Unsupervised Quality Control and Enhancement of Polyp Segmentation in Colonoscopy Videos using Spatiotemporal Consistency

Yujia Li[1], Tao Zhou[1], Ruixuan Wang[2], Shuo Wang[3], Yizhe Zhang[1]✉

[1] Nanjing University of Science and Technology, China
zhangyizhe@njust.edu.cn
[2] Sun Yat-sen University, China
[3] Fudan University, China

Abstract. Reliable polyp segmentation in colonoscopy videos is crucial for early detection and prevention of colorectal cancer. While deep learning-based segmentation models show promise, their performance can be inconsistent, and robust methods for assessing segmentation quality without ground-truth annotations are lacking. This paper presents a novel quality control framework for polyp segmentation that leverages the temporal consistency inherent in colonoscopy videos. Our framework utilizes the Segment Anything Model 2 (SAM2), a powerful video segmentation foundation model, to propagate segmentation predictions between adjacent frames. By evaluating the consistency between these propagated segmentations and the original model predictions, we obtain an unsupervised Segmentation Quality Assessment (SQA) score for each frame. Furthermore, we introduce a re-segmentation module that refines low-quality segmentations by leveraging information from high-quality frames, identified based on their SQA scores. Experiments on the SUN-SEG and PolypGen datasets demonstrate a moderate to strong correlation between the SQA scores produced by our framework and the ground-truth segmentation quality. The re-segmentation module also improves overall segmentation performance without requiring model retraining or fine-tuning. This work suggests a step towards building more reliable and trustworthy AI-assisted colonoscopy systems. The code is available at https://github.com/LYJ-NJUST/Seg-Quality-Control.

Keywords: Video Polyp Segmentation · Segmentation Quality Assessment · Quality Control · Colonoscopy · Foundation Model

1 Introduction

Colorectal cancer (CRC) ranks as the third most common malignancy and the second leading cause of cancer-related deaths globally, posing a substantial threat to public health [16]. Colorectal adenomatous polyps are recognized as the primary precursors to CRC. The timely identification and removal of these polyps are crucial for reducing CRC incidence and improving patient survival rates [16].

Colonoscopy remains the gold standard for polyp detection, but it is a resource-intensive procedure traditionally requiring significant physician time and expertise. Furthermore, the accuracy of colonoscopy can be highly dependent on the operator's skill and experience, with miss rates sometimes exceeding 20% [9, 8].

The advancements in deep learning have led to significant improvements in medical image segmentation accuracy. Automated polyp segmentation holds promise as a valuable tool to assist physicians, potentially reducing their workload and improving diagnostic consistency. However, existing polyp segmentation models often exhibit unstable performance, demonstrating high accuracy on familiar datasets but significantly reduced performance on unseen samples from diverse sources [3]. Even state-of-the-art polyp segmentation models (e.g., [18, 4, 14, 15]) can produce unreliable segmentations when presented with data variability. Crucially, these models typically lack mechanisms for self-assessment of segmentation quality. They generate predictions without any accompanying measure of confidence or accuracy, hindering trust and adoption by clinicians and patients. Therefore, a robust and reliable segmentation quality assessment system is critical for the trustworthy deployment of medical AI systems in clinical practice.

The Segment Anything Model 2 (SAM2) [13] is a powerful general-purpose segmentation foundation model trained on a massive dataset of videos and masks (50.9K videos and 35.5M masks). SAM2 demonstrates strong zero-shot generalization capabilities for object segmentation in both images and videos, across a variety of downstream tasks. In this paper, we leverage SAM2's video segmentation capabilities and promptable interface to develop a novel quality control framework for polyp segmentation in colonoscopy videos. This framework aims to assess and potentially enhance the reliability of automated polyp segmentation without the need for ground-truth annotations. Our approach operates as follows: given a polyp segmentation model's output for a colonoscopy video, we extract visual prompts from the predicted segmentation masks. These prompts are then used with SAM2 to propagate the segmentation results to adjacent frames. By measuring the spatiotemporal consistency (using the Dice coefficient) between the original segmentation and the SAM2-propagated segmentations, we obtain an indicator of the segmentation quality. Frames with high consistency scores are considered more reliable. Furthermore, these high-quality segmentations can be used to refine potentially lower-quality segmentations through a re-segmentation process. Our experiments on the SUN-SEG [10, 11] and PolypGen [2] datasets demonstrate that the proposed framework's quality assessment scores exhibit a moderate to strong positive correlation with actual segmentation quality (measured against ground truth). Moreover, the framework demonstrates the capability to improve the overall segmentation quality without requiring any model retraining or fine-tuning. This work makes the following key contributions: (1) **Unsupervised Quality Assessment:** We introduce a novel framework for assessing polyp segmentation quality in colonoscopy videos without relying on ground-truth annotations. (2) **Segmentation Enhancement:** Our framework can improve polyp segmentation quality without the need for additional train-

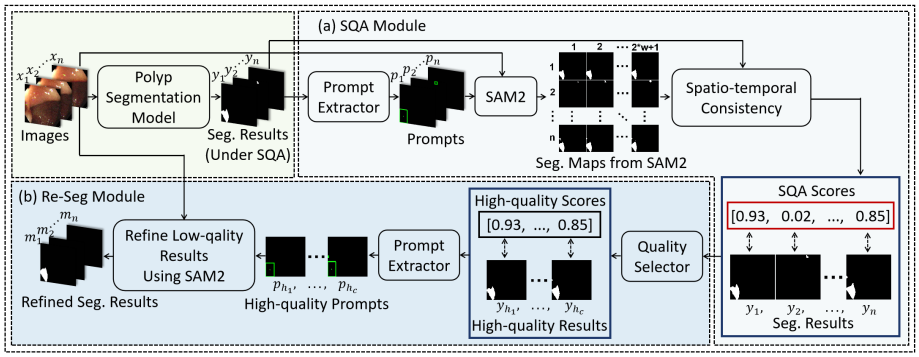

Fig. 1. Method overview: the SQA module evaluates video segmentation quality, and the Re-Seg module provides segmentation refinement.

ing or fine-tuning. (3) **Synergistic Model Integration:** We demonstrate the effective combination of a specialized polyp segmentation model with a general-purpose foundation model (SAM2). This approach is adaptable and can be extended to other medical video segmentation tasks and foundation models.

2 Methodology

The high-level motivation is to utilize the strong spatio-temporal mask propagation capability of video segmentation models (e.g., SAM2) to assess temporal segmentation consistency. This, in turn, helps evaluate segmentation quality and enhance or correct low-quality segmentations using high-quality ones along the temporal domain. Our proposed method, illustrated in Fig. 1, is designed to assess and enhance the quality of polyp segmentation in colonoscopy videos without relying on ground-truth annotations. The method operates in two primary stages: Segmentation Quality Assessment (SQA) and Re-Segmentation (Re-Seg).

2.1 Segmentation Quality Assessment (SQA)

The SQA module leverages the temporal consistency inherent in colonoscopy videos to evaluate the reliability of segmentation predictions. Given a colonoscopy video, we first decompose it into a sequence of n frames, denoted as $\{x_1, x_2, \ldots, x_n\}$, where each frame $x_i \in \mathbb{R}^{W \times H \times C}$ represents a color image with width W, height H, and C color channels (typically $C = 3$ for RGB). An already trained polyp segmentation model, denoted as f_{model}, processes each frame x_i to produce a corresponding binary segmentation mask y_i: $y_i = f_{\text{model}}(x_i)$, where $y_i \in \{0, 1\}^{W \times H}$. Note that we represent the binary mask as a single-channel image for notational simplicity, rather than the two-channel output mentioned in the original text.

The core of the SQA module lies in exploiting the video segmentation capabilities of the Segment Anything Model 2 (SAM2), denoted as f_{SAM2}. We utilize

SAM2 to propagate segmentation information between adjacent frames, thereby assessing the temporal consistency of the initial segmentation results.

More specifically, we operate on a window of frames centered around a given frame x_i. This window, with a radius of w, is defined as the set $\{x_{i-w}, \ldots, x_{i+w}\}$. For the central frame x_i, we first generate a visual prompt p_i from its predicted mask y_i. This prompt, created by an extractor $\mathcal{P}$, is composed of the bounding box coordinates and center points of the segmented object.

We use SAM2 to propagate a mask from a source frame x_i to other frames x_j within a specified temporal window. This process generates a propagated mask, $m_{i\to j}$, for each target frame x_j. The propagation function is designed to process the entire sequence of frames between the source and target, performing the segmentation propagation or tracking:

$$m_{i\to j} = f_{\text{SAM2}}(x_i, p_i, \{x_i, \ldots, x_j\}), \quad j \in \{i-w, \ldots, i+w\} \setminus \{i\}\,. \tag{1}$$

In this formulation, f_{SAM2} takes the initial prompt p_i and the sequence of frames from the source x_i up to the target x_j as input. The output, $m_{i\to j}$, represents the propagated segmentation mask for frame x_j, with pixel values ranging from $[0, 1]$.

We then measure how consistent the original segmentation, y_u, is with each of the propagated masks, $m_{v\to u}$. We do this using the Dice similarity coefficient (DSC), calculated as:

$$d_{v\to u} = \text{DSC}(y_u, m_{v\to u}) = \frac{2|y_u \cap m_{v\to u}|}{|y_u| + |m_{v\to u}|} \tag{2}$$

Here, $d_{v\to u}$ represents the Dice similarity between the original segmentation in frame u (y_u) and the mask propagated from frame v to frame u ($m_{v\to u}$). A higher DSC value means better overlap and consistency between the two masks.

Finally, we compute the SQA score, s_u, for the segmentation y_u of frame x_u. This score is simply the average DSC across all masks propagated to that specific frame:

$$s_u = \frac{1}{2w} \sum_{v=u-w \wedge v\neq u}^{u+w} d_{v\to u}\,. \tag{3}$$

The unsupervised SQA score, s_u, provides an evaluation of the segmentation mask's quality (y_u) for frame x_u. A higher s_u score indicates greater **spatiotemporal consistency** with neighboring frames. This consistency acts as a stand-in for higher segmentation quality, as a good segmentation should align well with its surroundings over time.

2.2 Re-Segmentation (Re-Seg)

The Re-Seg module leverages the SQA scores to selectively refine segmentation results. The process involves identifying low-quality frames and replacing their initial segmentations with improved ones derived from higher-quality frames.

First, we compute the average SQA score across all frames in the video: $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$. We then define the set of low-quality frames, L, as those frames whose SQA scores fall below the average: $L = \{x_i \mid s_i < \bar{s}\}$. To identify reliable high-quality frames, we introduce a stability criterion. For each frame x_i not in L, we calculate a self-consistency score $d_{i,i}$ by propagating the segmentation from x_i back to itself using SAM2: $m_{i\to i} = f_{\text{SAM2}}(x_i, p_i, \{x_i\})$, and $d_{i,i} = \text{DSC}(y_i, m_{i\to i})$. Here, the self-consistency score measures how well SAM2 can reproduce the initial segmentation y_i when provided with its own prompt.

We define the set of high-quality frames, H, as those frames that are not low-quality and whose self-consistency score exceeds a predefined threshold, t

$$H = \{x_i \mid s_i \geq \bar{s} \wedge d_{i,i} > t\} \tag{4}$$

The threshold $t \in [0, 1]$ is a hyperparameter controlling the stringency of the high-quality frame selection.

For each low-quality frame $x_l \in L$, we identify the c nearest high-quality frames within the video sequence, denoted as $\{x_{h_1}, x_{h_2}, \ldots, x_{h_c}\}$, where $x_{h_k} \in H$ for all k. "Nearest" is defined in terms of the frame index difference, minimizing:

$$\{h_1, h_2, \ldots, h_c\} = \underset{\{k_1,\ldots,k_c\}}{\arg\min} \sum_{j=1}^{c} |l - k_j|, \quad \text{subject to } x_{k_j} \in H \ \forall j\,. \tag{5}$$

We then propagate the segmentations from these c high-quality frames to the low-quality frame x_l using SAM2:

$$m_{h_k\to l} = f_{\text{SAM2}}(x_{h_k}, p_{h_k}, \{x_{h_k}, \ldots, x_l\}), \quad k \in \{1, 2, \ldots, c\}\,. \tag{6}$$

Finally, we combine these propagated masks to generate the refined segmentation m_l for the low-quality frame x_l: $m_l = \sigma\left(\frac{1}{c}\sum_{k=1}^{c} m_{h_k\to l}\right)$, where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function, applied element-wise. This averaging and sigmoid application produces a final, refined mask $m_l \in [0, 1]^{W\times H}$ for the low-quality frame. This refined mask m_l then replaces the original segmentation y_l.

3 Experiments and Results

For the primary experiments, we utilized a well-trained HSNet [18] (a leading polyp segmentation model, weights obtained from its official release) and SAM2 [13] (a state-of-the-art image and video segmentation foundation model). Two public colonoscopy datasets were used: PolypGen [1] and SUN-SEG [7, 12]. PolypGen, a multi-center dataset from Europe and Africa, contains 23 polyp-positive videos (2225 total frames). SUN-SEG comprises 285 positive colonoscopy videos (49136 total frames) with per-frame ground truth masks.

Within the SQA module, the temporal window radius is set to 10 by default, encompassing 21 frames (the current frame, plus 10 preceding and 10 subsequent frames). The stability threshold for high-quality sample re-segmentation is 0.85. For SQA assessment, Spearman's, Pearson's, and Kendall's correlation coefficients are used to measure the rank and linear correlations between predicted

Table 1. Correlation between predicted SQA scores and true Dice, comparing our method to other unsupervised approaches (higher is better).

SQA Methods	PolypGen			SUN-SEG		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Softmax [6]	0.389	0.297	0.195	0.363	0.262	0.191
Dropout(p=0.3) [5]	0.368	0.486	0.365	0.437	0.433	0.314
Dropout(p=0.5)	0.385	0.480	0.365	0.447	0.430	0.312
SPC(w=1) [19]	0.040	-0.001	-0.005	0.279	0.178	0.127
SPC(w=5)	0.014	-0.008	-0.012	0.287	0.164	0.119
SPC(w=10)	0.011	-0.017	-0.020	0.289	0.159	0.116
SQA-SAM [20]	0.389	0.481	0.356	0.408	0.413	0.300
Proposed	**0.649**	**0.514**	**0.384**	**0.662**	**0.524**	**0.387**
Dropout+Proposed	0.636	0.519	0.387	0.660	0.532	0.393
SQA-SAM+Proposed	0.648	0.525	0.393	0.655	0.526	0.388

SQA scores and ground truth DSC. Re-segmentation performance is evaluated using the mean Dice coefficient (mDice) before and after the operation.

Table 2. The segmentation quality (mDice) with and without re-segmentation.

Seg. Model	Re-Seg	PolypGen	SUN-SEG
HSNet	✗	0.719 ± 0.421	0.791 ± 0.315
	✓	**0.734 ± 0.419**	**0.804 ± 0.264**
PolypPVT	✗	0.723 ± 0.418	0.777 ± 0.287
	✓	**0.727 ± 0.418**	**0.786 ± 0.279**
CFANet	✗	0.705 ± 0.426	0.722 ± 0.338
	✓	**0.749 ± 0.427**	**0.751 ± 0.315**

3.1 Main Results

Quantitative and Qualitative Results of SQA. We quantitatively evaluated HSNet's segmentation quality on the PolypGen and SUN-SEG datasets using the proposed framework. SQA scores, generated by our framework, were compared against ground truth Dice similarity coefficients (DSC) via correlation analysis. Table 1 compares our framework with established methods, including Softmax entropy [6], Monte Carlo Dropout (various sampling probabilities) [5], Spatial Consistency (SPC, various window sizes) [19], and SQA-SAM [20]. While Softmax, Dropout, and SQA-SAM assess quality at the image level, SPC, like our method, evaluates video-level segmentation quality. Our method demonstrates better performance, achieving Pearson correlation coefficients > 0.6, Spearman rank correlation coefficients > 0.5, and Kendall's tau coefficients > 0.38 on both datasets. These results indicate a statistically significant, moderate-to-strong correlation between predicted SQA scores and ground truth DSC. Integrating

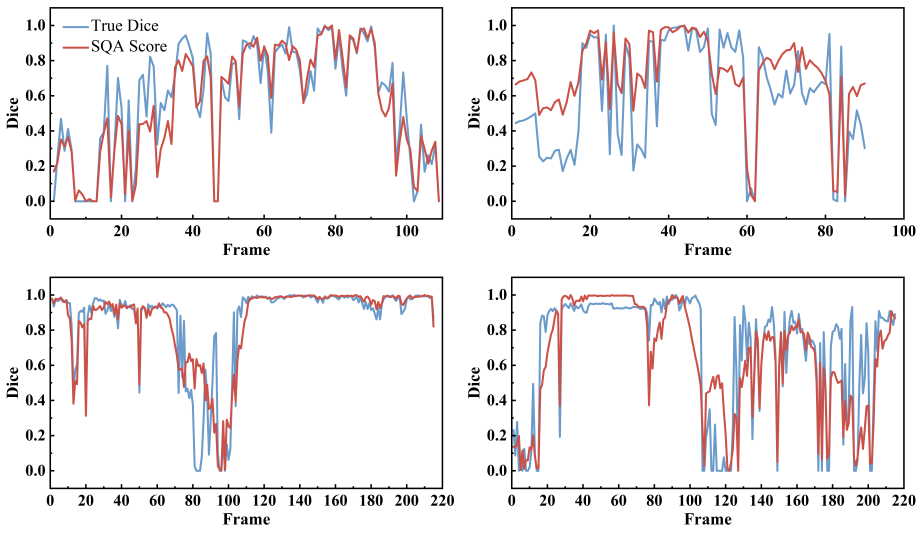

Fig. 2. Visualization of how well the SQA scores (in Dice) from our method align with the true Dice values obtained by comparing the segmentation to the ground truth.

Table 3. Performance of SQA for the proposed method using various options of video segmentation foundation models (VSFMs).

VSFM	PolypGen			SUN-SEG		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
SAM2(large)	0.649	0.514	0.384	0.662	0.524	0.387
SAM2(base+)	0.651	0.520	0.388	0.665	0.529	0.388
SAMURAI	0.646	0.497	0.370	0.627	0.500	0.366

Dropout or SQA-SAM with our method further improved segmentation quality assessment (SQA) performance in certain cases.

Fig. 2 visualizes the relationship between predicted SQA scores and ground truth DSC for representative video sequences. The observed trends and consistent relative magnitudes of SQA scores and DSC corroborate the strong correlative relationship. Notably, SQA scores exhibit a smoother response to abrupt DSC fluctuations, suggesting a potential area for future SQA refinement: capturing rapid temporal changes in segmentation quality.

On Improving Segmentation Performance. To quantify the impact of re-segmentation, we evaluated the initial segmentation performance of HSNet [18], PolypPVT [4], and CFANet [21] on PolypGen and SUN-SEG. Initial performance was measured using the mean Dice similarity coefficient (mDSC) between model predictions and ground truth. Subsequently, we applied our framework, encompassing SQA and re-segmentation. Table 2 presents a comparison of segmentation quality before and after applying the re-segmentation module. This evaluation considers only frames that contain polyps in the ground-truth anno-

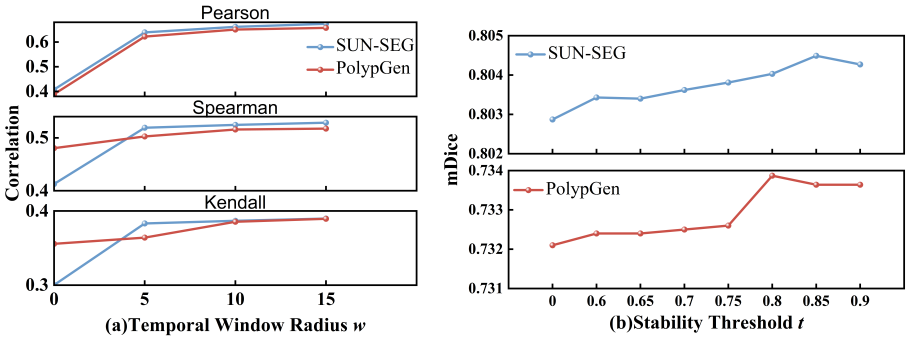

Fig. 3. Ablation studies of temporal window radius w (a) and stability threshold t (b).

tations. The results demonstrate consistent improvements in mDSC across all three SOTA polyp segmentation models, underscoring the effectiveness of the re-segmentation strategy.

On Improving Detection Performance. On the PolypGen dataset, HSNet initially generated 2,778 polyp detection instances. Among these, 1,257 instances (45.2%) were false positives (FPs), primarily due to the presence of 515 negative frames in the dataset. Additionally, 468 instances (16.8%) were false negatives (FNs). Our re-segmentation module effectively corrected 190 FPs (17.8%) and 70 FNs (17.6%), reducing the FP and FN rates to 40.5% and 15.1%, respectively. On the SUN-SEG dataset, HSNet produced 10,616 FPs (18.8%) and 4,652 FNs (8.2%). Through re-segmentation, 1,368 FPs (14.8%) and 534 FNs (13.0%) were corrected, resulting in reduced FP and FN rates of 16.6% and 7.4%, respectively.

3.2 Ablation and Additional Studies

We conduct ablation studies to investigate key hyperparameter influence and framework adaptability. Specifically, we examine the effects of the temporal window radius (w) in the SQA module and the stability threshold (t) in the re-segmentation module, and evaluate the framework's robustness to variations in the underlying foundation segmentation model.

Impact of Temporal Window Radius (w). Fig 3 (a) shows the effect of varying w on SQA module performance. A larger w expands the receptive field, enabling the model to incorporate broader temporal context, generally leading to a more accurate quality assessment. Setting $w = 0$, effectively disabling the temporal component, significantly degrades performance, underscoring spatiotemporal information's crucial role in video segmentation quality assessment.

Impact of Stability Threshold (t). Fig. 3 (b) shows the impact of the stability threshold, t, within the re-segmentation module. A non-zero threshold ($t > 0$) generally leads to better performance than no threshold ($t = 0$).

Working with Different Segmentation Foundation Models. To assess the framework's flexibility, we replaced the default SAM2 (large) foundation model

with SAM2 (base+) and SAMURAI [17], while keeping all other settings unchanged. Table 3 presents the SQA module performance with these different foundation models. The consistent performance across these variations demonstrates the robustness and adaptability of our framework to different underlying segmentation models. This indicates that our approach is not tightly coupled to any specific foundation model and can be easily adapted to incorporate advancements in foundational segmentation techniques.

4 Conclusion

We developed a novel, unsupervised framework for assessing and enhancing the quality of polyp segmentation in colonoscopy videos. By leveraging a video segmentation foundation model, we proposed an innovative quality control and performance enhancement approach that does not require ground-truth annotations or model retraining. Our results demonstrate a valuable correlation between the quality assessment scores and actual segmentation quality, with the resegmentation step further improving performance based on the SQA scores. The synergy between specialized polyp segmentation models and general-purpose foundation models like SAM2 opens new avenues for scalable, robust, and adaptable solutions in medical video analysis.

Acknowledgments. This research was supported in part by the Natural Science Foundation of Jiangsu Province (Grant BK20220949), National Natural Science Foundation of China (Grant 62201263, 62172228) and Major Research Plan of the National Natural Science Foundation of China (Grant 92370109).

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

References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023)
2. Ali, S., Zhou, F., Bratt, R., DaCosta, R., Shine, J., Dolan, K., Min, M., Beggs, A.D., Lovat, L.B., Tufano, A., et al.: Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment. arXiv preprint arXiv:2109.08565 (2021)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE transactions on medical imaging **36**(6), 1231–1249 (2017)
4. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)

5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
7. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
8. Kim, N.H., Jung, Y.S., Jeong, W.S., Yang, H.J., Park, S.K., Choi, K., Park, D.I.: Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. Intestinal research **15**(3), 411 (2017)
9. Leufkens, A., Van Oijen, M., Vleggaar, F., Siersema, P.: Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy **44**(05), 470–475 (2012)
10. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Ogawa, T., Maeda, Y., Oda, M., Mori, K.: Development and evaluation of an automated detection system for early gastric cancer in endoscopic video using deep learning. In: Gastrointestinal Endoscopy. vol. 93, p. AB109. Elsevier (2021)
11. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Ogawa, T., Maeda, Y., Oda, M., Mori, K.: Development of a computer-aided detection system for colonoscopy using a large dataset of various types of polyps (with video). In: Gastrointestinal Endoscopy. vol. 93, p. AB307. Elsevier (2021)
12. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy **93**(4), 960–967 (2021)
13. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
14. Sanderson, E., Matuszewski, B.J.: Fcn-transformer feature fusion for polyp segmentation. In: Annual conference on medical image understanding and analysis. pp. 892–907. Springer (2022)
15. Sharma, V., Kumar, A., Jha, D., Bhuyan, M.K., Das, P.K., Bagci, U.: Controlpolypnet: towards controlled colon polyp synthesis for improved polyp segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2325–2334 (2024)
16. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
17. Yang, C.Y., Huang, H.W., Chai, W., Jiang, Z., Hwang, J.N.: Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv preprint arXiv:2411.11922 (2024)
18. Zhang, W., Fu, C., Zheng, Y., Zhang, F., Zhao, Y., Sham, C.W.: Hsnet: A hybrid semantic network for polyp segmentation. Computers in biology and medicine **150**, 106173 (2022)
19. Zhang, Y., Borse, S., Cai, H., Wang, Y., Bi, N., Jiang, X., Porikli, F.: Perceptual consistency in video segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2564–2573 (2022)

20. Zhang, Y., Wang, S., Zhou, T., Dou, Q., Chen, D.Z.: Sqa-sam: Segmentation quality assessment for medical images utilizing the segment anything model. arXiv preprint arXiv:2312.09899 (2023)
21. 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**, 109555 (2023)