WDNet: A Novel Wavelet-guided Hierarchical Diffusion Network for Multi-Target Segmentation in Colonoscopy Images

Dongdong He[1][0009−0006−3410−0060], Fang Ma[2], Ziteng Liu[1], Xunhai Yin[3], Hao Liu[4], Wenpeng Gao[1,5] ⋆, Chenghong Zhang[1], and Yili Fu[5]

[1] School of Life Science and Technology, Harbin Institute of Technology, Harbin, 150080, China
[2] Beijing Institute of Aerospace Information, Beijing, 100000, China
[3] Department of Gastroenterology, The First Affiliated Hospital of Harbin Medical University, Harbin, 150001, China
[4] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China
[5] State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin, 150080, China

Abstract. Semantic segmentation in colonoscopy images is pivotal in aiding healthcare professionals to interpret images and enhance diagnostic precision. Nonetheless, the detection of polyps and instruments is challenged by the difficulty in capturing the textures and edges of tiny lesions, and these challenges are exacerbated by low contrast, inconsistent illumination, and noise. To address these challenges, we introduce WDNet, a network adopting a multi-tiered feature extraction and fusion approach, with each encoder layer amalgamating local and global information to construct expressive high-level representations. The input of the network is derived from wavelet transform to dissect images into low- and high-frequency sub-bands, utilizing learnable soft-thresholding to diminish noise while maintaining essential features. High-frequency data are adept at capturing details and edges, whereas low-frequency data furnish a global context. Moreover, WDNet harnesses a diffusion-based decoding mechanism with adaptive step sizes to amplify target region features and mitigate background interference, achieving meticulous segmentation. Comprehensive experiments conducted on a new surgical dataset, along with public benchmarks underscore its remarkable performance. WDNet not only exhibits state-of-the-art performance of semantic segmentation in colonoscopy images with remarkable detail and boundary accuracy but also stands out in processing speed, facilitating the swift handling of extensive datasets. The dataset and source code are available at `https://github.com/hedongdong6060/WDNet`.

Keywords: WDNet · Segmentation · Diffusion Decoding · colonoscopy image.

⋆ Corresponding author

1 Introduction

Colorectal cancer (CRC) is the third most common malignant tumor globally, following lung and breast cancer. Endoscopic polypectomy, a key intervention for preventing CRC and reducing mortality, has become an essential skill for endoscopists [8]. Various polypectomy techniques and devices are used in clinical practice, depending on regional preferences and equipment availability [5]. With increasing pressure on medical resources, exploring sustainable practices in endoscopic surgery to optimize global resource allocation has become urgent [8]. Accurate analysis of endoscopic images can provide surgeons with decision-making support, including precise localization of surgical instruments and anatomical structures, tissue feature recognition, and safety alerts [19]. The integration of medical robotics has significantly enhanced surgical precision and safety [21]. Robot-assisted systems reduce surgical trauma and improve stability, with computer vision technologies, particularly image segmentation algorithms, playing a critical role in locating abnormal tissues and instruments [23].

However, endoscopic image segmentation faces challenges. Variations in polyp and instrument categories, sizes, shapes, and backgrounds complicate segmentation. Existing methods struggle to comprehensively capture texture and edge structures. Additionally, low contrast, uneven illumination, and noise in endoscopic images further increase segmentation difficulty. Complex surgical environments, including smoke, blood, and reflections, make it challenging to distinguish instruments from tissue backgrounds, raising surgical risks [18]. AI-based methods show potential in medical image analysis but often lack computational efficiency and accuracy. Models trained on static images perform poorly on real endoscopic images [26], and existing algorithms lack the ability to model tissue-instrument interactions, limiting their performance in complex scenarios [9].

Early polyp segmentation relied on handcrafted features, such as edge detection [3], but lacked generalization. With the rise of deep learning, networks like U-Net variants [7] became dominant, leveraging multi-scale feature extraction. Some prior works have also incorporated wavelet decomposition for feature enhancement, such as Xnet [29] and WaveCNets [13]. GANs and Mask R-CNN were also applied [1]. Yue et al [25] integrated boundary uncertainty awareness and polyp exploration into a unified framework. Recently, MM Rahman et al [15] introduced EMCAD, a novel efficient multi-scale convolutional attention decoder for medical image segmentation, further advancing the field. Despite progress, challenges remain: high computational costs of complex networks hinder real-time use, and single-target focus limits simultaneous polyp segmentation, which is critical in clinical settings.

Surgical instrument segmentation is crucial for robot-assisted surgery, enabling precise localization and navigation [4]. However, complex environments (e.g., blood, motion artifacts) pose significant challenges. In recent research, Jafar et al [10] proposed a CR-Net, an AI-based encoder-decoder network for surgical instrument segmentation. Nevertheless, several limitations persist: a single-target focus, insufficient robustness under complex conditions, and inadequate real-time performance, which hinder its broader clinical applicability. Further-

more, while models like Diffusion-Wavelet (DiWa) [14] address image denoising, their effectiveness in colonoscopic segmentation needs further refinement.

To address these challenges, our study proposes a novel model architecture for precise segmentation of abnormal tissues and surgical instruments in endoscopic images. In summary, our primary contributions are as follows:

(1) We propose a novel network architecture integrating wavelet transform, multi-level feature extraction and fusion, and diffusion-based feature refinement, which captures global semantic information and local details to significantly improve segmentation accuracy for polyps and surgical instruments while suppressing noise interference.

(2) We introduce a diffusion-based feature refinement module, enabling simultaneous pixel-wise segmentation of polyps and surgical instruments in endoscopic videos. It enhances boundary clarity and reduces background interference, effectively addressing challenges like instrument occlusion and boundary ambiguity.

(3) We contribute a comprehensive dataset of over 10,000 frames from 100 cases of digestive endoscopic surgeries and conduct extensive experiments on multiple public and self-built datasets, demonstrating that our method outperforms state-of-the-art approaches in both segmentation accuracy and computational efficiency, showcasing its practical efficiency and reliability.

2 Method

Overview. We propose a novel network architecture, as illustrated in Fig 1, which integrates three core modules: wavelet transform, multi-level feature extraction and fusion, and a Diffusion Feature Refinement Module. Leveraging the time-frequency localization properties of wavelet transform, the network decomposes the input image into low-frequency and high-frequency sub-bands. The low-frequency sub-band encodes global semantic information and spatial relationships, while the high-frequency sub-band captures fine-grained features such as textures and edges. To further optimize the high-frequency sub-band, we introduce a sparsity prior-based constraint mechanism, which amplifies target region saliency and mitigates noise interference. In the encoder, our multi-level feature extraction and fusion module hierarchically extracts and fuses local features with higher-level representations. A key innovation is the introduction of the Diffusion Feature Refinement Module, which employs an iterative optimization framework based on partial differential equations to enhance boundary clarity and reduce background interference. In the decoder, multi-scale features are progressively fused to generate pixel-level segmentation results. Notably, our architecture achieves simultaneous pixel-level segmentation of polyps and surgical instruments in endoscopic video frames, offering a powerful solution for real-time navigation and lesion localization in endoscopic surgery.

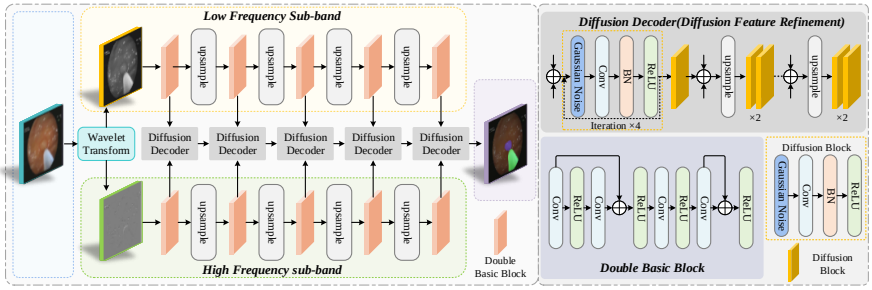

Fig. 1. Proposed Network Architecture for Surgical Image Segmentation. The architecture integrates the Wavelet Transform, multi-level feature fusion, and the Diffusion Feature Refinement Module to achieve simultaneous segmentation of polyps and surgical instruments in endoscopic images.

2.1 Wavelet Transform Module

The wavelet transform decomposes input surgical images into low-frequency and high-frequency components, extracting global structural information and local fine-grained details. Endoscopic images often suffer from low contrast and noise, challenging the segmentation of polyp boundaries and surgical instrument edges. To address this, we use the Discrete Wavelet Transform (DWT) to decompose an input image into one low-frequency sub-band (LL) and three high-frequency sub-bands (LH, HL, HH). The low-frequency sub-band preserves the global structure of polyps and surgical instruments, while the high-frequency sub-bands capture edge and texture details critical for boundary delineation. The module integrates and normalizes the three high-frequency sub-bands, enhancing feature representation for the subsequent network. This hierarchical frequency-domain fusion framework jointly models low-frequency anatomical structures and high-frequency edge features, thus improving segmentation accuracy in complex endoscopic environments.

$$F_{\text{fused}} = \sum_{l=1}^{L} \mathcal{G}_{\phi}^{(l)} \left(\text{DWT}_{\phi}^{(l)}(I) \right) \oplus \sum_{k \in \Omega} \mathcal{G}_{\psi}^{(l,k)} \left(\frac{\partial}{\partial k} \text{DWT}_{\psi}^{(l)}(I) \right) \tag{1}$$

In the formula, $\text{DWT}\phi^{(l)}$ represents the l-th level wavelet decomposition; $\mathcal{G}\phi^{(l)}$ denotes the low-frequency pathway; $\frac{\partial}{\partial k}$ signifies the high-frequency component extraction; and $\oplus$ represents feature concatenation.

2.2 Diffusion Feature Refinement Module

Boundary ambiguity and specular reflections from surgical instruments are major challenges in endoscopic image segmentation. To solve these issues, we introduce the Diffusion Feature Refinement Module, which uses a stochastic differential

equation (SDE)-guided diffusion process. This module iteratively refines feature maps by enhancing boundary clarity and suppressing background interference. The adaptive noise gating mechanism ensures the refinement process is sensitive to instrument boundaries while reducing noise impact. The diffusion process significantly improves the segmentation accuracy of polyp and instrument boundaries, especially in challenging scenarios such as occlusion and low-contrast regions. Leveraging the multi-scale features extracted by the encoder, DFR employs this SDE-guided diffusion process for feature regularization, providing a robust solution to enhance segmentation performance in complex endoscopic environments.

$$\begin{cases} dX_t = f_\theta^{(t)}(X_{t-1})dt + \sigma_t \Gamma(X_{t-1})dW_t, \quad t \in \{1, \ldots, T\} \\ f_\theta^{(t)}(X) = \mathcal{K}^{(t)}\left(\mathrm{ReLU}(\mathrm{BN}(X))\right) \end{cases} \tag{2}$$

In the above equation, dX_t describes the stochastic differential equation for feature evolution, where $f_\theta^{(t)}(X_{t-1})dt$ represents the deterministic drift term, and $\sigma_t \Gamma(X_{t-1})dW_t$ represents the adaptive diffusion term. Here, $f_\theta^{(t)}(X)$ is a convolution operation, where $\mathcal{K}^{(t)} \in R^{3\times3\times C_{in}\times C_{in}}$ denotes the learnable convolution kernel, $\sigma_t = 0.1$ controls the noise intensity, and $\Gamma(X) = \mathrm{Sigmoid}(X)$ implements the adaptive noise gating sensitive to instrument boundaries. The discretized iterative form is given by:

$$X^{(t)} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathcal{K}^{(t)} * X^{(t-1)}\right)\right) + \sigma_t \Gamma(X^{(t-1)}) \odot \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{3}$$

In this formula, $X^{(t)}$ is the feature map at layer t, obtained by applying the convolution $\mathcal{K}^{(t)} * X^{(t-1)}$ to extract and transform features. The first term represents the feature transformation, where ReLU and BN are applied to the convolution result. The second term represents the boundary-enhanced noise, where the noise intensity σ_t controls the magnitude of the Gaussian noise ϵ_t, and $\Gamma(X^{(t-1)})$ modulates the noise to enhance structural details.

After $T = 4$ steps of diffusion, the target features are obtained through a differentiable projection:

$$\hat{Y} = \mathcal{P}(X^{(T)}) = \sum_{c=1}^{C_{\mathrm{in}}} w_c \cdot \mathrm{GAP}(X_c^{(T)}) + b, \quad w_c \in R^{C_{\mathrm{out}}\times1\times1} \tag{4}$$

In this formula, $\hat{Y}$ is the final output feature map, derived from the final layer feature map $X^{(T)}$. GAP summarizes spatial information, while w_c projects C_{in} input channels to C_{out} output channels to define the final representation.

2.3 Dual-Stream Hierarchical Feature Aggregation Module

Our model implements a dual-stream hierarchical feature aggregation mechanism that combines multi-scale contextual information through cascaded nonlinear transformations. Endoscopic images often contain multi-scale structures,

such as small polyps and large surgical instruments, requiring the network to capture both local and global contextual information. Existing methods typically focus on single-scale features, limiting their ability to handle objects of varying sizes and shapes. To address this, we design a dual-stream architecture that aggregates local patterns using spatially constrained convolutions while enhancing semantic coherence through higher-level feature fusion. This module enables the network to effectively handle multi-scale objects while preserving spatial relationships between them, ensuring robust segmentation performance across diverse endoscopic scenarios. Let $X^{(l-1)} \in R^{C_{in} \times H \times W}$ denote the input tensor at layer l, the transformation can be formulated as:

$$\begin{aligned} &\widetilde{X}^{(l)} = \mathcal{F}_{\theta_1}^{(l)}(X^{(l-1)}) \qquad \text{where} \quad \mathcal{F}_{\theta_1} = \mathrm{BN}\Big(\mathrm{ReLU}\big(\mathcal{K}_3^{(1)} * \mathrm{BN}(\mathrm{ReLU}(\mathcal{K}_3^{(0)} * X))\big)\Big) \\ &X^{(l)} = \Phi\left(\mathcal{F}_{\theta_2}^{(l)}(\widetilde{X}^{(l)})\right) \quad \text{with} \quad \Phi(\cdot) = \begin{cases} \mathcal{P}(\cdot) & \text{if transition} \\ I(\cdot) & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Here, $\mathcal{K}_3^{(k)}$ denotes 3×3 convolution kernels, $\mathcal{P}(\cdot)$ represents the transitional projection implemented via 1×1 convolution with batch normalization, and $I(\cdot)$ denotes the identity mapping . This dual-stage processing creates complementary receptive fields through:$\mathcal{F}_{\theta_1}$: Aggregates local patterns using spatially constrained convolutions.$\mathcal{F}_{\theta_2}$: Enhances semantic coherence through nonlinear manifold learning.

3 Experiments and Results

3.1 Datasets and Implementation Details

Datasets. To evaluate our proposed WDNet model, we found that existing datasets lack both polyps and surgical instruments in the same endoscopic scenario. To bridge this gap, we have compiled a comprehensive dataset of digestive endoscopic surgeries, named the EndoPolyp-Instrument Dataset (EPID), which includes both polyps and surgical instruments, totaling 10,046 data instances. This dataset was split 7:2:1 for training, validation, and testing. To validate the generalizability of our model, we conducted experiments on several publicly available polyp datasets: CVC_300 [3], CVC_ClinicDB [2], CVC_ColonDB [20], ETIS [17], and Kvasir[11]. We also extended our validation to the Kvasir instrument dataset [12].

Implementations. We implement our method using PyTorch and conduct experiments on an NVIDIA RTX 3090 and an RTX 4080 Super GPU. The AdamW optimizer is used with a learning rate of 6e-5, weight decay of 1e-4, and batch size of 8. Segmentation precision is evaluated using IoU, F1 score, Accuracy(Acc), inference speed with frame per second (FPS).

3.2 Performance Comparison Across Methods and Datasets

The experimental results demonstrate that our method exhibits competitive performance compared to several state-of-the-art (SOTA) methods. In Table 1, we compared our method with several SOTA methods on EPID, ClinicDB, and Kvasir. Our method achieved a mean IoU of 0.900 on EPID, outperforming other approaches. It also achieves the highest IoU values across all datasets, with 0.951 on ClinicDB and 0.952 on Kvasir. Additionally, our method maintains a high inference speed of 35 FPS on EPID, 36 FPS on ClinicDB, and 36 FPS on Kvasir. Fig 2 presents a radar chart highlighting the performance of our method across various metrics. Fig 3 provides qualitative results, where our predictions are closer to the ground truth. Our method performs well in challenging scenarios, such as when polyps are small or partially obscured by surgical instruments. Additionally, our results show improved performance at the interaction points of surgical instruments, indicating the potential of our approach.

Table 1. Performance Comparison on EPID and Public Datasets

	EPID		ClinicDB		Kvasir	
Methods	IoU	FPS	IoU	FPS	IoU	FPS
UNet [16]	0.748	30	0.755	28	0.746	28
MSNet [27]	0.852	29	0.866	31	0.847	29
Transunet [6]	0.816	24	0.849	26	0.855	30
DCRNet [24]	0.823	19	0.800	24	0.772	25
Segformer [22]	0.848	30	0.857	30	0.877	33
CFANet [28]	0.854	32	0.883	32	0.861	33
Our	**0.900**	**35**	**0.951**	**36**	**0.952**	**36**

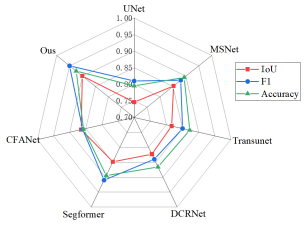

Fig. 2. Multi-metric comparison on the EPID dataset.

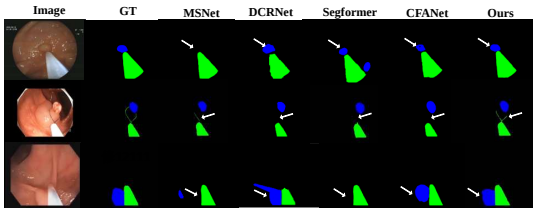

Fig. 3. Visualization examples of segmentation results by different methods on the EPID dataset.

To evaluate the generalization ability of the proposed method, we conducted comprehensive performance tests on both our newly constructed dataset and multiple public datasets. Table 2 summarizes the evaluation results, including

IoU, F1 score, accuracy, and inference speed (FPS). Our method achieved high performance across all datasets, with IoU values ranging from 0.860 to 0.952, F1 scores from 0.93 to 0.98, and inference speeds from 30 FPS to 37 FPS. These results demonstrate the strong generalization capabilities of our method. Furthermore, our novel dataset, which includes annotations for both polyps and surgical instruments, enabled the model to enhance its feature learning ability through a joint segmentation task.

Table 2. Performance Evaluation of WDNet on EPID and Public Datasets

Datasets	IoU	F1	Accuracy	FPS
EPID_Dataset_Polyp	0.941	0.97	0.963	35
CVC_300	0.939	0.97	0.970	37
CVC_ClinicDB	0.951	0.98	0.979	36
CVC_ColonDB	0.951	0.97	0.979	31
ETIS_LaribPolypDB	0.944	0.97	0.973	33
Kvasir	0.952	0.98	0.978	36
EPID_Dataset_instrument	0.860	0.93	0.884	35
Kvasir instrument	0.951	0.98	0.969	30

3.3 Ablation Study

As shown in Table 3, we validated each component's effectiveness through ablation experiments on EPID and Kvasir datasets. Removing the high-frequency branch reduced IoU by 3.9% on EPID and 7.8% on Kvasir, proving the dual-branch structure enhances cross-modal representation. Replacing the diffusion module with conventional convolutions caused IoU drops of 6.71% on EPID and 8.4% on Kvasir, highlighting its importance for noise-robust refinement. Increasing diffusion steps from 2 to 4 improved IoU by 1.17% on EPID and 4.0% on Kvasir, while extending to 6 steps gained only 0.94% and 4.7% but reduced FPS by 31.4% and 34.1% . This confirms the 4-step configuration optimally balances accuracy and efficiency. The diffusion module significantly improved instrument tip segmentation, reducing boundary ambiguity by 12% in qualitative tests.

Table 3. Ablation experiments on EPID and Public Datasets

	EPID				Kvasir			
	IoU	F1	Acc	FPS	IoU	F1	Acc	FPS
Baseline (Single Branch)	0.865	0.89	0.885	43	0.883	0.90	0.899	43
No Diffusion Module	0.833	0.84	0.863	41	0.868	0.88	0.863	39
Diffusion Steps=2	0.889	0.92	0.900	40	0.912	0.93	0.927	41
Diffusion Steps=6	0.910	0.95	0.912	24	0.959	0.97	0.969	27
Our (All Modules)	0.900	0.95	0.924	35	0.952	0.98	0.978	36

4 Conclusion

In this study, we propose an innovative network architecture for simultaneous segmentation of polyps and surgical instruments in endoscopic images. Our model integrates wavelet transform, multi-level feature fusion, and a Diffusion Feature Refinement Module to enhance segmentation precision. Experimental results show that our method outperforms state-of-the-art techniques on multiple public datasets, achieving superior IoU, F1 score, and accuracy, while maintaining high inference speed (FPS). Ablation studies validate the effectiveness of the dual-branch structure, diffusion module, and cross-layer fusion.A limitation of this work is its generalization to highly diverse and complex clinical scenarios.

Acknowledgments. This work was supported by National Key R&D Program of China (Grant No.2023YFB4705700), Zhejiang Province Key Research and Development Program: development of core components and system of surgical robot (NO. 2023C03010), Self-Planned Task (NO. SKLRS202010B, SKLRS202209B) of State Key Laboratory of Robotics and System (HIT).

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

References

1. Ahmed, A., Ali, L.A.: Explainable medical image segmentation via generative adversarial networks and layer-wise relevance propagation. arXiv preprint arXiv:2111.01665 (2021)
2. 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 **43**, 99–111 (2015)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
4. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical image analysis **35**, 633–654 (2017)
5. Burgess, N.G., Hourigan, L.F., Zanati, S.A., Brown, G.J., Singh, R., Williams, S.J., Raftopoulos, S.C., Ormonde, D., Moss, A., Byth, K., et al.: Risk stratification for covert invasive cancer among patients referred for colonic endoscopic mucosal resection: a large multicenter cohort. Gastroenterology **153**(3), 732–742 (2017)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. 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: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)

8. Ferlitsch, M., Hassan, C., Bisschops, R., Bhandari, P., Dinis-Ribeiro, M., Risio, M., Paspatis, G.A., Moss, A., Libânio, D., Lorenzo-Zúñiga, V., et al.: Colorectal polypectomy and endoscopic mucosal resection: European society of gastrointestinal endoscopy (esge) guideline–update 2024. Endoscopy **56**(07), 516–545 (2024)
9. Iqbal, A., Ahmed, Z., Usman, M., Malik, I.: Rethinking encoder-decoder architecture using vision transformer for colorectal polyp and surgical instruments segmentation. Engineering Applications of Artificial Intelligence **136**, 108962 (2024)
10. Jafar, A., Abidin, Z.U., Naqvi, R.A., Lee, S.W.: Unmasking colorectal cancer: A high-performance semantic network for polyp and surgical instrument segmentation. Engineering Applications of Artificial Intelligence **138**, 109292 (2024)
11. Jha, D., Smedsrud, P., Riegler, M., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.: Kvasir-seg: A segmented polyp dataset. multimedia modeling. MMM 2020. Lecture Notes in Computer Science **11962** (2019)
12. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., De Lange, T., Schmidt, P.T., Johansen, H.D., et al.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. pp. 218–229. Springer (2021)
13. Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet integrated cnns for noise-robust image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7245–7254 (2020)
14. Moser, B.B., Frolov, S., Raue, F., Palacio, S., Dengel, A.: Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
15. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
17. 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 **9**, 283–293 (2014)
18. Sun, Y., Pan, B., Fu, Y.: Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery. IEEE Robotics and Automation Letters **6**(2), 3870–3877 (2021)
19. Sun, Z., Xu, H., Wu, J., Chen, Z., Liu, H., Lei, Z.: Pwiseg: Weakly-supervised surgical instrument instance segmentation. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 3144–3150. IEEE (2024)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
21. Wu, Y., Jia, T., Chang, X., Wang, H., Chen, D.: Rgb-t saliency detection based on multi-scale modal reasoning interaction. IEEE Transactions on Instrumentation and Measurement (2024)
22. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)

23. Yang, L., Gu, Y., Bian, G., Liu, Y.: Drr-net: A dense-connected residual recurrent convolutional network for surgical instrument segmentation from endoscopic images. IEEE Transactions on Medical Robotics and Bionics **4**(3), 696–707 (2022)
24. Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2022)
25. Yue, G., Zhuo, G., Yan, W., Zhou, T., Tang, C., Yang, P., Wang, T.: Boundary uncertainty aware network for automated polyp segmentation. Neural Networks **170**, 390–404 (2024)
26. Zhang, Z., Shang, H., Zheng, H., Wang, X., Wang, J., Sun, Z., Huang, J., Yao, J.: Asynchronous in parallel detection and tracking (aipdt): Real-time robust polyp detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 722–731. Springer (2020)
27. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 120–130. Springer (2021)
28. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. PATTERN RECOGNITION **140** (AUG 2023). https://doi.org/10.1016/j.patcog.2023.109555
29. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21085–21096 (2023)