2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China

https://www.ic-icc.cn/2025/index.php

Hierarchical Refinement and Bilateral Attention Fusion for Polyp Segmentation

Pusheng An[1], JiYa MengHe[1], and He Yi[2]

[1] College of Computer, Inner Mongolia University Hohhot, 010020, China
anpusheng1@163.com, csmhiy@imu.edu.cn
[2] School of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China

Abstract. In the field of medical imaging, polyp segmentation is a crucial task as it enables doctors to accurately identify and segment polyps in endoscopic images and other medical images. Currently, numerous deep learning-based polyp segmentation models mainly rely on multi-scale feature fusion techniques to delineate the boundaries of polyps. However, these existing methods often fail to consider the interconnection between the model's localization and segmentation processes. Generally, when searching for polyps, people first determine the approximate location of the polyps and then gradually obtain detailed feature information of the polyps. In view of this, we propose a hierarchical refinement multi-scale feature fusion Model named HRFFNet. First, we design a hierarchical refinement feature extraction method to precisely optimize the initially located polyp regions. Then, we develop a feature fusion block named FB, which relies on the overall lesion information to form multi-scale feature representations. Through extensive experiments on four commonly used benchmark datasets, we find that HRFFNet performs outstandingly in polyp segmentation, and its performance significantly surpasses that of existing top-notch models.

Keywords: Polyp Segmentation, Hierarchical Refinement, Bilateral Attention, Feature Fusion.

1 Introduction

In the field of medical image analysis, polyp detection [1,2] is of utmost significance for the early detection of colorectal cancer, prevention of disease progression, and alleviation of symptoms and complications associated with polyps. When there is abnormal hyperplasia in parts of the patient's body, such as the gastrointestinal tract, clinicians will determine whether precise polyp segmentation is required based on the patient's specific condition and professional judgment. Therefore, the use of automated techniques to achieve accurate polyp recognition can provide doctors with more powerful diagnostic support, thereby improving the accuracy and efficiency of diagnosis. With the rapid development of artificial intelligence and image recognition technologies, automated recognition systems have been widely applied in various fields, such as industrial defect detection [3-5], crop pest [6] and disease monitoring [7], greatly enhancing work efficiency and accuracy. In the medical field, numerous researchers are committed to developing efficient polyp segmentation models [8-14]. These models have demonstrated excellent performance on multiple public datasets, providing strong technical support for clinical diagnosis [15-19].

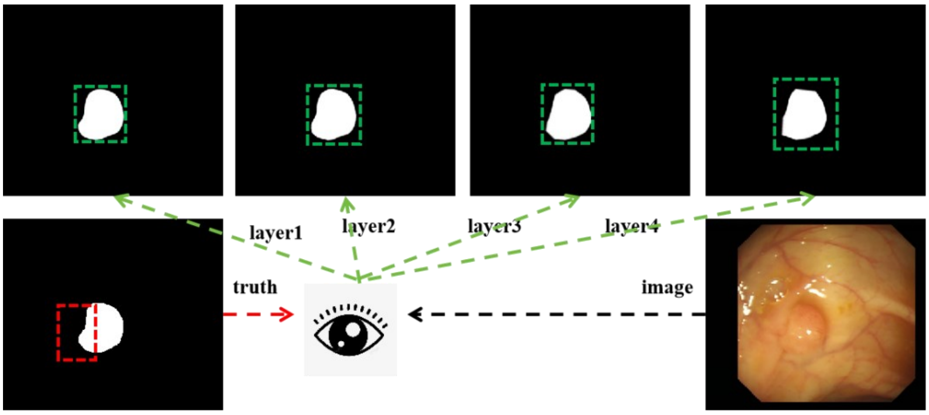

Fig. 1. Deep and shallow visual perception.

Current polyp segmentation models primarily rely on feature extraction and enhancement techniques to achieve segmentation. However, these models often fail to consider the visual processing mechanisms humans employ when observing polyps. As illustrated in Fig. 1, the deepest prediction map of a deep segmentation model can accurately identify the target region, and then progressively refine the edge details from layer4 to layer1. Based on these considerations, we aim to emulate the way humans observe polyps to enhance the segmentation accuracy of the target in tumor segmentation models. We hope to achieve a transition from initial localization of the polyp region to detailed segmentation through the gradual fusion of multi-scale features. However, in this process, we face two major challenges. First, how to effectively integrate deep and shallow information [20] during the multi-layer feature extraction process. Secondly, the existing feature fusion methods often ignore the transmission of background information, which will bring noise interference to the positive sample information. This requires suppression during the fusion process.

In summary, to effectively address the aforementioned challenges, this paper proposes a hierarchical multi-scale feature fusion model for polyp segmentation called HRFFNet. We first perform layer by layer fusion of the decoding layers to simulate the process of human observation of polyps. Subsequently, we design a feature fusion block to capture global lesion information, thereby further enhancing segmentation accuracy. As shown in Fig. 1, in the four feature output layers, the deepest feature map of the decoder primarily focuses on the overall localization area of the polyp. As the feature extraction process progresses layer by layer, the boundary details of the polyp will gradually be refined. Finally, we conducted experiments on the publicly available datasets Kvasir-SEG, ClinicDB, ColonDB, and ETIS. Compared with existing state-of-the-art (SOTA) polyp segmentation methods, our approach achieves better accuracy.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation [21-23] is a crucial task in the field of computer vision. It aims to classify each pixel in an image into different semantic categories, thus enabling in depth understanding of the image content. Since 2012, several important semantic segmentation models have been proposed one after another, each with its unique advantages and disadvantages.In 2012, the FCN [24] model first applied the Convolutional Neural Network (CNN) to the semantic segmentation task. It achieved end to end pixel-level classification through fully convolutional layers. Its advantage lies in its ability to process input images of any size and output segmentation results of the same size as the input image. However, FCN does not adequately consider the relationships between pixels, lacks spatial consistency, produces relatively rough segmentation results, and is insensitive to details in the image. Subsequently, in 2014, the DeepLab [25] series of models introduced dilated convolutions and fully connected conditional random fields, significantly enhancing the ability to capture multi-scale contextual information. Its advantage is that the dilated convolutions expand the receptive field, and at the same time, the conditional random fields are used to optimize the segmentation boundaries, resulting in excellent performance in complex scenes. Nevertheless, this series of models has a high computational complexity and requires a large amount of hardware resources.In 2015, the SegNet [26] model adopted an encoder decoder structure and restored boundary details by retaining pooling indices. The advantage of SegNet lies in its efficient memory usage and lightweight model design, making it well suited for embedded devices. However, SegNet performs poorly when dealing with multi-scale objects, and the restoration of boundary details depends on pooling indices, resulting in limited detail richness.In 2017, RefineNet [27] improved the decoder structure. By fusing low-level and high-level features through long distance residual connections, it further enhanced the segmentation accuracy. RefineNet demonstrated high accuracy and flexibility on multiple datasets and could adjust the network structure according to specific requirements. But it has a high computational complexity and relatively high requirements for hardware resources. In the same year, PSPNet [28] proposed a pyramid pooling module, which effectively captured global contextual information through multi-scale feature fusion. PSPNet performed outstandingly in scene parsing tasks and significantly improved the segmentation performance. However, this model has a large number of parameters and consumes a lot of computational resources.In 2019, DFANet [29] mined high-level features through a deep multi-layer aggregation structure and used a lightweight encoder to aggregate information, effectively reducing the computational load. DFANet performs well in scenarios with high real time requirements, but its performance in handling extremely small targets may be inferior to some more complex models.To address the insufficient ability to represent global information, SegFormer [30] was proposed in 2021. It is built based on the Transformer architecture, removes position encoding, and designs a lightweight decoder. Through multi-scale feature representation and efficient feature fusion, SegFormer achieves a simple and efficient design and demonstrates strong generalization ability in various tasks. However, SegFormer may face challenges when dealing with smaller objects and has a high demand for the amount of data.

2.2 Polyp Segmentation

In 2015, U-Net [10] proposed a U-shaped structure and achieved remarkable results in the field of biomedical image segmentation for the first time. Its advantage lies in the preservation of the image's spatial information through skip connections, which leads to excellent performance in segmentation accuracy and boundary details, especially suitable for small scale datasets. However, U-Net may face the problem of overfitting when dealing with large scale datasets, and its relatively simple network structure makes it difficult to handle complex multi-scale features. The following year, SFANet [31] designed a new selective feature aggregation network, which includes region and boundary constraints. The network predicts the region and boundary of polyps through a shared encoder and two mutually constrained decoders. The main advantage of SFANet is its ability to optimize both region segmentation and boundary detection simultaneously, significantly improving the segmentation accuracy. Nevertheless, its complex network structure may result in longer training times and higher requirements for hardware resources. In 2020, PraNet [11] introduced a global inverse attention mechanism, which further refined the edge information of polyp segmentation. PraNet effectively suppressed background noise through the inverse attention mechanism and improved the accuracy of edge details. However, this mechanism may reduce the segmentation efficiency when dealing with complex multi target scenarios. In 2021, HarDNet-MSEG [32] designed a lightweight backbone network in the field of polyp segmentation, with a speed of up to 86fps while maintaining good segmentation results. Its advantages are high efficiency and real time performance, making it suitable for clinical scenarios that require rapid processing. However, its lightweight design may perform slightly weaker when dealing with complex textures and details.In the same year, EU-Net [33] used a new semantic feature enhancement module (SFEM) to enhance semantic information to assist feature extraction and introduced an adaptive global context module. By enhancing semantic information and global context awareness, EU-Net significantly improved the segmentation accuracy. However, its complex module design may lead to an increase in model training and inference times. Also in 2021, SANet designed a color swapping operation to eliminate the impact of color on polyp segmentation and proposed a shallow layer attention module to filter out the background noise of shallow layer features. The main advantage of SANet is its ability to effectively eliminate color interference and improve the segmentation accuracy. However, its shallow layer attention mechanism may have limited effects on the segmentation of extremely small targets. In 2023, Polyp-PVT [34] used a Transformer based backbone network, which has more powerful segmentation capabilities and robustness. Polyp-

PVT extracts multi-scale long distance dependency features through the Pyramid Vision Transformer (PVT), significantly improving the segmentation performance. However, the Transformer architecture has high demands for computational resources, and its adaptability to small scale datasets still needs to be improved.

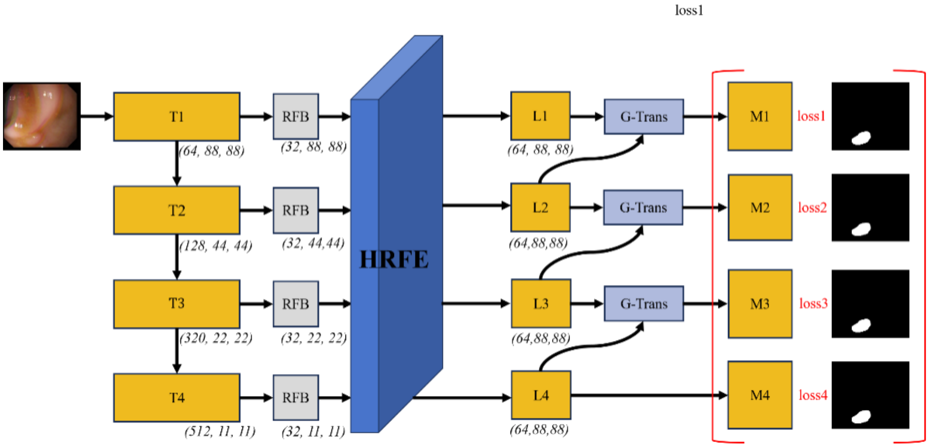

Fig. 2. Structure of HRFFNet.

3 Method

Fig.2 shows the polyp segmentation model HRFFNet that we designed. Firstly, the model adopts a hierarchical refined feature extraction strategy to mine the information of the polyp area from coarse to fine and from shallow to deep. In the first stage, the network acquires the global rough feature map for roughly locating the suspicious area. Subsequently, in each subsequent layer, these features are gradually refined. Through smaller receptive fields and more precise convolution operations, the resolution of the polyp edge and morphology is continuously improved, ultimately achieving precise positioning and description of the target area. This process is equivalent to simulating the visual recognition approach of human doctors, which is to "first scan the overall situation and then gradually focus", ensuring the integrity of positioning while also enhancing the depiction of details. Secondly, in order to make full use of multi-scale context information, HRFFNet introduces Feature Fusion blocks (FB). This module establishes multiple information paths among feature maps at different levels, fusing features from the shallow layer (high resolution, low semantics) and the deep layer (low resolution, high semantics) to obtain a richer multi-scale expression. Specifically, FB organically combines the global structure of the entire lesion with local details through adaptive weighting and pixel-by-pixel fusion. This not only enhances the semantic coherence of the feature map but also improves the robustness of the model to changes in the shape, size, and contrast of polyps. Through the synergy of these two core modules, HRFFNet can achieve high-precision segmentation of the polyp area while maintaining efficient computing.

3.1 Hierarchical Refinement Feature Extraction Strategy

Numerous SOTA polyp segmentation networks at present, such as U-Net, PraNet, and Polyp-PVT, are founded upon the encoding decoding architecture. These networks have already achieved relatively high segmentation accuracy. Inspired by this, our model also adheres to this well-validated strategy.Initially, for the baseline model, we select PolypNet. In the encoding phase, PVT-v2 is employed as the backbone network. Through the enhanced perception module, the dimensions of the feature maps at each layer are optimized and adjusted. Specifically, the number of features per layer is streamlined to 32, thereby significantly enhancing the model's inference efficiency. Subsequently, we obtain four feature maps $T_i(i \in [1,2,3,4])$ for subsequent decoding computations. The detailed model architecture is depicted in Fig.3. As illustrated, we perform layer by layer fusion of the feature maps from the bottom-up direction. Initially, T_4 and T_3 are merged to yield the output of the third layer, and this process is repeated accordingly. As a result, we acquire four feature maps$L_i(i \in [1,2,3,4])$. Ultimately, by leveraging the Transformer (G-Trans) to decode both deep level and shallow level features, we obtain four mask prediction maps $M_i(i \in [1,2,3,4])$.

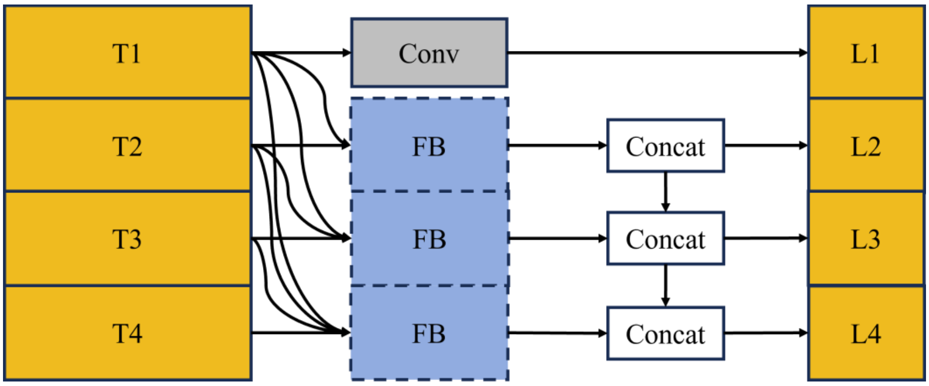

Fig. 3. Structure of HRFE.

3.2 A bilateral attention-based feature fusion block

Feature fusion is vital for progressive feature extraction. An effective module can enhance contextual information extraction, thereby improving feature map predictions across layers. Our designed feature fusion module excels at global feature extraction. As depicted in Fig.4, given the shallow feature f1 and deep feature f2, we first concatenate f1 and f2 along the channel dimension. Then, a volume-based layer processes them to yield the preliminary fused feature fp. The formula is as follows:

$$f_p = Conv(cat[f_1, f_2]) \tag{1}$$

where Conv represents the convolutional layer; cat represents the merging of two features along the channel dimension. Then, we perform a concatenated global pooling operation (GCBlock) on f_p to obtain the feature map f_p' containing global attention information, as shown in the following formula:

$$f_p' = Sigmoid(f_{p1}' + f_{p2}') \tag{2}$$

where f_{p1}' is calculated as follows:

$$f_{p1}' = PConv(Relu(PConv(GAP(f_p)))) \tag{3}$$

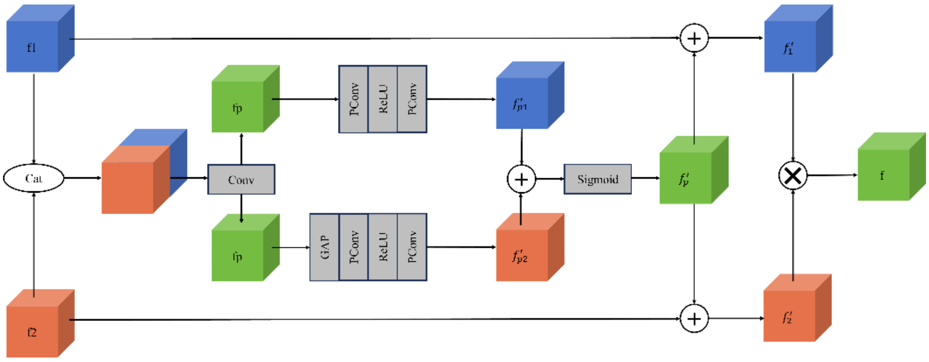

Fig. 4. Structure of FB.

where PConv stands for point-wise convolution; Relu stands for activation function; and GAP stands for global average pooling layer. The corresponding formula for f_{p2}' is as follows:

$$f_{p2}' = PConv(Relu(PConv(f_p))) \tag{4}$$

Then, we add f_p' with the initial feature map to obtain two feature maps f1' and f2' containing global contextual feature information. We fuse f_{p1}' and f_{p2}' to further extract the effective feature point information, and the specific formula is shown below:

$$f = Conv(f_{p1}' \odot f_{p2}') \tag{5}$$

3.3 Loss Function

Our loss function consists of two parts ($Loss_{BCE}$and $Loss_{IOU}$). $Loss_{BCE}$ (Binary Cross-Entropy Loss) is a special case of cross entropy loss function, and the specific calculation formula can be formulated as:

$$Loss_{BCE} = -[y \cdot log\hat{y} + (1-y)log(1-\hat{y})] \tag{6}$$

y represents the true value (0 or 1); $\hat{y}$ means the probability value of belonging to this class. $Loss_{IOU}$ is an area-related loss function, and its calculation formula is as follows:

$$Loss_{IOU} = 1 - \frac{y \cap \hat{y}}{y \cup \hat{y}} \tag{7}$$

Our method will eventually output four forecast images, and the final loss calculation formula is as follows:

$$Loss = \sum_{i=1}^{4} Loss_{BCE} + \sum_{i=1}^{4} Loss_{IOU} \tag{8}$$

4 Experments

4.1 Experimental Setup

Our model is implemented in PyTorch and accelerated by an NVIDIA 4060 Ti GPU. All inputs are uniformly resized to 352×352 and trained with a multi-scale strategy of (0.75, 1, 1.25). To verify different models' validity, the same training strategy and image preprocessing are applied. We use the Adam optimization algorithm to optimize all parameters, setting the learning rate at 0.0001.

4.2 Comparison with State-of-the-arts

To evaluate our model's performance, we carried out comparative experiments on two test sets, Kvasir-SEG and ClinicDB, that were part of the training set. We selected 10 SOTA models (U-Net, U-Net++, SFA, MSEG, DCRNet, ACSNet, PraNet, EU-Net, SANet, and PolypNet) for comparison. Table 1 below presents the results.

Table 1. Quantitative Results of the Test Datasets, i.e., KVASIR-SEG AND CLINICDB.

Model	Kvasir-SEG						ClinicDB					
	mDic	mIoU	F	S	M	MAE	mDic	mIoU	F	S	M	MAE
U-Net	0.818	0.746	0.794	0.858	0.881	0.055	0.823	0.755	0.811	0.889	0.913	0.019
U-Net++	0.821	0.743	0.808	0.862	0.886	0.048	0.794	0.729	0.785	0.873	0.891	0.022
SFA	0.723	0.611	0.670	0.782	0.834	0.075	0.700	0.607	0.647	0.793	0.840	0.042
MSEG	0.897	0.839	0.885	0.912	0.942	0.028	0.909	0.864	0.907	0.938	0.961	0.007
DCRNet	0.886	0.825	0.868	0.911	0.933	0.035	0.896	0.844	0.890	0.933	0.964	0.010
ACSNet	0.898	0.838	0.882	0.920	0.941	0.032	0.882	0.826	0.873	0.927	0.947	0.011
PraNet	0.898	0.840	0.885	0.915	0.944	0.030	0.899	0.849	0.896	0.936	0.963	0.009
EU-Net	0.908	0.854	0.893	0.917	0.951	0.028	0.902	0.846	0.891	0.936	0.959	0.011
SANet	0.904	0.847	0.892	0.915	0.949	0.028	0.916	0.859	0.909	0.939	0.971	0.012
PolypNet	0.912	0.862	0.908	0.924	0.956	0.023	0.931	0.883	0.931	0.945	0.977	0.011
Ours	**0.921**	**0.869**	**0.911**	**0.925**	**0.961**	**0.023**	**0.933**	**0.885**	**0.931**	**0.949**	**0.982**	**0.007**

Table 2. Quantitative Results of the Test Datasets, i.e., ColonDB AND ETIS.

Model	ColonDB						ETIS					
	mDic	mIoU	F	S	M	MAE	mDic	mIoU	F	S	M	MAE
U-Net	0.512	0.444	0.498	0.712	0.696	0.061	0.398	0.355	0.366	0.684	0.643	0.036
U-Net++	0.483	0.410	0.467	0.691	0.680	0.064	0.401	0.344	0.390	0.683	0.629	0.035
SFA	0.469	0.347	0.379	0.634	0.675	0.094	0.297	0.217	0.231	0.557	0.531	0.109
MSEG	0.735	0.666	0.724	0.834	0.859	0.038	0.700	0.630	0.671	0.828	0.854	0.015
DCRNet	0.704	0.631	0.684	0.821	0.840	0.052	0.556	0.496	0.506	0.736	0.742	0.096
ACSNet	0.716	0.649	0.697	0.829	0.839	0.039	0.578	0.509	0.530	0.754	0.737	0.059
PraNet	0.712	0.640	0.699	0.820	0.847	0.043	0.628	0.567	0.600	0.794	0.808	0.031
EU-Net	0.756	0.681	0.730	0.831	0.863	0.045	0.687	0.609	0.636	0.793	0.807	0.067
SANet	0.753	0.670	0.726	0.837	0.869	0.043	0.750	0.654	0.685	0.849	0.881	0.015
PolypNet	0.808	0.727	0.795	0.865	0.913	0.031	0.717	0.649	0.687	0.831	0.852	0.024
Ours	**0.813**	**0.734**	**0.796**	**0.867**	**0.914**	**0.030**	**0.769**	**0.690**	**0.732**	**0.861**	**0.883**	**0.019**

In the experiments conducted on the Kvasir-SEG and ClinicDB datasets, our proposed HRFFNet demonstrates significant quantitative improvements over a wide range of state-of-the-art methods. On Kvasir-SEG, HRFFNet attains a Dice coefficient of 0.921, which represents an absolute improvement of 0.103 over the U-Net baseline (0.818) and an enhancement of 0.009 relative to the best competing method, PolypNet (0.912). This advancement is accompanied by a notable increase in the mean Intersection-over-Union from 0.746 in U-Net to 0.869 in HRFFNet (an absolute gain of 0.123), while the F-measure rises from 0.794 to 0.911, reflecting an increase of 0.117. Additionally, structural similarity (S) and model robustness (M) metrics improve from 0.858 and 0.881 in U-Net to 0.925 and 0.961 in HRFFNet, respectively, and the mean absolute error (MAE) is reduced by more than 50\% compared to U-Net, reaching 0.023. On ClinicDB, HRFFNet achieves a Dice coefficient of 0.933, which is 0.110 higher than that of U-Net (0.823), and an IoU of 0.885 compared to 0.755, corresponding to a substantial absolute increase of 0.130. The F-measure remains robust at 0.931, while the S and M metrics are elevated to 0.949 and 0.982, respectively. Moreover, the MAE in HRFFNet is reduced to 0.007, marking a significant decline in segmentation error relative to the 0.019 observed in U-Net. These comprehensive numerical evaluations across multiple performance indices substantiate that HRFFNet not only achieves superior segmentation accuracy but also offers enhanced model stability and error minimization across both datasets. To further assess generalization capability, we performed comparative experiments on two test sets not included in the training set, namely ColonDB and ETIS. The experimental results are presented in Table 2 below.

As shown in Table 2, HRFFNet demonstrates marked improvements on both ColonDB and ETIS datasets. On ColonDB, the Dice coefficient reaches 0.813, which is a substantial increase over U-Net's 0.512 and slightly higher than PolypNet's 0.808. The mean IoU is measured at 0.734, representing an absolute improvement of 0.290 compared to U-Net and a modest gain of 0.007 over PolypNet. In addition, the F-measure is recorded at 0.796, while the structural similarity and model robustness metrics are 0.867 and 0.914, respectively; these values exceed those of U-Net and are marginally higher than those of PolypNet. Notably, the mean absolute error is reduced to 0.030, nearly half the error observed in U-Net. On the ETIS dataset, HRFFNet continues to outperform its counterparts. It achieves a Dice coefficient of 0.769, which is 0.371 higher than U-Net's 0.398 and 0.052 above PolypNet's score. The mean IoU improves to 0.690, and the F-measure reaches 0.732, both of which signify notable advancements. Structural similarity and model robustness are elevated to 0.861 and 0.883, respectively, and the mean absolute error further decreases to 0.019. Collectively, these diverse performance gains underscore the robustness and superior segmentation capabilities of HRFFNet.

Overall, these results highlight HRFFNet's ability to consistently capture both global lesion structure and fine-grained boundary details across diverse colonoscopy datasets, translating into substantial gains in segmentation accuracy and robustness. Such performance improvements not only demonstrate the model's practical potential for real-time polyp detection in clinical settings but also suggest that the hierarchical refinement and multi-scale fusion strategies could be broadly applicable to other challenging medical imaging tasks.

4.3 Ablation Studies

To further validate the effectiveness of our proposed module, we conducted ablation experiments. The specific results are shown in Table 3:

Model	**Kvasir-SEG**						**ClinicDB**					
	mDic	mIoU	F	S	M	MAE	mDic	mIoU	F	S	M	MAE
b	0.890	0.828	0.872	0.908	0.945	0.031	0.908	0.844	0.897	0.936	0.971	0.126
b+1	0.907	0.852	0.894	0.917	0.951	0.029	0.910	0.852	0.906	0.930	0.969	0.014
b+1+2	**0.921**	**0.869**	**0.911**	**0.925**	**0.961**	**0.023**	**0.933**	**0.885**	**0.931**	**0.949**	**0.982**	**0.007**

In Table 3, we present an ablation study conducted on the Kvasir-SEG and ClinicDB datasets. Here, "b" denotes the baseline model, "b+1" corresponds to the baseline augmented with the HRFE module, and "b+1+2" represents the baseline further enhanced by incorporating the FB module. On the Kvasir-SEG dataset, the baseline achieves a

Dice coefficient of 0.890, which increases to 0.907 with the addition of HRFE and further to 0.921 when the FB module is integrated. Similar trends are observed for the mean IoU, which improves from 0.828 (baseline) to 0.852 with HRFE, and reaches 0.869 with the FB module; corresponding gains are noted in the F-measure, structural similarity, and model robustness metrics, while the mean absolute error is progressively reduced from 0.031 to 0.029 and finally to 0.023. On the ClinicDB dataset, the baseline model records a Dice coefficient of 0.908, which marginally increases to 0.910 with HRFE and then rises to 0.933 after the integration of the FB module. Notably, the mean IoU, F-measure, and other performance indicators also exhibit similar improvements, with the MAE experiencing a dramatic decline from 0.126 at the baseline to 0.014 with HRFE and further to 0.007 upon combining the FB module. These results clearly demonstrate that the progressive integration of HRFE and FB modules not only enhances segmentation accuracy but also improves overall robustness, thereby validating the effectiveness of our proposed modifications.

4.4 Qualitative analysis

Furthermore, we have visualized (as shown in Fig~\ref{fig:fig_pre}.) and analyzed five authoritative polyp detection datasets: CVC-300, CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB, and Kvasir. We selected typical polyp scene images from each dataset to assess the segmentation performance of different models. The prediction results are displayed in a unified format for intuitive comparison: the first column shows original images, the second column shows ground-truth (GT) images, and the third to eighth columns show the prediction results of HRFFNet, PolypNet, PraNet, SFA, U-Net, and U-Net++ models, respectively. HRFFNet excels in the CVC-300 dataset with no extra false detections and the most complete polyp shapes, indicating strong feature extraction and boundary definition. In the CVC-ClinicDB dataset, HRFFNet accurately segments polyps, even those with low contrast, leaving no internal regions undetected. It also performs well in the CVC-ColonDB dataset, producing clean and accurate segmentations without false positives, proving its stability across diverse data distributions. HRFFNet stands out in the ETIS-LaribPolypDB dataset as the only model that precisely predicts polyp shapes and sizes, even in challenging cases like irregular shapes or poor image quality. In the Kvasir dataset, HRFFNet again demonstrates its capability by fully segmenting polyps without missing any internal regions, showing its adaptability to complex intestinal environments. Overall, HRFFNet shows significant advantages in polyp segmentation across all five datasets, outperforming other SOTA methods in shape accuracy, internal region completeness, and adaptability to diverse conditions. These strengths make HRFFNet a promising tool for clinical practice, offering reliable and accurate imaging support for early polyp detection and diagnosis, and enhancing the prevention and treatment of colorectal cancer and other intestinal diseases.

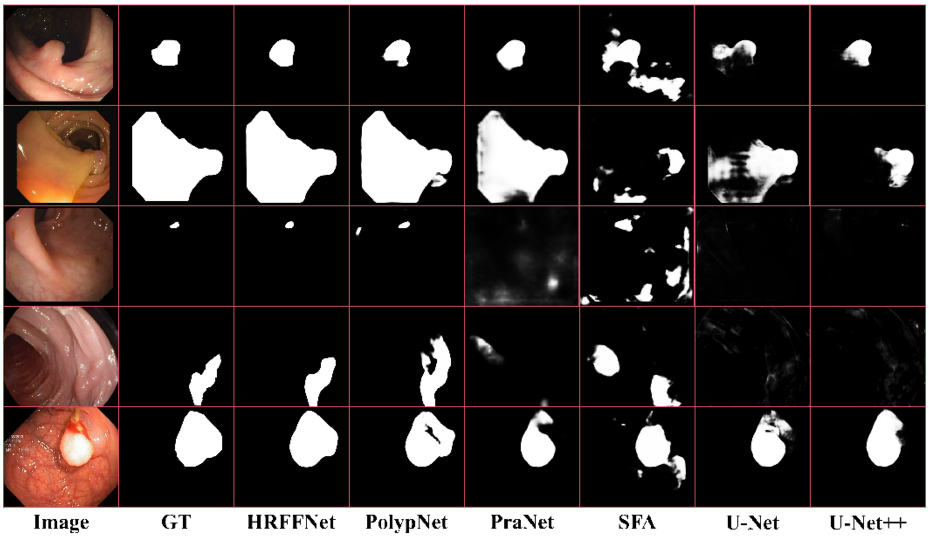

Fig. 4. Visualization of the Predictions.

5 Conclusion

In summary, to simulate the process of doctors searching for polyps, we propose HRFFNet for polyp localization and segmentation. HRFFNet includes two main components: firstly, the Hierarchical Refinement Feature Extraction Strategy (HRFE) imitates doctors' gradual polyp search through stepwise thinking. It achieves effective detail segmentation via incremental feature fusion and extraction. Secondly, to minimize redundant information during fusion, a Bilateral Attention-based Feature Fusion Block (FB) is designed. This block can suppress background information and further refine foreground edge information. HRFFNet reaches state-of-the-art performance on four image-level datasets. In the future, we plan to apply this method to polyp segmentation, providing efficient medical assistance and enhancing model accuracy by integrating video feature extraction.

References

1. Nappi, J., Yoshida, H.: Computer-aided detection of polyps in CT colonography: effect of feature-guided polyp segmentation. In: CARS 2002 Computer Assisted Radiology and Surgery. pp. 749–754 (2002). https://doi.org/10.1007/978-3-642-56168-9_125.

2. Rahim, T., Usman, M., Shin, S.: A Survey on Contemporary Computer-Aided Tumor, Polyp, and Ulcer Detection Methods in Wireless Capsule Endoscopy Imaging. arXiv: Image and Video Processing, (2019).

3. Casner, S.M., Hutchins, E.L., Norman, D.: The challenges of partially automated driving. Communications of the ACM. 59, 70–77 (2016). https://doi.org/10.1145/2830565.

4. Marti, E., de Miguel, M.A., Garcia, F., Perez, J.: A Review of Sensor Technologies for Perception in Automated Driving. IEEE Intelligent Transportation Systems Magazine. 94–108 (2019). https://doi.org/10.1109/mits.2019.2907630.

5. Xia, X., Meng, Z., Han, X., Li, H., Tsukiji, T., Xu, R., Zheng, Z., Ma, J.: An automated driving systems data acquisition and analytics platform. Transportation Research Part C: Emerging Technologies. 104120 (2023). https://doi.org/10.1016/j.trc.2023.104120.

6. Anwar, Z., Masood, S.: Exploring Deep Ensemble Model for Insect and Pest Detection from Images. Procedia Computer Science. 2328–2337 (2023). https://doi.org/10.1016/j.procs.2023.01.208.

7. Um, T.T., Pfister, F.M.J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., Kulić, D.: Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (2017). https://doi.org/10.1145/3136755.3136817.

8. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Richard Tsai, Y.-H.: Automated polyp detection in colon capsule endoscopy. IEEE Transactions on Medical Imaging. 1488–1502 (2014). https://doi.org/10.1109/tmi.2014.2314959.

9. Maghsoudi, O.H.: Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). pp. 1–4 (2017). https://doi.org/10.1109/spmb.2017.8257027.

10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Lecture Notes in Computer Science, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28.

11. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science. pp. 263–273 (2020). https://doi.org/10.1007/978-3-030-59725-2_26.

12. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow Attention Network for Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Lecture Notes in Computer Science. pp. 699–708 (2021). https://doi.org/10.1007/978-3-030-87193-2_66.

13. Cai, L., Wu, M., Chen, L., Bai, W., Yang, M., Lyu, S., Zhao, Q.: Using Guided Self-Attention with Local Information for Polyp Segmentation.

14. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science. pp. 3–11 (2018). https://doi.org/10.1007/978-3-030-00889-5_1.

15. Jha, D., Smedsrud, Pia H., Riegler, M., Halvorsen, P., Lange, T., Johansen, D., Johansen, Havard D.: Kvasir-SEG: A Segmented Polyp Dataset. arXiv: Image and Video Processing, arXiv: Image and Video Processing. (2019).

16. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics. 99–111 (2015). https://doi.org/10.1016/j.compmedimag.2015.02.007.

17. Vazquez, D., Bernal, J., Sánchez, F., Fernández-Esparrach, G., López, A., Romero, A., Drozdzal, M., Courville, A.: A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition. (2016).

18. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. IEEE Transactions on Medical Imaging. 630–644 (2016). https://doi.org/10.1109/tmi.2015.2487997.

19. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery. 283–293 (2014). https://doi.org/10.1007/s11548-013-0926-3.

20. Qin, Y., Kamnitsas, K., Ancha, S., Nanavati, J., Cottrell, Garrison W., Criminisi, A., Nori, Aditya V.: Autofocus Layer for Semantic Segmentation. Cornell University - arXiv, Cornell University - arXiv. (2018).

21. Thoma, M.: A Survey of Semantic Segmentation. arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition. (2016).

22. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing. 41–65 (2018). https://doi.org/10.1016/j.asoc.2018.05.018.

23. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep Semantic Segmentation of Natural and Medical Images: A Review. Artificial Intelligence Review. 137–178 (2021). https://doi.org/10.1007/s10462-020-09854-1.

24. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). https://doi.org/10.1109/cvpr.2015.7298965.

25. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, Alan L.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Le Centre pour la Communication Scientifique Directe - HAL - Diderot, Le Centre pour la Communication Scientifique Directe - HAL - Diderot. (2015).

26. Badrinarayanan, V., Handa, A., Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. Computer Vision and Pattern Recognition, Computer Vision and Pattern Recognition. (2015).

27. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). https://doi.org/10.1109/cvpr.2017.549.

28. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). https://doi.org/10.1109/cvpr.2017.660.

29. Li, H., Xiong, P., Fan, H., Sun, J.: DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00975.

30. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, Jose M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition. (2021).

31. Fang, Y., Chen, C., Yuan, Y., Tong, K.: Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation. In: Lecture Notes in Computer Science, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 302–310 (2019). https://doi.org/10.1007/978-3-030-32239-7-34.

32. Huang, C., Wu, H.-H., Lin, Y.-L.: HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition. (2021).

33. Patel, K., Bur, A.M., Wang, G.: Enhanced U-Net: A Feature Enhancement Network for Polyp Segmentation. In: 2021 18th Conference on Robots and Vision (CRV) (2021). https://doi.org/10.1109/crv52889.2021.00032.

34. Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H., Shao, L.: Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. Cornell University - arXiv, Cornell University - arXiv. (2021).