Scs-net: A scale and context sensitive network for retinal vessel segmentation. Med. Image Anal. **70**, 102025 (2021)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Proceedings of the IEEE/CVF international conference on computer vision **34**, 12077–12090 (2021)
27. Yang, Y.H., Huang, T.E., Sun, M., Bulò, S.R., Kontschieder, P., Yu, F.: Dense prediction with attentive feature aggregation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 97–106 (2023)
28. Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H.: Disentangled non-local neural networks. In: Proceedings of the European Conference on Computer Vision. pp. 191–207. Springer (2020)
29. Zhu, G., Wang, R., Liu, Y., Zhu, Z., Gao, C., Liu, L., Sang, N.: An adaptive post-processing network with the global-local aggregation for semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2023)