ESPNet: Edge-Aware Feature Shrinkage Pyramid for Polyp Segmentation

Raneem Toman[1] [0009−0008−1145−8152], Venkataraman Subramanian[2,3] [0000−0003−3603−0861], and Sharib Ali(✉)[1] [0000−0003−1313−3542]

[1] School of Computer Science, Faculty of Engineering and Physical Sciences, University of Leeds, LS2 9JT, Leeds, UK
[2] Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
[3] Leeds Teaching Hospitals NHS Trust, Leeds, UK
{scrmat,s.s.ali}@leeds.ac.uk

Abstract. Despite numerous techniques developed for polyp segmentation, the issue of generalizability to new centers and populations persists. To address these issues, we compile a multicenter train set consisting of 4,000 polyp frames and propose a novel approach toward generalizing to different data centers, difficult polyp morphologies (e.g., flat or small), and inflammatory conditions such as inflammatory bowel disease (IBD). In this regard, we propose a transformer-based polyp segmentation model to leverage global contextual information, and enhancement of local feature interactions through a novel feature decoding and fusion method, and polyp edge features. This combines the vision transformers' strong contextual understanding with enhanced locality modeling through graph-based relational understanding and multiscale feature aggregation. We compare our model with eight recent state-of-the-art methods under five widely used metrics on the following benchmark datasets: Kvasir-Sessile, SUN-SEG-Easy (Seen), ETIS-LaribPolypDB, CVC-ColonDB, PolypGen-C6, and our in-house IBD dataset. Extensive experiments show that our model outperforms state-of-the-art methods on out-of-distribution datasets with mIoU improvements of 2.84% on ETIS-LaribPolypDB, 1.26% on CVC-ColonDB, 1.90% on PolypGen-C6, and 3.52% on the in-house IBD polyp dataset compared to the most accurate recent method. The code is available at https://github.com/Raneem-MT/ESPNet.

Keywords: Polyp segmentation · Feature shrinkage · Edge-Aware Segmentation· Generalization

1 Introduction

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer-related deaths worldwide [14]. Also, there exist variations among clinicians in their ability to accurately detect and delineate these polyps, mainly due to their different levels of expertise, specialties, and the high variability in polyps' morphologies [17,27]. For example, inflammatory bowel disease

(IBD) has a worse prognosis and higher morbidity compared to the non-IBD-CRC. This is because IBD dysplasia often presents as flatter and poorly defined polyps blended with inflamed surrounding mucosa, making them harder to detect endoscopically [16]. Automated methods for early detection and segmentation of polyps during colonoscopy screening can significantly reduce the current incidences and help endoscopists remove polyps before they become life-threatening. However, there is a big gap in the existing state-of-the-art methods [6,25,5] as they are not rigorously assessed on difficult out-of-distribution scenarios important for complex polyp segmentation tasks. Hence, their adoption in routine clinical settings and across multiple centers remains questionable.

The need to alleviate operator dependency has made automated polyp segmentation an active area of research. CNN-based methods like FCN and UNet have laid the foundation for dense prediction tasks [13,15], but their local receptive fields have limited their capturing of long-range dependencies. PraNet tackled this by generating global maps and introducing a parallel reverse attention network that aggregates high-level features, but it was not robust to noise [6]. SANet used a shallow attention module to filter out background noise, but it did not effectively learn the features [22]. LDNet proposed dynamic kernel adaptive learning based on the input features to improve lesion detection, but it was prone to false positives [25]. CFANet applied cross-level feature aggregation through a boundary prediction module and two-stream segmentation module that fuses multi-scale features and boundary features [28]; however, it struggled with challenging scenes like large polyps and complex backgrounds. Transformer-based models, on the other hand, while strongly capable of modeling global context, suffer from the limitation of locality modeling. Polyp-PVT utilized a pyramid vision transformer (PVT), cascaded fusion, and similarity aggregation modules to effectively extract and fuse multi-scale features, but it still faced over-detection (a large number of false positives) in cases of shadows and light reflections [5]. TransNetR leveraged shortcut connections and residual transformer blocks to better capture long-range dependencies [10]; however, it still presented false positives. ASPS integrated both CNN and ViT features through a cross-branch feature augmentation module to improve feature representation, but it did not detect accurate boundaries [12]. MEGANet used an edge-guided attention (EGA) module that preserves boundary information at different scales, but it was limited by the inherent noise in the edge features, especially in complex background scenarios [4]. Despite the numerous techniques developed in the field of polyp segmentation, the issue of generalizability to different acquisition systems, centers, and patient populations persists. This is because supervised models are trained on publicly available datasets, which usually represent either a single center and one population cohort, making them tend to struggle when assessed on different data distributions [3,1].

To overcome these limitations, we propose a novel transformer-based edge-aware polyp segmentation method that leverages a feature shrinkage pyramid (FSPNet) [7], integrates residual connections for feature enhancement through scales, and mask-level (high-level) and edge-level (low-level) features with an attention mechanism for robust feature learning.

Our contributions can be summarized as follows: 1) We propose to include a Feature Shrinkage Decoder (FSD) network to learn fine-grained features by regressing the polyp edges. The learned edge features complement the mask predictions across four different scales. 2) We apply channel and spatial attention to the aggregated feature maps to enhance high-level information (e.g., shape) and low-level features (e.g., edges). 3) We introduce residual connections during feature shrinkage to preserve cues from earlier scales. 4) We ablate these experiments demonstrating improvements over the baseline FSPNet. 5) Finally, we assess the generalizability of methods in two contexts: i) on data from 3 data centers not seen during model training and ii) on an unseen center and different patient population dataset (IBD) compared to the trained model (non-IBD).

2 Method

2.1 Overview

Figure 1 outlines the architecture of our proposed ESPNet. The main components include a ViT encoder, a token-enhancement module (TEM), and mask and edge feature shrinkage decoders (FSDs). We also include feature fusion attention blocks in both TEM and FSD layers. Specifically, the input image is serialized into patch tokens that are fed into ViT's self-attention mechanism for global context modeling. The TEM is then applied to explore features within the tokens and the interactions between them through a graph fusion module (GFM). The enhanced features are then passed to the mask and edge FSDs, where each aggregates adjacent features while upscaling them until reaching the output. To retain information, residual connections are applied between scales. Finally, to accumulate the best object features from the masks and edges, an attention block is applied to the concatenated features.

2.2 Vision Transformer Encoder

The data-efficient image transformer (DeiT) [20] is kept as the encoder as they are less prone to overfitting. In DeiT, the input image $I \in R^{C\times H\times W}$ is first divided into non-overlapping patches of size $p\times p$ (16×16 in our case). Each patch is flattened into a vector and then linearly projected into a 1D sequence of token embeddings for the transformer input T^0. A learnable position embedding T^{pe} that refers to the location of each patch in the image is added to the embeddings. Finally, a distillation token T^d is added to allow learning from the teacher's predictions, so the final token becomes $T^f = T^0 + T^{pe} + T^d$. The tokens are then fed into the encoder's transformer layers, each containing a multi-head self-attention (MSA) and a multi-layer perceptron (MLP) block (MSA+MLP layers), giving a sequence of refined tokens as the output from the encoder, capturing long-range relations in the image:

$$T = \mathrm{MLP}(\mathrm{MSA}(T^f)) \qquad (1)$$

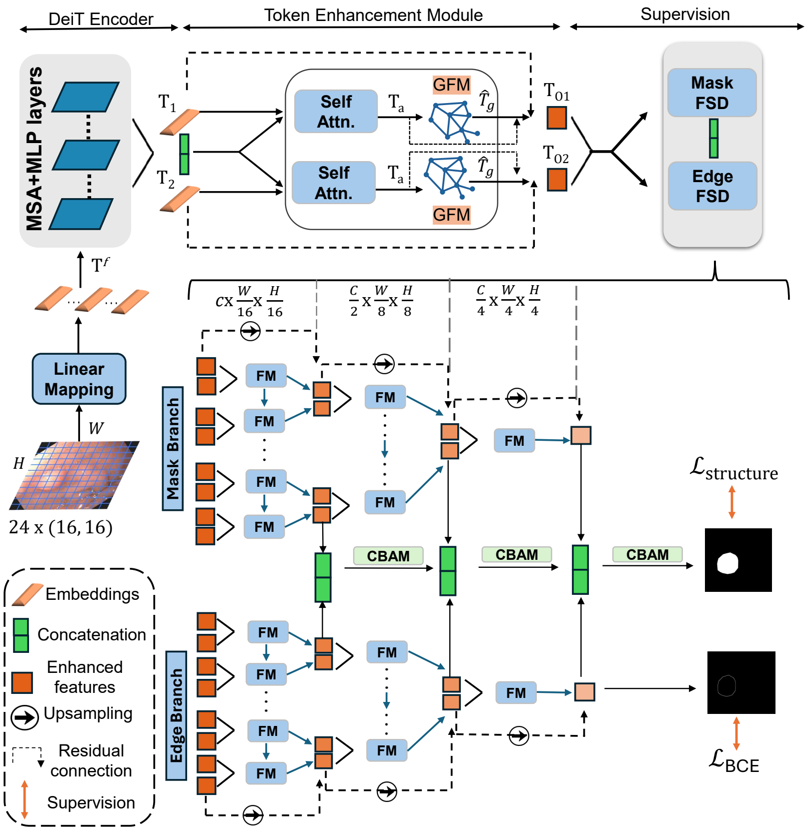

Fig. 1. Overall architecture of ESPNet includes a DeiT encoder, token enhancement module, edge and mask FSDs with residual connections, and attention-guided fusion.

2.3 Token Enhancement Module

While ViTs capture strong contextual information, they lack mechanisms for understanding local information. FSPNet provides a local enhancement mechanism that fuses adjacent token pairs through attention and graph-based refinement. First, two adjacent tokens T_1, T_2 are fused into one token that interacts with them to produce an attention-weighted feature map T_a and explore correlations with them. The tokens are then refined through a single-layer graph convolutional network (GCN) that learns high-level semantic relations between regions to capture global representations within tokens (GFM), outputting $\hat{T}_g$. Finally, a skip connection combines the original token T_1 with T_a and $\hat{T}_g$, followed by deserialization D to produce the enhanced token as 2D image features with the same dimension as the original features for decoding, giving the output token T_{O_1}, and the same can be applied to get T_{O_2} and other token pairs:

$$T_{O_1} = D(\hat{T}_g \otimes T_a^\top + T_1) \tag{2}$$

2.4 Edge-Aware Feature Shrinkage Decoder

FSPNet's FSD facilitates an intra- and inter-layer feature fusion module (FM), where it hierarchically aggregates feature pairs from consecutive scales in a four-scale pyramid (16, 8, 4, 1), in addition to combining features within the same scale, achieving a smooth flow for cue accumulation. We build on FSPNet's decoder by creating another FSD for edge prediction E to guide the mask segmentation M. The edge pyramid has the same FSD structure, where we concatenate the learned edge features to the mask features at different FSD scales s, and we apply a convolutional block attention module (CBAM) to refine the aggregated features [24], in addition to adding residual connections between scales to retain information throughout aggregation. Figure 1 illustrates our edge-aware shrinkage module. Finally, we use structure loss ($\mathcal{L}_{structure}$) for mask prediction [23] compared with the ground truth (GT) mask G_M, and binary cross-entropy loss ($\mathcal{L}_{BCE}$) for edge prediction compared with GT edge G_E. Both losses contribute equally to the joint loss. The pyramid is supervised at the four different scales through GT feature maps PixelShuffle upsampling operations:

$$\mathcal{L}_{\text{joint}} = \sum_{s=0}^{3} \mathcal{L}_{\text{structure}}(M_s, G_M) + \sum_{s=0}^{3} \mathcal{L}_{\text{BCE}}(E_s, G_E) \tag{3}$$

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

To push generalizability to new centers and populations, we compiled a 4000-image white light endoscopy (WLE) training dataset: 804 images from Kvasir-SEG [9], 1,182 images from centers 1-5 in the PolypGen dataset [2], and 2,014 images from the SUN-SEG training dataset [11]. SUN-SEG is a video colonoscopy dataset that includes 100 positive cases with polyp frames. Around 20 non-consecutive frames were selected per positive case, such that the polyps are viewed from different angles. To evaluate our model against other methods, we train our model on the compiled dataset, and use five publicly available polyp segmentation datasets for evaluation: Kvasir-Sessile [8] and SUN-SEG-Easy (Seen) [11] are used to assess performance on test sets from centers seen during training. ETIS-LaribPolypDB [18], CVC-ColonDB [19], and PolypGen-C6 [2] are used for out-of-distribution center testing (not seen during training). Moreover, we use our private in-house IBD dataset to test on an unseen center and population during training. Our dataset consists of 95 images collected between 2017 and 2020 from 44 patients with IBD at Leeds Teaching Hospitals NHS Trust (REC ethical approval reference 17/EM/0033). All frames contain visible polyps and are annotated with masks validated independently by two expert endoscopists, each with over 10 years of clinical experience.

We use five widely used polyp segmentation metrics to evaluate the performance: mean Intersection over Union score (mIoU), mean Dice score (mDice), S-measure (S_α), weighted F-measure (F_β), and mean absolute error (MAE) [6].

Table 1. Quantitative comparison with eight methods on similar distribution (**seen**) datasets. ↑ / ↓ denote larger/smaller is better, respectively, with best results in bold.

Data	Methods	$mIoU\uparrow$	$mDice\uparrow$	$S_\alpha\uparrow$	$F_{\beta w}\uparrow$	$MAE\downarrow$
Kvasir-Sessile	PraNet [MICCAI'20] [6]	0.772	0.845	0.891	0.829	0.027
	SANet [MICCAI'21] [22]	0.700	0.793	0.860	0.753	0.044
	LDNet [MICCAI'22] [26]	0.770	0.849	0.904	0.828	0.026
	CFANet [PR'23] [28]	0.749	0.832	0.887	0.802	0.033
	TransNetR [PMLR'23] [10]	0.774	0.850	0.894	0.836	0.024
	Polyp-PVT [AIR'23] [5]	0.807	0.878	0.912	0.863	0.020
	ASPS [MICCAI'24] [12]	0.475	0.600	0.726	0.548	0.064
	MEGANet [WACV'24] [4]	**0.819**	**0.883**	0.911	0.870	0.021
	Ours (ESPNet)	**0.819**	**0.883**	**0.913**	**0.874**	**0.019**
SUN-SEG-Easy	PraNet [MICCAI'20] [6]	0.795	0.861	0.905	0.845	0.019
	SANet [MICCAI'21] [22]	0.720	0.808	0.875	0.763	0.036
	LDNet [MICCAI'22] [26]	0.818	0.883	0.930	0.861	0.016
	CFANet [PR'23] [28]	0.767	0.844	0.902	0.806	0.025
	TransNetR [PMLR'23] [10]	0.819	0.881	0.920	0.870	0.014
	Polyp-PVT [AIR'23] [5]	0.858	0.910	0.933	0.901	0.015
	ASPS [MICCAI'24] [12]	0.53	0.629	0.726	0.622	0.044
	MEGANet [WACV'24] [4]	**0.864**	**0.917**	**0.936**	**0.907**	**0.011**
	Ours (ESPNet)	0.859	0.912	**0.936**	0.902	**0.011**

3.2 Implementation Details

Our proposed ESPNet is implemented in PyTorch. We used DeiT-trained ViT as an encoder, with an input image size of 384 × 384 and a patch size of 16 × 16. Geometric augmentations (vertical and horizontal flips, and random rotations) were applied to all the training data. Adam was used as the optimizer, and the learning rate was initialized to $1e^{-4}$ and then scaled down by 10 at 50 epochs in our model and the baseline. The complete training process for 100 epochs with a batch size of 8 took around 9 hours on 4 NVIDIA L40S 48GB GPUs.

3.3 Results

To demonstrate the effectiveness of our method, we compare it with eight popular state-of-the-art (SOTA) polyp segmentation models. The models were retrained on our dataset with the same implementation details to ensure fairness. Table 1 shows the performance on the seen datasets: Kvasir-Sessile (same center as Kvasir-SEG) and SUN-SEG-Easy (Seen). Table 2 shows the performance on the unseen datasets.

While ESPNet has similar performance to SOTAs in the seen datasets, it achieves the best metrics on all unseen datasets at frames per second (FPS) ranging between 25 to 35, which is around the real-time requirement in colonoscopy (30 FPS) [21]. Specifically, we show the following mIoU improvements on the different test sets: 2.84% on ETIS-LaribPolypDB, 1.26% on CVC-ColonDB, 2.01% on PolypGen-C6, and 3.52% on our IBD dataset. It can be seen that there is

Table 2. Quantitative comparison with 8 methods on **unseen** datasets. Notes: ↑ / ↓ denote that larger/smaller is better, respectively, with best results in bold.

Data	Methods	$mIoU$↑	$mDice$↑	S_{α}↑	$F_{\beta w}$ ↑	MAE↓
ETIS-LaribPolypDB	PraNet [MICCAI'20] [6]	0.662	0.722	0.847	0.704	**0.016**
	SANet [MICCAI'21] [22]	0.579	0.677	0.803	0.611	0.026
	LDNet [MICCAI'22] [26]	0.631	0.704	0.840	0.665	0.024
	CFANet [PR'23] [28]	0.634	0.721	0.843	0.672	**0.016**
	TransNetR [PMLR'23] [10]	0.643	0.717	0.837	0.684	0.016
	Polyp-PVT [AIR'23] [5]	0.685	0.753	0.854	0.731	0.020
	ASPS [MICCAI'24] [12]	0.333	0.418	0.666	0.389	0.033
	MEGANet [WACV'24] [4]	0.739	0.811	0.872	0.783	0.026
	Ours (ESPNet)	**0.760**	**0.827**	**0.890**	**0.808**	**0.016**
CVC-ColonDB	PraNet [MICCAI'20] [6]	0.679	0.752	0.844	0.746	**0.032**
	SANet [MICCAI'21] [22]	0.622	0.712	0.811	0.677	0.050
	LDNet [MICCAI'22] [26]	0.702	0.783	0.865	0.771	0.038
	CFANet [PR'23] [28]	0.684	0.765	0.849	0.747	0.036
	TransNetR [PMLR'23] [10]	0.650	0.725	0.823	0.718	0.038
	Polyp-PVT [AIR'23] [5]	0.700	0.775	0.848	0.767	0.037
	ASPS [MICCAI'24] [12]	0.405	0.497	0.693	0.488	0.052
	MEGANet [WACV'24] [4]	0.716	0.794	0.845	0.782	0.047
	Ours (ESPNet)	**0.725**	**0.802**	**0.856**	**0.794**	**0.032**
PolypGen-C6	PraNet [MICCAI'20] [6]	0.696	0.759	0.880	0.737	0.019
	SANet [MICCAI'21] [22]	0.629	0.706	0.840	0.662	0.034
	LDNet [MICCAI'22] [26]	0.698	0.757	0.884	0.738	0.022
	CFANet [PR'23] [28]	0.663	0.735	0.865	0.705	0.026
	TransNetR [PMLR'23] [10]	0.695	0.751	0.876	0.733	0.014
	Polyp-PVT [AIR'23] [5]	0.739	0.795	0.903	0.780	**0.012**
	ASPS [MICCAI'24] [12]	0.521	0.605	0.783	0.570	0.043
	MEGANet [WACV'24] [4]	0.738	0.800	0.883	0.777	0.031
	Ours (ESPNet)	**0.753**	**0.802**	**0.890**	**0.791**	0.024
IBD	PraNet [MICCAI'20] [6]	0.646	0.728	0.819	0.703	0.053
	SANet [MICCAI'21] [22]	0.142	0.191	0.490	0.174	0.128
	LDNet [MICCAI'22] [26]	0.647	0.738	0.833	0.703	0.046
	CFANet [PR'23] [28]	0.585	0.68	0.794	0.646	0.044
	TransNetR [PMLR'23] [10]	0.614	0.710	0.803	0.676	0.044
	Polyp-PVT [AIR'23] [5]	0.658	0.750	0.827	0.720	0.054
	ASPS [MICCAI'24] [12]	0.460	0.572	0.710	0.542	0.054
	MEGANet [WACV'24] [4]	0.682	0.767	0.832	0.736	0.057
	Ours (ESPNet)	**0.706**	**0.795**	**0.851**	**0.773**	**0.032**

a drop in performance across all models on the IBD dataset as opposed to the other unseen datasets, demonstrating the generally more challenging scenarios encountered in IBD. Figure 2 demonstrates the qualitative results on the unseen test set; the top four performing models were selected for visualization: ESPNet, MEGANet, PolypPVT, and LDNet. It shows that ESPNet is able to detect challenging cases more accurately, such as flat, multiple, small, and large polyps

Table 3. Ablation experiments of ESPNet on CVC-ColonDB and ETIS-LaribPolypDB

Experiment	**ETIS-LaribPolypDB**		**CVC-ColonDB**	
	mIoU	*mDice*	*mIoU*	*mDice*
FSPNet Baseline	0.683	0.763	0.692	0.770
+Edge FSD	0.733	0.809	0.706	0.783
+CBAM	0.730	0.807	0.714	0.794
+Residual Connections	0.760	0.827	0.725	0.802

(top to bottom). MEGANet seems to have more false positives, and LDNet has more false negatives.

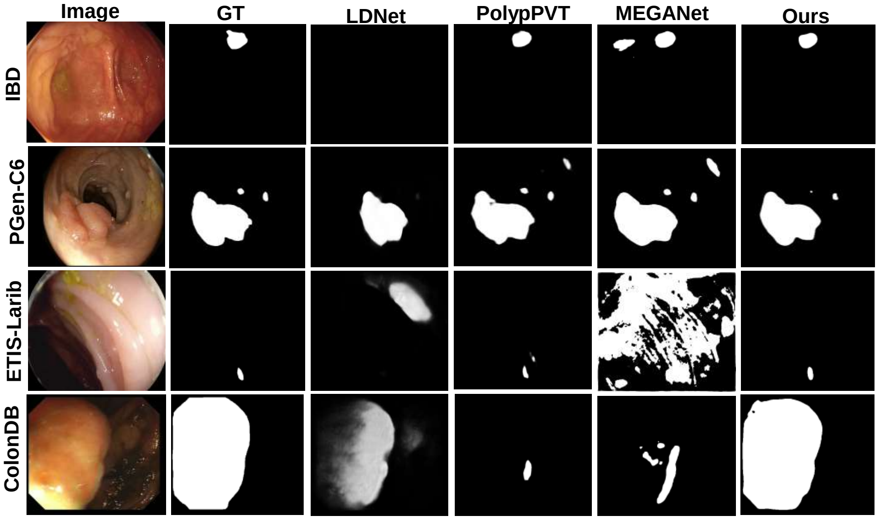

Fig. 2. Qualitative results of the top four methods on unseen dataset samples.

3.4 Ablation Study

To understand the generalizability to unseen data, we chose two external datasets, CVC-ColonDB and ETIS-LaribPolypDB. Our ablation shows improvement on both datasets. We compare with FSPNet as our baseline and show that the performance increases in both mIoU and mDice when adding the edge FSD, then attention after edge-mask feature concatenation, and finally adding residual connections between consecutive scales. Our overall improvement from the FSPNet baseline to our final ESPNet is 11.27% in mIoU on ETIS-LaribPolypDB and 4.77% in mIoU on CVC-ColonDB, showing that our incorporation of edge features with attention and residual connections is beneficial.

4 Conclusion

In this work, we proposed a novel edge-aware feature shrinkage decoding network, ESPNet, for effective polyp segmentation in varied seen and unseen polyp datasets. The key idea is to improve low-level morphological feature learning by adding an additional feature shrinkage decoder for edge detection and concatenating these edge features with mask features, before refining them through attention and adding residual connections between consecutive scales, thereby retaining important features across various scales. Our experiments demonstrated improved model performance and generalizability to out-of-distribution datasets, including four different population cohorts compared to the training dataset.

Acknowledgments

The work was supported in part by UK Research and Innovation (UKRI) [CDT grant number EP/S024336/1], Crohn's & Colitis UK (M2023-5), and the Academy of Medical Sciences (SBF0010\1191). This work was undertaken on the Aire High-Performance Computing (HPC) system at the University of Leeds, UK.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ali, S.: Where do we stand in AI for endoscopic image analysis? deciphering gaps and future directions **5**(1), 1–13. https://doi.org/10.1038/s41746-022-00733-3, publisher: Nature Publishing Group
2. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., Petlund, A., Halvorsen, P., Rittscher, J., de Lange, T., East, J.E.: A multi-centre polyp detection and segmentation dataset for generalisability assessment **10**(1), 75. https://doi.org/10.1038/s41597-023-01981-y, publisher: Nature Publishing Group
3. Bar, O., Neimark, D., Zohar, M., Hager, G.D., Girshick, R., Fried, G.M., Wolf, T., Asselmann, D.: Impact of data on generalization of AI for surgical intelligence applications **10**(1), 22208. https://doi.org/10.1038/s41598-020-79173-6, publisher: Nature Publishing Group
4. Bui, N.T., Hoang, D.H., Nguyen, Q.T., Tran, M.T., Le, N.: MEGANet: Multiscale edge-guided attention network for weak boundary polyp segmentation. pp. 7985–7994. https://doi.org/https://doi.org/10.48550/arXiv.2309.03329
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-PVT: Polyp segmentation with pyramid vision transformers **2**. https://doi.org/10.26599/AIR.2023.9150015
6. 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: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 263–273. Springer International Publishing. https://doi.org/10.1007/978-3-030-59725-2_26

7. Huang, Z., Dai, H., Xiang, T.Z., Wang, S., Chen, H.X., Qin, J., Xiong, H.: Feature shrinkage pyramid for camouflaged object detection with transformers. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5557–5566. IEEE. https://doi.org/10.1109/CVPR52729.2023.00538
8. Jha, D., Smedsrud, P.H., Johansen, D., de Lange, T., Johansen, H.D., Halvorsen, P., Riegler, M.A.: A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation **25**(6), 2029–2040. https://doi.org/10.1109/JBHI.2021.3049304
9. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T.d., Johansen, D., Johansen, H.D.: Kvasir-SEG: A segmented polyp dataset. https://doi.org/10.48550/arXiv.1911.07069
10. Jha, D., Tomar, N.K., Sharma, V., Bagci, U.: TransNetR: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In: Medical Imaging with Deep Learning. pp. 1372–1384. PMLR. https://doi.org/https://doi.org/10.48550/arXiv.2303.07428, ISSN: 2640-3498
11. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Gool, L.: Video polyp segmentation: A deep learning perspective **19**, 1–19. https://doi.org/10.1007/s11633-022-1371-y
12. Li, H., Zhang, D., Yao, J., Han, L., Li, Z., Han, J.: ASPS: Augmented segment anything model for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, vol. 15009, pp. 118–128. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72114-4_12
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. https://doi.org/10.48550/arXiv.1411.4038
14. Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C.J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N., Bray, F.: Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN **72**(2), 338–344. https://doi.org/10.1136/gutjnl-2022-327736
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. https://doi.org/10.48550/arXiv.1505.04597
16. Shah, S.C., Itzkowitz, S.H.: Colorectal cancer in inflammatory bowel disease: Mechanisms and management **162**(3), 715–730.e3. https://doi.org/10.1053/j.gastro.2021.10.035
17. Shine, R., Bui, A., Burgess, A.: Quality indicators in colonoscopy: an evolving paradigm **90**(3), 215–221. https://doi.org/10.1111/ans.15775
18. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer **9**(2), 283–293. https://doi.org/10.1007/s11548-013-0926-3
19. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information **35**(2), 630–644. https://doi.org/10.1109/TMI.2015.2487997, conference Name: IEEE Transactions on Medical Imaging
20. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. https://doi.org/10.48550/arXiv.2012.12877
21. Wang, P., Berzin, T.M., Brown, J.R.G., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., Li, Y., Xu, G., Tu, M., Liu, X.: Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study . https://doi.org/10.1136/gutjnl-2018-317500, https://gut.bmj.com/content/68/10/1813, publisher: BMJ Publishing Group Section: Endoscopy

22. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. https://doi.org/10.48550/arXiv.2108.00882
23. Wei, J., Wang, S., Huang, Q.: F^{3}net: Fusion, feedback and focus for salient object detection **34**(7), 12321–12328. https://doi.org/10.1609/aaai.v34i07.6916, number: 07
24. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, vol. 11211, pp. 3–19. Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_1, series Title: Lecture Notes in Computer Science
25. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. pp. 99–109. https://doi.org/10.1007/978-3-031-16437-8_10
26. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. https://doi.org/10.48550/arXiv.2301.04904, http://arxiv.org/abs/2301.04904
27. Zhao, S., Wang, S., Pan, P., Xia, T., Chang, X., Yang, X., Guo, L., Meng, Q., Yang, F., Qian, W., Xu, Z., Wang, Y., Wang, Z., Gu, L., Wang, R., Jia, F., Yao, J., Li, Z., Bai, Y.: Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: A systematic review and meta-analysis **156**(6), 1661–1674.e11. https://doi.org/10.1053/j.gastro.2019.01.260
28. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation **140**, 109555. https://doi.org/10.1016/j.patcog.2023.109555