Probabilistic Modeling Ensemble Vision Transformer Improves Complex Polyp Segmentation

Tianyi Ling[1,2], Chengyi Wu[2,3], Huan Yu[2,4], Tian Cai[5], Da Wang[1], Yincong Zhou[2], Ming Chen[2(✉)], and Kefeng Ding[1(✉)]

[1] Department of Colorectal Surgery and Oncology (Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education, Key Laboratory of Molecular Biology in Medical Sciences, Zhejiang Province, China), The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
dingkefeng@zju.edu.cn

[2] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, China
mchen@zju.edu.cn

[3] Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China

[4] Department of Thoracic Surgery, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China

[5] Department of Hepatobiliary and Pancreatic Surgery, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

Abstract. Colorectal polyps detected during colonoscopy are strongly associated with colorectal cancer, making polyp segmentation a critical clinical decision-making tool for diagnosis and treatment planning. However, accurate polyp segmentation remains a challenging task, particularly in cases involving diminutive polyps and other intestinal substances that produce a high false-positive rate. Previous polyp segmentation networks based on supervised binary masks may have lacked global semantic perception of polyps, resulting in a loss of capture and discrimination capability for polyps in complex scenarios. To address this issue, we propose a novel Gaussian-Probabilistic guided semantic fusion method that progressively fuses the probability information of polyp positions with the decoder supervised by binary masks. Our **P**robabilistic Modeling **E**nsemble Vision **T**ransformer **Net**work(***PETNet***) effectively suppresses noise in features and significantly improves expressive capabilities at both pixel and instance levels, using just simple types of convolutional decoders. Extensive experiments on five widely adopted datasets show that *PETNet* **outperforms** existing methods in identifying polyp

T. Ling and C. Wu—Equal contributions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_54.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 572–581, 2023.
https://doi.org/10.1007/978-3-031-43990-2_54

camouflage, appearance changes, and small polyp scenes, and achieves a speed about **27FPS** in edge computing devices. Codes are available at: https://github.com/Seasonsling/PETNet.

Keywords: Colonoscopy · Polyp Segmentation · Vision Transformer

1 Introduction

Colorectal cancer (CRC) remains a major health burden with elevated mortality worldwide [1]. Most cases of CRC arise from adenomatous polyps or sessile serrated lesions in 5 to 10 years [9]. Colonoscopy is considered the gold standard for the detection of colorectal polyps. Polyp segmentation is a fundamental task in the computer-aided detection (CADe) of polyps during colonoscopy, which is of great significance in the clinical prevention of CRC.

Traditional machine learning approaches in polyp segmentation primarily focus on learning low-level features, such as texture, shape, or color distribution [14]. In recent years, encoder-decoder based deep learning models such as U-Net [12], UNet++ [22], ResUNet++ [6], and PraNet [4] have dominated the field. Furthermore, Transformer [3,17,19,20] models have also been proposed for polyp segmentation, and achieve the state-of-the-art(SOTA) performance.

Despite significant progress made by these binary mask supervised models, challenges remain in accurately locating polyps, particularly in complex clinical scenarios, due to their insensitivity to complex lesions and high false-positive rates. More specifically, most polyps have an elliptical shape with well-defined boundaries. However, supervised segmentation learning solely based on binary masks may not be effective in discriminating polyps in complex clinical scenarios. Endoscopic images often contain pseudo-polyp objects with strong boundaries, such as colon folds, blood vessels, and air bubbles, which can result in false positives. In addition, sessile and flat polyps have ambiguous and challenging boundaries to delineate. To address these limitations, Qadir et al. [11] proposed using Gaussian masks for supervised model training. This approach reduces false positives significantly by assigning less attention to outer edges and prioritizing surface patterns. However, this method has limitations in accurately segmenting polyp boundaries, which are crucial for clinical decision-making.

Therefore, the primary challenge lies in enhancing polyp segmentation performance in complex scenarios by precisely preserving the polyp segmentation boundaries, while simultaneously maximizing the decoder's attention on the overall pattern of the polyps.

In this paper, we propose a novel transformer-based polyp segmentation framework, *PETNet*, which addresses the aforementioned challenges and achieves SOTA performance in locating polyps with high precision. Our contributions are threefold:

- We propose a novel Gaussian-Probabilistic guided semantic fusion method for polyp segmentation, which improves the decoder's global perception of polyp locations and discrimination capability for polyps in **complex scenarios**.

- We evaluate the performance of *PETNet* on five widely adopted datasets, demonstrating its superior ability to identify polyp camouflage and small polyp scenes, achieving **state-of-the-art** performance in locating polyps with high precision. Furthermore, we show that *PETNet* can achieve a speed of about **27FPS** in edge computing devices (Nvidia Jetson Orin).
- We design several polyp instance-level evaluation metrics, considering that conventional pixel-level calculation methods cannot explicitly and comprehensively evaluate the overall performance of polyp segmentation algorithms.

2 Methods

2.1 Architecture Overview

As shown in Fig. 1, *PETNet* is an end-to-end polyp segmentation framework consists of three core module groups. (1) The **Encoder Group** employs a vision transformer backbone [18] cascaded with a mixed transformer attention layer to encode long-range dependent features at four scales. (2) The **Gaussian-Probabilistic Modeling Group** consists of a Gaussian Probabilistic Guided UNet-like decoder branch(GUDB) and Gaussian Probabilistic-Induced transition(GIT) modules. (3) The **Ensemble Binary Decoders Group** includes a UNet-like structure branch(UDB) [12], a fusion module(Fus), and a cascaded fusion module(CFM) [3].

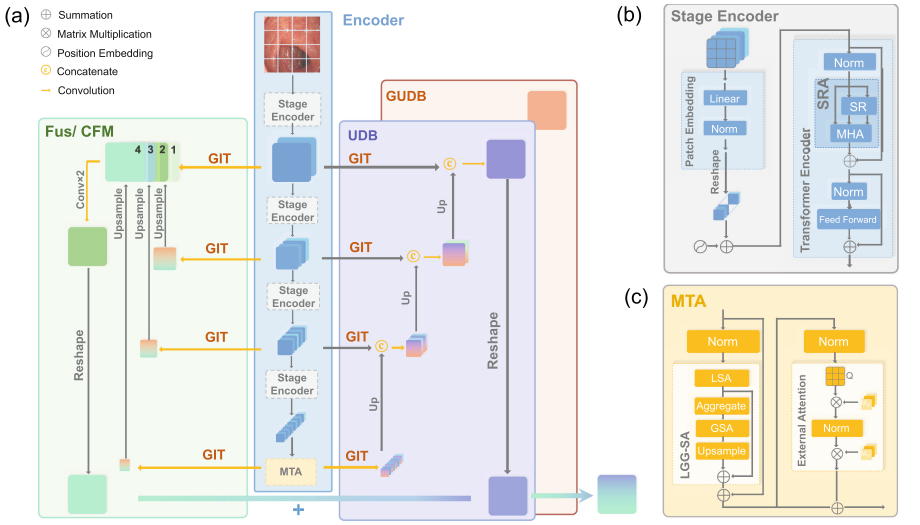

Fig. 1. Proposed PETNet Framework: (a) Comprises three critical module groups. (b) Depicts the Stage Encoder. (c) Illustrates the Mixed Transformer Attention layer (MTA).

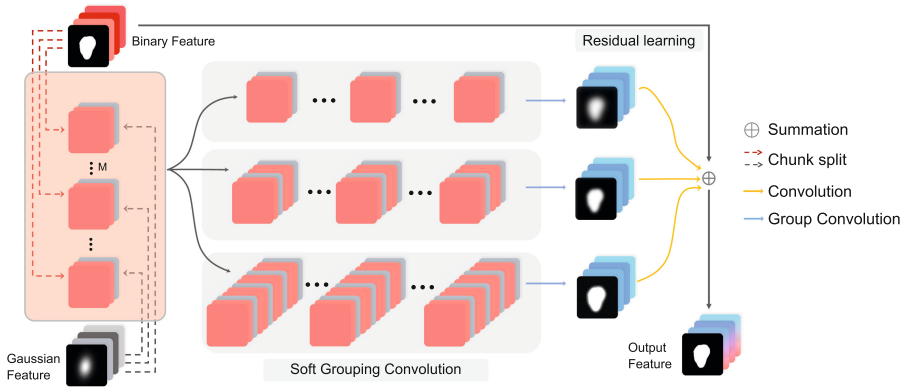

Fig. 2. Illustration of Gaussian Probabilistic-induced transition (GIT).

2.2 Encoder Group

To balance the trade-off between computational speed and feature representation capability, we utilize the pre-trained PVTv2-B2 model [18] as the backbone. Mixed transformer attention(MTA) layer is composed of Local-Global Gaussian-Weighted Self-Attention (LGG-SA) and External Attention (EA). We add a MTA layer to encode the last level features, enhancing the model's semantic representation and accelerating the training process [16]. Moreover, the encoder output features are presented as $\{\mathbf{X}_i^E\}_{i=1}^4$ with channels of $[2C, 4C, 8C, 16C]$.

2.3 Gaussian-Probabilistic Modeling Group

To incorporate both polyp location probability and surface pattern information in a progressive manner, we propose the Gaussian Probabilistic-induced Transition (GIT) method. This method involves the interaction between a Gaussian auxiliary decoder and multiple binary decoders in a layer-wise fashion, as shown in Fig. 2.

Gaussian Probabilistic Mask. Inspired by [11] and [21], in addition to utilizing binary representation, polyps can also be represented as probability heatmaps with blurred borders. We present a method of converting the binary polyp mask $f(x, y) \in \{0, 1\}^{W \times H \times 1}$ into Gaussian masks $Y(x, y) \in [0, 1]^{W \times H \times 1}$ by utilizing elliptical Gaussian kernels. Specifically, for every polyp in a binary mask, after masking other polyp pixels as background, we calculate

$$Y = \exp\left(-\left(a(x - x_o)^2 + 2b(x - x_o)(y - y_o) + (y - y_o)^2\right)\right) \tag{1}$$

where (x_o, y_o) is the mass of each polyp in the binary image $f(x, y)$. To rotate the output 2D Gaussian masks according to the orientation, we set a, b, c as followings,

$$a = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2}, \tag{2}$$

$$b = \frac{-sin(2\theta)}{4\sigma_x^2} + \frac{sin(2\theta)}{4\sigma_y^2}, \tag{3}$$

$$c = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2}, \tag{4}$$

where σ_x^2 and σ_y^2 are the polyp size-adaptive standard deviations [21], and θ is the orientation of each polyp [11]. Finally, we determine the final Gaussian probabilistic mask $\mathcal{P}_\mathcal{G}$ for all polyps within an image mask by computing the element-wise maximum.

Gaussian Guided UNet-Like Decoder Branch. The Gaussian Guided UNet-like decoder branch(GUDB) module is a simple UNet-like decoding branch supervised by Gaussian Probabilistic masks. We employ four levels of encoder output features, and adjust encoder features $\{\mathbf{X}_i^E\}_{i=1}^4$ to the features $\{\mathbf{X}_i^G\}_{i=1}^4$ with channels of $[C, 2C, 2C, 2C]$ in each level. At the final layer, a 1×1 convolution is used to convert the feature vector to one channel, producing a size of $\mathcal{H} \times \mathcal{W} \times 1$ Gaussian mask.

Gaussian Probabilistic-Induced Transition Module. We use the Gaussian probabilistic-induced transition module(GIT) to achieve transition between binary features and gaussian features. Given the features originally sent to the Decoder as binary features $\{\mathbf{X}_i^B\}_{i=1}^4$, and the transformed encoder features sent to GUDB as $\mathbf{X}^G$. We first splits 4 levels of $\mathbf{X}^B$ and $\mathbf{X}^G$ into fixed groups as:

$$\{\mathbf{X}_{i,m}^D\}_{m=1}^M \in \mathbb{R}^{C_d/M \times H_i \times W_i} \leftarrow \mathbf{X}_i^D \in \mathbb{R}^{C_d \times H_i \times W_i}, \tag{5}$$

where M is the corresponding number of groups. Then, we periodically arrange groups of $\mathbf{X}_{i,m}^B$ and $\mathbf{X}_{i,m}^G$ for each level, and generate the regrouped feature $\mathbf{Q}_i \in \mathbb{R}^{(C_i+C_g) \times H_i \times W_i}$ in an Multi-layer sandwiches manner. Soft grouping convolution [7] is then applied to provide parallel nonlinear projections at multiple fine-grained sub-spaces (Fig. 2). We further introduce residual learning in a parallel manner at different group-aware scales. The final output $\{\mathbf{Z}_i^T\}_{i=1}^4 \in \mathbb{R}^{C_i \times H_i \times W_i}$ is obtained for the UDB decoder. Considering the computation cost, The binary features $\mathbf{X}^B \leftarrow \mathbf{X}^E$ for the Fus decoder have channel numbers of $[4C, 4C, 4C, 4C]$. The Fus decoder and CFM share identical transited output features, while CFM exclusively utilizes the last three levels of features.

2.4 Ensemble Binary Decoders Group

During colonoscopy, endoscopists often use the two-physician observation approach to improve the detection rate of polyps. Building on this manner, we propose the ensemble method that integrates multiple simple decoders to enhance

the detection and discrimination of difficult polyp samples. We demonstrate the effectiveness of our approach using three commonly used convolutional decoders. After GIT process, diverse level of Gaussian probabilistic-induced binary features were sent to these decoders. The output mask $\mathcal{P}$ is obtained by element-wise summation of $\mathcal{P}_i$, where i represents the binary decoder index.

Fusion Module. As shown in Fig. 1, Set $\mathcal{X}i$, $i\in(1,2,3,4)$ represent multi-scale mixed features. Twice convolution following with bilinear interpolation are applied to transform these feature with same 4C channels as $\mathcal{X}_1{}', \mathcal{X}_2{}', \mathcal{X}_3{}', \mathcal{X}_4{}'$. Afterward, we get $\mathcal{X}_{out}$ with the resolution of $\mathcal{H}/4 \times \mathcal{W}/4 \times \mathcal{C}1$ through following formula, where $\mathcal{F}$ represents twice 3×3 convolution:

$$\mathcal{X}_{out} = \mathcal{F}(Concat(\mathcal{X}_1{}', \mathcal{X}_2{}', \mathcal{X}_3{}', \mathcal{X}_4{}') \tag{6}$$

UNet Decoder Branch and CFM Module. The structure of the UDB is similar to that of the GUDB, except for the absence of channel reduction prior to decoding. In our evaluation, we also examine the decoder CFM utilized in [3], which shares the same input features (excluding the first level) as the Fus.

3 Experiments

3.1 Datasets Settings

To evaluate models fairly, we completely follow *PraNet* [4] and use five public datasets, including 548 and 900 images from ClinicDB [2] and Kvasir-SEG [5] as training sets, and the remaining images as validation sets. We also test the generalization capability of all models on three unseen datasets (ETIS [13] with 196 images, CVC-ColonDB [8] with 380 images, and EndoScene [15] with 60 images). Training settings are the same as [3].

3.2 Loss Setting

Our loss function formulates as $\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_i + \lambda \mathcal{L}_g$, and

$$\mathcal{L}_i = \sum_{i=1}^{n} (\mathcal{L}_{IOU}(\mathcal{P}_i, \mathcal{G}_B) + \mathcal{L}_{BCE}(\mathcal{P}_i, \mathcal{G}_B)) \tag{7}$$

where N is the total number of binary decoders, $\mathcal{L}_g$ represents the L1 loss between the ground truth Gaussian mask $\mathcal{G}_G$ and GUDB prediction mask $\mathcal{P}_G$. λ is a hyperparameter used to balance the binary and Gaussian losses. Furthermore, we employ intermediate decoder outputs to calculate auxiliary losses for convergence acceleration.

3.3 Evaluation Metrics

Conventional evaluation metrics for polyp segmentation are typically limited to pixel-level calculations. However, metrics that consider the entire polyp are also crucial. Here we assess our model from both **pixel-level** and **instance-level** perspectives.

Pixel-level evaluation is based on mean intersection over union ($mIoU$), mean Dice coefficient ($mDic$), and weighted F_1 score (wF_m). For polyp instance evaluation, a true positive (TP) is defined when the detection centroid is located within the polyp mask. False positives (FPs) occur when a wrong detection output is provided for a negative region, and false negatives (FNs) occur when a polyp is missed in a positive image. Finally, we compute sensitivity $n\mathcal{S}en = TP/(TP + FN) \times 100$, precision $n\mathcal{P}re = TP/(TP + FP) \times 100$, and $n\mathcal{F}_1 = 2 \times (\mathcal{S}en \times \mathcal{P}re)/(\mathcal{S}en + \mathcal{P}re) \times 100$ based on the number count for instance evaluation.

3.4 Results

Training and Learning Ability. Table S1 displays the results of our model's training and learning performance. Our model achieves comparable performance to the SOTA model on the Kvasir-SEG and ClinicDB datasets. Notably, our model yields superior results in false-positive instance evaluation.

Generalization Ability. The generalization results are shown in Table 1. We conduct three unseen datasets to test models' generalizability. Results show that $PETNet$ achieves excellent generalization performance compared with previous models. Most importantly, our false-positive instance counts(45 in ETIS and 55 in CVC-ColonDB) reduce significantly of other models. We also observe a performance mismatch phenomenon in pixel-level evaluation and instance-level evaluation.

Small Polyp Detection Ability. The detection capability results of small polyps are shown in Table 2. Diminutive polyps are hard to precisely detect, while they are the major targets of optical biopsies performed by endoscopists. We selected images from two unseen datasets with 0∼2% polyp labeled area to perform the test. As shown, $PETNet$ demonstrates great strength in both datasets, which indicates that one of the major advantages of our model lies in detecting small polyps with lower false-positive rates.

Ablation Analysis. Table 3 presents the results of our ablation study, where we investigate the contribution of the two key components of our model, namely the Gaussian-Probabilistic Guided Semantic Fusion method and ensemble decoders. We observe that while the impact of each binary decoder varies, all sub binary decoders contribute to the overall performance. Furthermore, the GIT method significantly enhances instance-level evaluation without incurring performance penalty in pixel-level evaluation, especially in unseen datasets.

Table 1. Quantitative results of the test datasets EndoScene, CVC-ColonDB and ETIS-LaribPolypDB.

	Methods	$mDic$	$mIoU$	$nSen$	$nPre$	nF_1	TP	FP	FN
EndoScene	PraNet (MICCAI'20) [4]	0.748	0.634	**1.000**	0.526	0.690	60	54	**0**
	EU-Net (CRV'21) [10]	0.882	0.809	**1.000**	0.938	0.968	60	4	**0**
	LDNet (MICCAI'22) [19]	0.888	0.814	**1.000**	0.984	0.992	60	1	**0**
	SSFormer-S (MICCAI'22) [17]	0.881	0.812	**1.000**	0.938	0.968	60	4	**0**
	Polyp-PVT(CAAI AIR'23) [3]	0.894	0.829	**1.000**	0.968	0.984	60	2	**0**
	***PETNet* (Ours)**	**0.899**	**0.834**	**1.000**	**1.000**	**1.000**	**60**	**0**	**0**
ColonDB	PraNet (MICCAI'20) [4]	0.643	0.542	0.882	0.499	0.638	351	352	47
	EU-Net (CRV'21) [10]	0.757	0.672	0.889	0.760	0.819	354	112	44
	LDNet (MICCAI'22) [19]	0.751	0.667	0.905	0.726	0.805	360	136	38
	SSFormer-S (MICCAI'22) [17]	0.787	0.708	0.900	0.840	0.869	358	68	40
	Polyp-PVT(CAAI AIR'23) [3]	0.808	0.724	**0.937**	0.772	0.847	**373**	110	**25**
	***PETNet* (Ours)**	**0.817**	**0.740**	0.935	**0.871**	**0.902**	372	**55**	26
ETIS	PraNet (MICCAI'20) [4]	0.584	0.482	0.904	0.390	0.545	188	294	20
	EU-Net (CRV'21) [10]	0.687	0.610	0.870	0.557	0.679	181	144	27
	LDNet (MICCAI'22) [19]	0.702	0.611	0.909	0.659	0.764	189	98	19
	SSFormer-S (MICCAI'22) [17]	0.744	0.672	0.856	0.674	0.754	178	86	30
	Polyp-PVT(CAAI AIR'23) [3]	0.765	0.687	**0.923**	0.667	0.774	**192**	96	**16**
	***PETNet* (Ours)**	**0.782**	**0.703**	0.904	**0.807**	**0.853**	188	**45**	20

Table 2. Quantitative results of the **small polyp detection** in ETIS and CVC-ColonDB dataset. Small polyps are defined as the polyp area accounts for 0∼2% of the entire image.

	Methods	$mDic$	$mIoU$	$nSen$	$nPre$	nF_1	TP	FP	FN
ETIS	PraNet [4]	0.43	0.34	0.87	0.34	0.49	87	167	13
	Polyp-PVT [3]	0.68	0.60	**0.93**	0.61	0.74	**93**	60	**7**
	***PETNet* (Ours)**	**0.69**	**0.60**	**0.88**	**0.73**	**0.80**	**88**	**33**	**12**
ColonDB	PraNet [4]	0.45	0.34	0.82	0.40	0.54	80	120	17
	Polyp-PVT [3]	**0.68**	**0.58**	**0.93**	0.71	0.80	**90**	37	**7**
	***PETNet* (Ours)**	0.67	**0.58**	**0.93**	**0.76**	**0.84**	**90**	**28**	**7**

Table 3. Ablation study for *PETNet* on five datasets. wF_m: pixel-based weighted F1 score, nF_1: instance-based weighted F1 score. w/o: without.

Dataset	Metric	Baseline	w/o Fus	w/o UDB	w/o CFM	w/o GUDB	PETNet
ClinicDB	wF_m	0.819	0.929	0.929	**0.941**	0.933	0.932
	nF_1	0.883	**0.978**	0.964	0.971	0.971	0.964
Kvasir	wF_m	0.804	0.904	**0.914**	0.899	0.911	0.912
	nF_1	0.853	0.882	0.879	**0.895**	0.881	0.891
EndoScene	wF_m	0.705	0.859	0.877	0.876	0.874	**0.884**
	nF_1	0.851	0.992	0.952	0.976	0.976	**1.000**
ColonDB	wF_m	0.635	0.800	0.783	0.775	**0.812**	0.802
	nF_1	0.761	0.898	0.894	0.881	0.882	**0.902**
ETIS	wF_m	0.467	0.738	0.744	0.725	0.736	**0.747**
	nF_1	0.577	0.840	0.810	0.805	0.850	**0.853**

3.5 Comparative Analysis

Fig.S1 shows that our proposed model, *PETNet*, outperforms SOTA models in accurately identifying polyps under complex scenarios, including lighting disturbances, water reflections, and motion blur. Faluire cases are shown in Fig.S2.

3.6 Running in the Real World

Furthermore, we deployed *PETNet* on the edge computing device Nvidia Jetson Orin and optimized its performance using TensorRT. Our results demonstrate that *PETNet* achieves real-time denoising and segmentation of polyps with high accuracy, achieving a speed of 27 frames per second on the device(Video S1).

4 Conclusion

Based on intrinsic characteristics of the endoscopic polyp image, we specifically propose a novel segmentation framework named ***PETNet*** consisting of three key module groups. Experiments show that *PETNet* consistently outperforms most current cutting-edge models on five challenging datasets, demonstrating its solid robustness in distinguishing other intestinal analogs. Most importantly, *PETNet* shows better sensitivity to complex lesions and diminutive polyps.

Acknowledgement. This work was supported by the National Natural Sciences Foundation of China (Nos. 31771477, 32070677), the Fundamental Research Funds for the Central Universities (No. 226-2022-00009), and the Key R&D Program of Zhejiang (No. 2023C03049).

References

1. Ahmed, A.M.A.A.: Generative adversarial networks for automatic polyp segmentation. arXiv:2012.06771 [cs, eess] (2020)
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. Comput. Med. Imaging Graph.: Official J. Comput. Med. Imaging Soc. **43**, 99–111 (2015). https://doi.org/10.1016/j.compmedimag.2015.02.007
3. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-PVT: polyp segmentation with pyramid vision transformers. arXiv:2108.06932 [cs] (2021)
4. Fan, D.P., et al.: PraNet: parallel reverse attention network for polyp segmentation. arXiv:2006.11392 [cs, eess] (2020)
5. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. arXiv:1911.07069 [cs, eess] (2019)
6. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. arXiv:1911.07067 [cs, eess] (2019)
7. Ji, G.P., Fan, D.P., Chou, Y.C., Dai, D., Liniger, A., Van Gool, L.: Deep gradient learning for efficient camouflaged object detection. Tech. Rep. arXiv:2205.12853, arXiv (2022)

8. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE Trans. Med. Imaging **33**(7), 1488–1502 (2014). https://doi.org/10.1109/TMI.2014.2314959, http://arxiv.org/abs/1305.1912
9. National Health Commission of the People's Republic of China: [Chinese Protocol of Diagnosis and Treatment of Colorectal Cancer (2020 edition)]. Zhonghua Wai Ke Za Zhi [Chinese Journal of Surgery] **58**(8), 561–585 (2020). https://doi.org/10.3760/cma.j.cn112139-20200518-00390
10. Patel, K., Bur, A.M., Wang, G.: Enhanced U-Net: a feature enhancement network for polyp segmentation. In Proceedings of the International Robots & Vision Conference. International Robots & Vision Conference **2021**, 181–188 (2021). https://doi.org/10.1109/crv52889.2021.00032, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8341462/
11. Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I.: Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction. Med. Image Anal. **68**, 101897 (2021). https://doi.org/10.1016/j.media.2020.101897, https://linkinghub.elsevier.com/retrieve/pii/S1361841520302619
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs] (2015), http://arxiv.org/abs/1505.04597
13. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2013). https://doi.org/10.1007/s11548-013-0926-3
14. Tajbakhsh, N., Gurudu, S.R., Liang, J.: A comprehensive computer-aided polyp detection system for colonoscopy videos. Inf. Process. Med. Imaging **24**, 327–38 (2015). https://doi.org/10.1007/978-3-319-19992-4_25
15. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. **2017**, 4037190 (2017). https://doi.org/10.1155/2017/4037190
16. Wang, H., et al.: Mixed transformer U-Net for medical image segmentation. arXiv:2111.04734 [cs, eess] (2021)
17. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: local guides global. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 110–120. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_11
18. Wang, W., et al.: PVTv2: improved baselines with pyramid vision transformer. arXiv:2106.13797 [cs] (2022)
19. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 99–109. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_10
20. Zhang, Y., Liu, H., Hu, Q.: TransFuse: fusing transformers and CNNs for medical image segmentation. arXiv:2102.08005 [cs] (2021)
21. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points (2019). https://doi.org/10.48550/arXiv.1904.07850
22. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. arXiv:1807.10165 [cs, eess, stat] (2018)

Clinical Applications – Ophthalmology