Towards a Benchmark for Colorectal Cancer Segmentation in Endorectal Ultrasound Videos: Dataset and Model Development

Yuncheng Jiang[1,2,3,4†], Yiwen Hu[1,2†], Zixun Zhang[1,2,3†], Jun Wei[1,2], Chun-Mei Feng[3], Xuemei Tang[5], Xiang Wan[4], Yong Liu[3], Shuguang Cui[2,1], and Zhen Li[2,1✉]

[1] Shenzhen Future Network of Intelligence Insititute (FNii), China
[2] SSE, The Chinese University of Hong Kong, Shenzhen, China
[3] IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[4] Shenzhen Research Institute of Big Data (SRIBD), Shenzhen, China
[5] Affiliated Hospital of North Sichuan Medical College, Sichuan, China
yunchengjiang@link.cuhk.edu.cn, lizhen@cuhk.edu.cn

Abstract. Endorectal ultrasound (ERUS) is an important imaging modality that provides high reliability for diagnosing the depth and boundary of invasion in colorectal cancer. However, the lack of a large-scale ERUS dataset with high-quality annotations hinders the development of automatic ultrasound diagnostics. In this paper, we collected and annotated the first benchmark dataset that covers diverse ERUS scenarios, *i.e.* colorectal cancer segmentation, detection, and infiltration depth staging. Our ERUS-10K dataset comprises 77 videos and 10,000 high-resolution annotated frames. Based on this dataset, we further introduce a benchmark model for colorectal cancer segmentation, named the **A**daptive **S**parse-context **TR**ansformer (**ASTR**). ASTR is designed based on three considerations: scanning mode discrepancy, temporal information, and low computational complexity. For generalizing to different scanning modes, the adaptive scanning-mode augmentation is proposed to convert between raw sector images and linear scan ones. For mining temporal information, the sparse-context transformer is incorporated to integrate inter-frame local and global features. For reducing computational complexity, the sparse-context block is introduced to extract contextual features from auxiliary frames. Finally, on the benchmark dataset, the proposed ASTR model achieves a 77.6% Dice score in rectal cancer segmentation, largely outperforming previous state-of-the-art methods.

Keywords: Endorectal ultrasound · Segmentation · Transformer.

1 Introduction

Colorectal cancer (CRC) has become the second leading cause of cancer death worldwide [4]. Accurate early detection of CRC is crucial for making therapeu-

† Equal contributions.

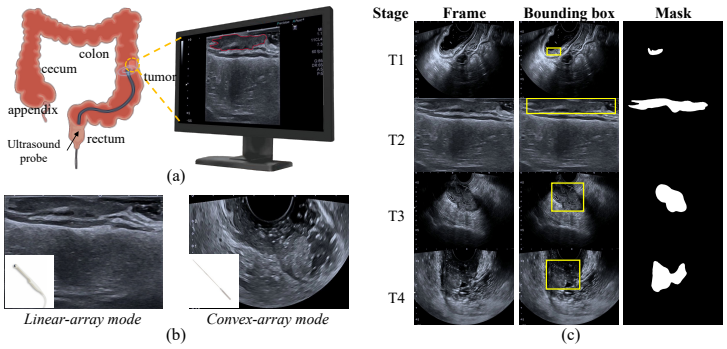

Fig. 1. (a) Schematic diagram of ERUS operation. (b) Different scanning modes of ultrasound. (c) Examples of our ultrasound video dataset with corresponding labels.

tic decisions and improving the survival rate. Currently, Endorectal ultrasound (ERUS) is adopted as a routine imaging modality for diagnosing and staging colorectal cancer [8]. As shown in Fig. 1(a), ERUS provides in-depth assessments of tumor infiltration, precisely depicting the cancer's location, size, and its relationship with surrounding tissues [15]. However, the sonographers's level of experience or fatigue during long duty hours contributes to a non-negligible rate of missed detections in clinical diagnosis. Thus, it is important to develop an automatic system for computer-aided diagnosis of CRC from ERUS videos.

Previous works have proposed well-annotated ultrasound datasets and corresponding methods, covering various organs such as the breast, pancreas, and thyroid [7, 10, 13]. Li *et al.* [10] pioneer the application of deep neural networks in the diagnosis of pancreatic diseases using Endoscopic Ultrasound (EUS). Wang *et al.* introduce a 3D feature pyramid network to conduct inter-frame collaboration. Li *et al.* [11] propose a memory bank and dynamically update the memory to establish long-term temporal correlation. Building upon this, Lin *et al.* [13] extend the approach by incorporating Fourier transforms to aggregate multiple features from the frequency domain. Despite significant advancements achieved by those methods, accurate colorectal cancer segmentation in ERUS remains challenging due to (1) the scarcity of large-scale endorectal ultrasound datasets to train well-converged segmentation models, (2) the intrinsic scanning mode discrepancy obtained from different ultrasound sensors, as illustrated in Fig. 1(b), and (3) the motion blur resulted from a rapidly moving ultrasound probe.

To address the above issues, we collected 77 endorectal ultrasound videos with 10,000 well-annotated frames and propose the ***first*** **large-scale ERUS video colorectal cancer segmentation dataset,** of which 57 videos contain tumor infiltration depth staging, contributing to realistic clinical scenarios. Apart from the benchmark dataset, we further propose a benchmark model for colorectal cancer segmentation, termed the **A**daptive **S**parse-context **TR**ansformer (**ASTR**). ASTR is designed based on three main considerations: scanning mode discrepancy, temporal information, and computational complexity. To gener-

alize to different scanning modes, the adaptive scanning mode augmentation (ASMA) is proposed, which converts images between linear and convex scanning modes through coordinate transformation (*i.e.*, Polar and Cartesian coordinate systems). To exploit temporal information between frames, the Sparse-context Transformer is incorporated to integrate inter-frame local and global features. To reduce computational complexity during temporal fusion, the Sparse-context Block (SCB) is introduced to extract contextual features while eliminating irrelevant noises from reference frames.

In summary, our contributions include three aspects: (1) We present the first well-annotated endorectal ultrasound dataset with comprehensive annotations, laying the foundation for advancements in automatic ultrasound diagnosis of colorectal cancer. (2) We build the first adaptive sparse-context transformer model tailored for video colorectal cancer segmentation. (3) We have conducted extensive experiments and established a comprehensive benchmark for video colorectal cancer segmentation. Experimental results demonstrate that our proposed method achieves state-of-the-art performance.

2 Method

2.1 Adaptive Scanning Mode Augmentation

In practical endorectal ultrasound examinations, sonographers employ different probes depending on specific patient conditions. These probes exhibit varying scanning modes (*i.e.*, linear and convex) and result in images of different forms. To ensure our model can generalize to these images under different modes, we propose a simple yet effective data augmentation method called Adaptive Scanning Mode Augmentation (ASMA).

For the convex-array and linear-array modes, there exists a coordinate mapping relationship between them. Therefore, we can obtain the image transformation between different modes through Polar-Cartesian transformation. Specifically, we establish the polar coordinates for the convex-array mode image with the origin point (r_o, θ_o) set at the top center of the image and the x-axis at the top edge. Meanwhile, Cartesian coordinates are set with (x_o, y_o) as the origin, the top edge as the x-axis, and the vertical central axis as the y-axis. **For the convex-array mode**, we obtain the value and position of each pixel (x, y) on the transformed image according to its corresponding pixel (r, θ) on the original convex-array mode image by the following polar-to-cartesian transformation:

$$x = x_o + r\cos\theta, \quad y = y_o + r\sin\theta. \tag{1}$$

Similarly, **for the linear-array mode**, the transformed position (r, θ) of convex-array mode image is calculated using the cartesian-to-polar transformation:

$$r = \sqrt{x^2 + y^2}, \quad \theta = \arctan(\frac{y}{x}). \tag{2}$$

After augmentation, the number of images in the original dataset will be balanced across different scanning modes, reducing the risk of model overfitting to any particular mode due to the imbalanced dataset.

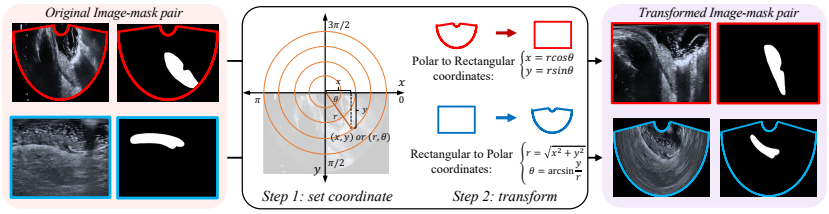

Fig. 2. Schematic illustration of the adaptive scanning mode augmentation (ASMA). The original frame of linear-array/convex-array mode is transformed to the frame of convex-array/linear-array mode by Polar-Cartesian coordinate system transformation, enhancing the model's generalization ability on different scanning modes.

2.2 Sparse-context Transformer

Per-frame context learning. The Sparse-context Transformer can be divided into two key stages: context learning and context fusion. In the per-frame context learning phase, coarse per-frame context feature $f_t \in \mathbb{R}^{c\times h\times w}$ are initially extracted through the backbone network. Then, we refine f_t via stacked transformer layers with self-attention (SA) operation:

$$SA(f_t) = \text{softmax}(\frac{q(f_t)k(f_t)^T}{\sqrt{d}})v(f_t), \quad p = SA(f_t), \tag{3}$$

where $p \in \mathbb{R}^{c\times h\times w}$ is the refined per-frame context and q, k, v are the linear projection functions, *i.e.* $q(f_t) \in \mathbb{R}^{c\times d}$.

Reference-frame context learning. For the reference-frame context learning stage, we enhance the coarse backbone features $\{f_{t-1}, f_{t-2}, \cdots, f_{t-T}\}$ via the sparse-context block (SCB), producing sparse representations for reference-frame contexts $r_{t-1} \in \mathbb{R}^{c\times n_{t-1}}, \cdots, r_{t-T} \in \mathbb{R}^{c\times n_{t-T}}$. Those features are concatenated along the channel dimension to form the reference-frame context $r \in \mathbb{R}^{c\times m}$, $m = (n_{t-1} + \cdots + n_{t-T})$. Then, in the context fusion stage, we adopt a temporal transformer architecture equipped with cross-attention (CA) as described in Li *et al.*[12] to generate multi-frame contexts y:

$$CA(p, r) = \text{softmax}(\frac{q(p)k(r)^T}{\sqrt{d}})v(r), \quad y = CA(p, r). \tag{4}$$

Finally, the fused features are fed to a sequential of up-sampling layers and a segmentation head to generate the predicted mask.

Sparse-context Block. Historical evidence suggests that adjacent frames encapsulate valuable temporal contexts, thereby aiding in bolstering the model's generalization capabilities. However, these temporal contexts are sparsely distributed in adjacent frames [17]. Namely, most pixels are redundant, thereby increasing the computational burden unnecessarily. To reduce the computational cost, we summarized two sparse principles: (1) nearby reference frames provide limited distinguished information, (2) target regions are crucial information for

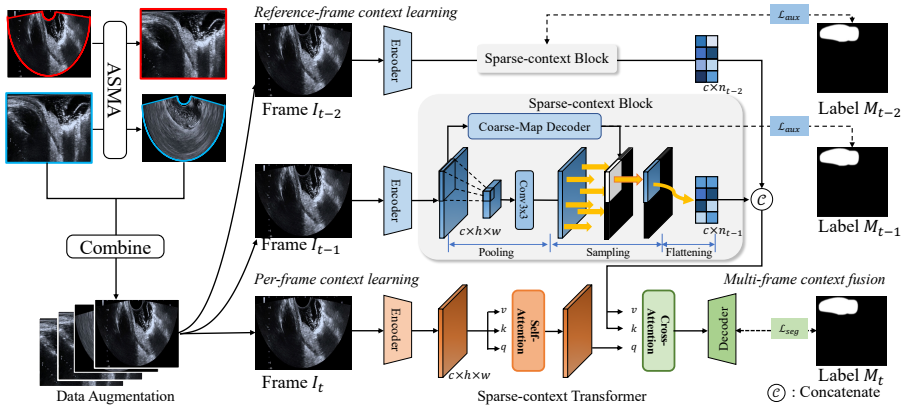

Fig. 3. Pipeline of the proposed ASTR. To generalize to different scanning modes, we first conduct data augmentation by interconverting the linear-array mode and convex-array mode in the adaptive scanning mode augmentation (ASMA). During training, the Sparse-context Transformer extracts inter-frame contexts to exploit spatiotemporal information. Furthermore, we devise a Sparse-context Block to eliminate the irrelevant background noise and reduce computational cost. Finally, the multi-frame contexts from all samples are fused for segmentation mask prediction.

learning. Based on these two principles, we have devised a sparse-context block for selecting key information from reference features. Following principle (1), features undergo pooling operations to adjust the receptive field, followed by a 3×3 convolutional layer to refine the features. Here, the pooling size K is determined by proximity: $K = 2^i$, where i is the index of the reference frame. This process ensures that distant frames can learn features from a larger receptive field, contributing meaningful information. Following principle (2), we introduce explicit attention to obtain sparse representations. Specifically, features pass through a coarse-map decoder, which is a 1×1 convolution layer, to generate a coarse mask of the lesion region. We sample feature values at positions corresponding to the non-zero position in the coarse mask, ensuring that only regions likely to contain lesions contribute to subsequent multi-frame context fusion. Through sparse representation, the computational complexity in the context fusion stage decreases from $\mathcal{O}(h^2w^2tc)$ to $\mathcal{O}(hwmc)$, where $m \ll hwt$.

2.3 Loss Function

We adopt the combination of binary cross-entropy loss $\mathcal{L}_{bce}$, Dice loss $\mathcal{L}_{dice}$, and mean absolute error loss $\mathcal{L}_{mae}$ for pixel-level supervision. Additionally, to refine the localization of lesion regions within the sparse attention module, we incorporate an auxiliary loss function $\mathcal{L}_{aux}$ to supervise the coarse-map decoder. This auxiliary loss, using the ground truth label as supervision, ensures more accurate guidance for SCB during the feature selection process, aligning the

model's focus with the actual lesion areas. The total loss function is defined as follows:

$$\begin{aligned} \mathcal{L}_{seg} = \mathcal{L}_{aux} = \mathcal{L}_{bce} + \mathcal{L}_{dice} + \mathcal{L}_{mae}, \\ \mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{aux} \cdot \mathcal{L}_{aux}, \end{aligned} \tag{5}$$

where the hyper-parameter $\mathcal{L}_{aux} = 0.3$ is used to balance the weight between the segmentation loss and auxiliary loss.

3 Experiment

Endorectal Ultrasound Dataset To facilitate advancements in colorectal cancer segmentation and the staging of tumor infiltration depth, we collected and annotated the first endorectal ultrasound dataset, named **ERUS-10K**, consisting of 77 endorectal ultrasound videos with a total of 10,000 annotated frames. All patients underwent endorectal ultrasound examinations at the Affiliated Hospital of North Sichuan Medical College using CANNO-type color Doppler ultrasound diagnostic apparatus. Among 77 videos, 19 videos were recorded using the linear-array scanning mode by 11CL4 rectal cavity probe, while the remaining utilized the convex-array scanning mode by the vaginal probe. Fig. 1(b) illustrates the differences between these two scanning modes. Manual annotations of colorectal cancers were performed by experienced sonographers. The provided annotations include colorectal lesion masks and bounding boxes, comprehensively covering clinical scenarios ranging from colorectal lesion detection to segmentation. Furthermore, 57 videos implemented pathological examinations via percutaneous biopsy to determine the tumor infiltration depth (*i.e.* stage T1, T2, T3, T4), laying the foundation for automated and precise colorectal cancer staging. Fig. 1(c) displays sample images along with their corresponding labels. The entire dataset is divided into training, validation, and test sets in a ratio of 7:1:2, enabling a comprehensive benchmark evaluation of our proposed methods.
Implementation Details. Our network was implemented using the PyTorch framework on two NVIDIA V100 GPUs. We adopt the Res2Net-50 [5] pre-trained on ImageNet as the backbone. We sample a video clip with $T = 3$ frames for training and inference. All input images are uniformly resized to 352×352 and employ random flip as data augmentation. The whole network is trained in an end-to-end manner using Adam optimizer. The initial learning rate is set to 0.0001. We train the entire model for 24 epochs with batch size 24.

3.1 Comparisons with State-of-the-arts

We conduct comparisons between our method with **10** state-of-the-art segmentation methods, comprising **5** image-based methods: UNet [16], SANet [19], TransUNet [1], SETR [21], and MedSAM [14], and **5** video-based methods: DAFNet [18], PNS-Net [9], DCF-Net [20], SLT-Net [2], and FLA-Net [13]. We compare the performances using various metrics, including the Dice similarity (Dice), Intersection over Union (IoU), Mean Absolute Error (MAE), Sensitivity (Sen), Specificity (Spe), and inference Frame Per Second (FPS).

Table 1. Quantitative comparison of our ASTR and other state-of-the-art methods on the ERUS video lesion segmentation dataset.

	Pubs.	Methods	MAE↓	IoU↑	Dice↑	Sen↑	Spe↑	FPS↑
Image-based	*MICCAI15*	UNet	5.0	59.3	72.6	74.4	98.1	65.2
	MICCAI20	SANet	4.5	59.8	73.2	72.0	98.5	56.6
	CVPR21	TransUNet	5.2	58.0	71.7	70.9	98.0	58.1
	CVPR21	SETR	7.4	56.0	69.8	75.1	97.8	27.5
	Nature24	MedSAM	2.9	62.8	75.7	75.5	98.3	2.2
Video-based	*TMI19*	DAF-Net	4.2	59.9	73.6	73.7	98.1	30.1
	MICCAI21	PNS-Net	3.6	61.2	74.5	75.8	98.3	16.5
	CVPR21	DCF-Net	3.2	61.6	74.7	73.2	98.4	6.6
	CVPR22	SLT-Net	3.0	62.2	75.2	74.0	98.4	10.2
	MICCAI23	FLA-Net	3.2	61.7	74.8	76.0	98.3	35.5
		ASTR	**2.7**	**65.7**	**77.6**	**78.5**	**98.6**	**40.8**

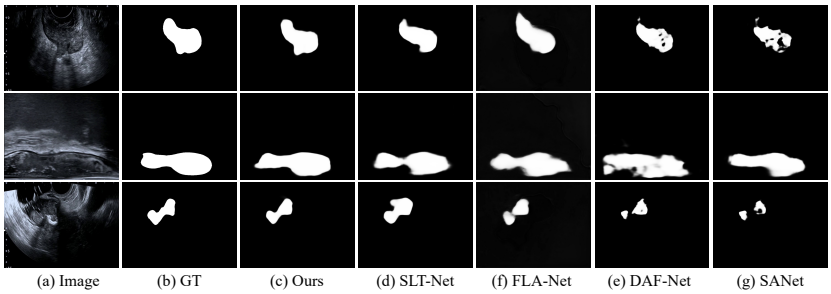

Fig. 4. Qualitative comparisons. "GT" denotes the ground truth. See *Suppl.* for more visualization results.

Quantitative Comparisons. As shown in Table 1, among the compared methods, video-based methods generally outperformed image-based methods since temporal information is considered. With the proposed modules, our ASTR model exhibited significant improvements across all metrics, surpassing the state-of-the-art methods by a large margin. Specifically, it increased the Dice score from 75.7% to 77.6%, the IoU score from 62.8% to 65.7%, the sensitivity from 76.0% to 78.5%, and reduced the MAE score from 2.9% to 2.7%.

Qualitative Comparisons. Fig. 4 visualize the segmentation results of our method on the ERUS dataset. In ultrasound videos, highly differentiated tumors exhibit diverse shapes and sizes, often with indistinct boundaries. Compared to other methods, our approach demonstrates superior robustness, enabling accurate localization and segmentation of tumors even in challenging scenarios. More importantly, our method offers higher detection rates and lower false detection rates, which are crucial for clinical decision-making.

Table 2. Ablation study on different components.

Design	MAE↓	Dice↑	FPS↑
baseline	4.5	73.6	
w/ CopyPaste	$4.1_{\downarrow 0.4}$	$74.1_{\uparrow 0.5}$	32.6
w/ ArSDM	$4.0_{\downarrow 0.5}$	$74.5_{\uparrow 0.9}$	
w/ ASMA	$3.2_{\downarrow 1.3}$	$75.6_{\uparrow 2.0}$	
w/ SCB	$3.8_{\downarrow 0.7}$	$74.9_{\uparrow 1.3}$	
w/ SCB+$\mathcal{L}_{aux}$	$2.8_{\downarrow 1.7}$	$75.8_{\uparrow 2.2}$	$40.8_{\uparrow 8.2}$
Ours	$\mathbf{2.7}_{\downarrow 1.8}$	$\mathbf{77.6}_{\uparrow 4.0}$	

Table 3. Ablation study of ASMA as data augmentation.

Design	MAE↓	Dice↑	Sen↑
SANet	4.5	73.2	72.0
w/ ASMA	$3.6_{\downarrow 0.9}$	$74.6_{\uparrow 1.4}$	$74.8_{\uparrow 2.8}$
SLT-Net	3.0	75.2	74.0
w/ ASMA	$2.7_{\downarrow 0.3}$	$77.0_{\uparrow 1.8}$	$77.7_{\uparrow 3.7}$
FLA-Net	3.2	74.8	78.0
w/ ASMA	$2.9_{\downarrow 0.3}$	$76.4_{\uparrow 1.6}$	$77.2_{\downarrow 0.8}$

3.2 Ablation Study

Effectiveness of components. We investigate the contribution of each proposed component, as shown in Tab. 2. We employed a network without the ASMA and SCB modules as the "baseline". After individually integrating the designed ASMA and SCB into the network, significant performance improvements can be observed. More importantly, the sparsing operations reduce the redundant computations, thereby accelerating the inference FPS. Furthermore, combining these two modules results in a 1.8% decrease in MAE, 4% increase in Dice, and a 0.5% increase in Sensitivity compared to the baseline. These results indicate that our method provides richer ultrasound features and mitigates noise interference during multi-frame context fusion. Additionally, we conducted a comparison with two advanced data augmentation techniques: CopyPaste[6] and ArSDM[3]. CopyPaste involves pasting foreground regions of each linear/convex-array mode onto a random convex/linear-array background, which yielded less performance improvement compared to our ASMA. On the other hand, compared to ArSDM, which trains a diffusion model to generate simulation data, our method saves computational resources and achieves higher performance gains.

Effectiveness of adaptive scanning mode augmentation. We further investigate the effectiveness of adaptive scanning mode augmentation by applying it to three different segmentation methods. As shown in Tab. 3, all methods equipped with our proposed ASMA achieved observable segmentation performance without any extra cost during the inference stage.

4 Conclusion

In this paper, we explore the potential of computer-aided automated segmentation of colorectal cancer in endorectal ultrasound videos and contribute the first well-annotated endorectal ultrasound dataset with segmentation and infiltration depth staging labels. Besides, we evaluate 10 state-of-the-art cancer segmentation methods on the proposed dataset and establish benchmark performance

metrics. Furthermore, we devise the first colorectal cancer segmentation model, ASTR, tailored for endorectal ultrasound videos and achieves state-of-the-art performance. We hope this work can inspire researchers and pave the way for future works on computer-aided automatic endorectal ultrasound diagnosis.

Acknowledgments. This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, %The Chinese University of Hong Kong, Shenzhen CHUK-Shenzhen, by the NSFC 61931024&12326610, by Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), by Tencent & Huawei Open Fund, by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), and by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141.

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

References

1. 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)
2. Cheng, X., Xiong, H., Fan, D.P., Zhong, Y., Harandi, M., Drummond, T., Ge, Z.: Implicit motion handling for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13864–13873 (2022)
3. Du, Y., Jiang, Y., Tan, S., Wu, X., Dou, Q., Li, Z., Li, G., Wan, X.: Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In: International conference on medical image computing and computer-assisted intervention. pp. 339–349. Springer (2023)
4. Favoriti, P., Carbone, G., Greco, M., Pirozzi, F., Pirozzi, R.E.M., Corcione, F.: Worldwide burden of colorectal cancer: a review. Updates in surgery **68**, 7–11 (2016)
5. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence **43**(2), 652–662 (2019)
6. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2918–2928 (2021)

7. Han, H., Liao, H., Zhang, D., Kong, W., Chen, F.: Thyroid nodule diagnosis in dynamic contrast-enhanced ultrasound via microvessel infiltration awareness. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 169–179. Springer (2023)
8. Hünerbein, M.: Endorectal ultrasound in rectal cancer. Colorectal Disease **5**(5), 402–405 (2003)
9. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
10. Li, J., Zhang, P., Wang, T., Zhu, L., Liu, R., Yang, X., Wang, K., Shen, D., Sheng, B.: Dsmt-net: Dual self-supervised multi-operator transformation for multi-source endoscopic ultrasound diagnosis. IEEE Transactions on Medical Imaging (2023)
11. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–400. Springer (2022)
12. Li, J., Wang, W., Chen, J., Niu, L., Si, J., Qian, C., Zhang, L.: Video semantic segmentation via sparse temporal transformer. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 59–68 (2021)
13. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 497–507. Springer (2023)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Rieger, N., Tjandra, J., Solomon, M.: Endoanal and endorectal ultrasound: applications in colorectal surgery. ANZ journal of surgery **74**(8), 671–675 (2004)
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. Sun, G., Liu, Y., Ding, H., Probst, T., Van Gool, L.: Coarse-to-fine feature mining for video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3126–3137 (2022)
18. Wang, Y., Dou, H., Hu, X., Zhu, L., Yang, X., Xu, M., Qin, J., Heng, P.A., Wang, T., Ni, D.: Deep attentive features for prostate segmentation in 3d transrectal ultrasound. IEEE transactions on medical imaging **38**(12), 2768–2778 (2019)
19. 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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 699–708. Springer (2021)
20. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1553–1563 (2021)
21. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)