




HalF-SAM: SAM-based Haustral Fold Detection In Colonoscopy with Debris Suppression and Temporal Consistency

Mayank Golhar^{*} , Luojie Huang^{*} , and Nicholas J. Durr[✉] 

Department of Biomedical Engineering, Johns Hopkins University, USA
ndurr@jhu.edu

Abstract. Haustral folds can serve as important landmarks to localize and navigate colonoscopes through the colon. Fold edges can be utilized for tracking in 3D reconstruction algorithms to generate colonoscopy coverage maps and ultimately reduce missed lesions. Current haustral fold detection models struggle with debris-filled colonoscopy videos and fail to maintain high temporal consistency due to their single-frame input. We introduce HalF-SAM, a Haustral Fold detection model utilizing the Segment Anything Model (SAM) image encoder, which suppresses edges from specular reflection and fecal debris. The SAM2-based memory module enhances temporal consistency, which is essential for tracking. Our experiments have shown significant improvements in haustral fold extraction accuracy and stability. We also release a training dataset with automatically annotated haustral fold edges in debris-filled high-fidelity colon phantom videos. The dataset and code will be available at: <https://github.com/DurrLab/HalFSAM>.

Keywords: Colonoscopy · Edge Detection · Segment Anything Model.

1 Introduction

Colorectal cancer is the second leading cause of cancer deaths in the United States [1]. Colonoscopy remains the gold standard for screening and detecting colorectal cancer. However, studies utilizing tandem colonoscopy have shown that approximately 25% of lesions go undetected, highlighting the need for enhanced diagnostic quality of colonoscopy [2].

One primary reason for undetected polyps is ‘missed regions’ during colonoscopic surveillance. Researchers have proposed employing 3D reconstruction methods, such as Simultaneous Localization and Mapping (SLAM), to generate a topography of colon lumen, where “holes” would represent missed regions [3]. Unfortunately, traditional 3D reconstruction algorithms using standard feature tracking methods have exhibited poor performance in colonoscopy. Challenging factors include the textureless and highly specular nature of the colon surface, as

^{*} Equal Contribution

well as dynamic imaging conditions caused by artifacts such as debris, bubbles, and mucus, along with the deformation of the colon itself [4]. An alternative strategy potentially improves tracking by utilizing haustral folds—protrusions of the colon walls which act as reliable anatomical landmarks [5]. Haustral fold edges remain relatively stable and distinct across frames, potentially serving as reliable replacements for conventional feature points. Haustral folds can also aid in determining the centerline of the colon lumen, offering valuable information for navigating autonomous endoscopes [6]. Additionally, these folds represent key points of depth discontinuities, and incorporating them as priors might enhance depth estimation models [7]. Overall, haustral folds can play a critical role as landmarks to assist with localization, tracking, and navigation during colonoscopic procedures.

In this work, we present HalF-SAM, a novel model for haustral fold edge detection in colonoscopy videos, along with a synthetic data set incorporating challenging colonoscopy conditions. Our contributions are as follows.

1. **HalF-SAM Model:** A new haustral fold edge detection model that incorporates the powerful SAM (Segment Anything Model) [8,9] image encoder and memory into the DexiNed [10] framework. The model enhances edge detection of haustral folds while reducing false positives from debris and improving temporal consistency, even in cases with poor bowel preparation.
2. **Multi-layer Memory Mechanism and Temporal Consistency:** Our model processes video sequences rather than individual frames, incorporating a multi-layer memory mechanism to enhance temporal consistency in haustral fold detection. This approach could improve performance in downstream tasks such as longitudinal fold tracking for 3D reconstruction.
3. **Haustral Fold Edge Detection Dataset:** A comprehensive dataset of 113 videos featuring high-fidelity silicone colon phantoms. It includes challenging conditions like fecal debris and dirty optics, aimed at improving algorithms for reducing false positives in edge detection.

2 Related Works

Haustral Fold Detection Methods Early works primarily relied on traditional edge extraction techniques, often leveraging hand-crafted geometric and curvature constraints for identifying haustral folds [11,12]. In contrast, recent deep learning-based methods don’t require explicit geometric priors, potentially improving generalizability across different haustral fold shapes and imaging conditions. FoldIt model by Matthew et al. [13] utilized a generative adversarial network to translate optical colonoscopy images into virtual colonoscopy images with haustral fold overlays. They approached the task as a haustral fold semantic segmentation problem. However, some limitations of this model included temporal inconsistencies in segmentations, subpar performance in the presence of colonic debris, and ambiguous haustral fold boundary annotations used during training.

Jin et al. [5] proposed a self-supervised method for haustral fold detection. Their approach involves generating pseudo-labels using the DexiNed edge detection model, coupled with temporal inpainting to address specular reflections. These pseudo-labels are then utilized to retrain the DexiNed model, incorporating a triplet loss to enhance the temporal consistency of edge detection. However, this model was primarily trained and tested on cases with clean bowel preparation, resulting in a decline in performance under challenging conditions. Additionally, the temporal consistency is upper-bounded since the model’s input is a single image. In contrast, HalF-SAM takes video as input and was trained on a dataset containing fecal debris.

Haustral Fold Detection Datasets A significant bottleneck in the development of haustral fold detection models is the lack of substantial datasets. Jin et al. [5] released a pseudo-labeled training dataset comprising 31 videos. However, since the temporal inpainting method primarily addresses the artifact of specular reflection, the resulting pseudo-labels contain false positive edges attributed to debris, bubbles, and other factors. Recently, the SegCol challenge [14] introduced a subtask focused on haustral fold detection, which utilized a dataset of 8,440 images. Unfortunately, access to this dataset was restricted to challenge participants. Our dataset can complement the SegCol dataset by increasing its quantity and comprehensiveness, particularly in cases involving challenging debris-filled environments.

3 HalF-SAM

We propose HalF-SAM, a novel edge detection framework designed for detecting haustral folds in colonoscopy videos. The model overview is shown in Fig. 1. Our approach leverages the powerful pre-trained Segment Anything Model 2 (SAM2) [9] as the frame encoder backbone, while incorporating trainable adaptors for domain adaptation. Compared to existing edge extraction models, the proposed HalF-SAM takes videos as inputs and enhances the temporal consistency of haustral fold detection with the novel multi-layer memory mechanism. This feature is particularly beneficial for downstream tasks such as longitudinal fold tracking for full colon 3D reconstruction.

In the following sections, we elaborate on the adaptation of the SAM2 encoder for colonoscopy videos in Sec. 3.1, the multi-layer memory mechanism in Sec. 3.2, and the improved sequence loss for video-based edge detection in Sec. 3.3.

3.1 Adapted SAM2 Encoder for Colonoscopy Videos

Due to its large-scale pre-training on real-world video datasets, SAM2 is optimized for sequential inputs and has demonstrated superior performance in object and scene understanding. Its encoder has been successfully generalized to various domains, particularly in different medical tasks. However, a significant challenge in adapting SAM2 is the substantial computational burden resulting from its

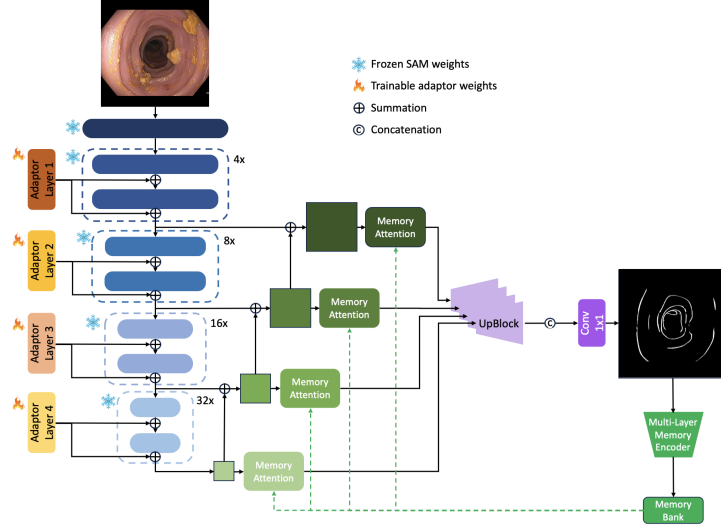


Fig. 1. HalF-SAM Overview. The proposed model follows an encoder-decoder architecture, consisting of three major modules: 1) SAM-based frame encoder with adaptors; 2) Haustral fold decoder with multi-layer Upblocks; 3) Multi-layer edge memory attention modules.

large model size. To achieve efficient SAM2 adaptation on small datasets, the SAM-adaptor has been proposed [15], and demonstrated impressive performance on various tasks. Inspired by this work, we employ the frozen SAM encoder, pre-trained Hiera [16], as our backbone and integrate the trainable adaptor layers for domain-specific fine-tuning of SAM2 for haustral fold detection. As illustrated in Fig. 1, we incorporate four adaptor layers (Adaptor Layer 1–4), injecting fine-tuning weights A_{0-3} at different backbone blocks with downsampling scales from 4 to 32.

For each colonoscopy frame input $I_t \in \mathbb{R}^{H \times W \times 3}$, the adapted encoder outputs four layers of feature maps $F_{k \times} \in \mathbb{R}^{H/k \times W/k \times D}$ that represent features extracted at different resolutions:

$$F_{4 \times}, F_{8 \times}, F_{16 \times}, F_{32 \times} = \text{Hiera}(I_t; A_{0-4})$$

To generate final haustral fold detection, we apply the UpBlocks, similar to US-Net in DexiNed, to upsample feature maps to the original resolution $Up(F_{k \times}) \in \mathbb{R}^{H \times W \times 1}$ through deconvolution layers. And then the upsampled features are fused to generate the edge prediction $\mathcal{E}(I_t) \in \mathbb{R}^{H \times W \times 1}$. The edge decoding process could be expressed as:

$$\mathcal{E}(I_t) = \text{Conv}_{1 \times 1}(\text{Conct}[Up(F_{4 \times}), Up(F_{8 \times}), Up(F_{16 \times}), Up(F_{32 \times})])$$

3.2 Multi-layer Memory Mechanism

As HalF-SAM takes colonoscopy videos as input, we introduce a temporal memory mechanism that propagates across video frames to enhance feature learning. Inspired by SAM2, this mechanism addresses the challenge of inconsistent edge predictions across frames, due to debris occlusion, reflection, and motion blurs, by cross-attending current features with edge memory from nearby frames.

After the model extracts the edges at each timepoint, we use a memory encoder from SAM2, $\mathcal{M}()$, to encode the current prediction into edge memory. Since the haustral fold detection fuse features from different levels of resolution, we modify the memory encoder to generate multi-level memories $M_{k\times} \in \mathbb{R}^{H/k \times W/k \times D_m}$ to match the features:

$$M_{4\times}^t, M_{8\times}^t, M_{16\times}^t, M_{32\times}^t = \mathcal{M}(\mathcal{E}(I_t))$$

The encoded memories are then stored in the memory bank, a First-In, First-out queue. To save space for long video inference, the memory bank is limited to storing memories up to N recent frames. When the memory bank is full, the earliest memories would be discarded.

When the memory bank is not empty at a timepoint t_p , the model would retrieve all memories from the memory bank. The memories are then stacked within each level to $M_{k\times}^{t_p-1 \sim t_p-n-1} \in \mathbb{R}^{n \times H/k \times W/k \times D_m}$. Then, the retrieved memories are used to condition current features, before haustral fold decoding, with vanilla attention of alternative self- and cross-attention [17]:

$$F_{k\times}^{att} = \text{Att}(F_{k\times}^{t_p}, M_{k\times}^{t_p-1 \sim t_p-n-1}) \in \mathbb{R}^{H/k \times W/k \times D}$$

Memory-conditioned features are fused to the output edge map prediction through the same decoder.

3.3 Sequence Loss for Video-based Haustral Fold Detection

We supervised our network using the BDCN loss [18], a powerful loss function for highly imbalanced detection tasks, such as edge extraction, between the predicted and ground-truth values. Given ground truth \mathcal{E}_{gt} , the loss is calculated including all upsampled maps and final fused edge prediction:

$$\mathcal{L}_{multilevel}(\mathcal{E}(I_i), \mathcal{E}_{gt}) = \sum_k^{4,8,16,32} w_k \mathcal{L}_{BDCN}(\text{Up}(F_{k\times}), \mathcal{E}_{gt}) + w_0 \mathcal{L}_{BDCN}(\mathcal{E}(I_i), \mathcal{E}_{gt})$$

To ensure our memory mechanism works effectively for the input sequence, we calculate sequence loss over the full sequence of N predictions, with exponentially increasing weights ($\gamma = 0.8$). The sequence loss is defined as

$$\mathcal{L}_{seq} = \sum_{i=1}^N \gamma^{N-i} \mathcal{L}_{multilevel}(\mathcal{E}(I_i), \mathcal{E}_{gt})$$

4 Experiments & Results

4.1 Dataset

We present a synthetic dataset for haustral fold detection that incorporates challenging conditions commonly encountered in real-world colonoscopy procedures [19]. This dataset features scenarios such as the presence of fecal debris, mucous pools, bubbles, rapid camera movements, and dirty lenses with lens cleaning events. To create this dataset, we followed the procedure described by Bobrow et al. [20], utilizing high-fidelity silicone colon phantoms to produce pixel-wise paired RGB, depth, and flow frames.

Our data collection process involved recording two videos of the same phantom using identical camera trajectories and settings: one clean colon and another filled with debris. After synchronization, we obtained paired colonoscopy frames corresponding to clean and debris-filled colon conditions. We then employed DexiNed to extract edge frames from the clean colon RGB frames, debris-filled colon RGB frames, and the registered depth frames. Since the depth frames primarily contain edges from haustral folds, we used them as region of interest (ROI) masks. Only edges present in all three images were retained, as they likely belong to haustral folds. The dataset collection protocol is described in detail in Golhar et al. [19].

Our dataset consists of 113 videos, with each video averaging 424 frames (standard deviation: 175). We divided the dataset into training, validation, and test splits, comprising 74, 15, and 24 videos, respectively. Each split contains paired videos, with half of the videos depicting a clean colon and the other half containing debris. Each debris-filled video contains fecal debris across all frames. The dataset encompasses a range of haustral fold shapes (e.g., triangular, elliptical, and irregular) and various camera poses (parallel, normal, oblique, and laterally shifted relative to the centerline).

4.2 Evaluation Metrics

We frame haustral fold detection as an edge detection problem. We use standard metrics prevalent in the field - Optimal Dataset Scale (ODS), Optimal Image Scale (OIS), and Average Precision (AP) to assess the performance of our edge detection models. ODS measures the best F-score achieved at a fixed threshold across the entire dataset, while OIS reports the average best F-scores for each individual image at an optimal threshold. Higher ODS and OIS indicate the model has stronger edge extraction performance. And OIS is always higher than ODS, since it allows flexible thresholds. If OIS is significantly higher than ODS, it indicates that the algorithm requires more fine-tuning for each image, thereby reducing its generalization ability and stability. On the other hand, if OIS is roughly equal to ODS, the algorithm is robust and works well without much threshold adjustment. AP evaluates the trade-off between precision and recall across different thresholds.

4.3 Implementation

HalF-SAM is trained on four NVIDIA RTX A5000 GPUs. During training, we set the input sequence length to 5 frames and the number of memories to 4 frames. During inference, we input the whole video frame by frame, and the number of memories is set to 4. The model is trained using the Adam optimizer with an initial learning rate of $1e-4$, which is decreased by a factor of 0.1 at epochs 10 and 15. The weight decay for training was set to $1e-8$. For the BDCN loss, we set the layer weights $[w_0, w_4, w_8, w_{16}, w_{32}]$ to $[1.5, 0.7, 1.1, 0.7, 0.3]$. For comparative experiments, we adapt the current SOTA haustral fold detection models- DexiNed and the Self-Supervised DexiNed model by Jin et al. [5], on the proposed dataset. We do not present results from the FoldIt model, as it produces a broad haustral fold segmentation instead of thin edges, as discussed by Jin et al. [5].

Quantitative Results of Comparative Experiments Table 1 illustrates the superior performance of the Half-SAM model when compared to current SOTA models. Notably, the Self-Supervised DexiNed model struggles to adapt to our dataset, likely due to its training on relatively clean bowel cases. The videos with debris introduce a new challenge to haustral fold detection. On the other hand, supervised DexiNed demonstrated reasonable performance on our dataset. Compared to DexiNed, HalF-SAM showed significant improvements in ODS and OIS, suggesting more accurate segmentation in both clean video and videos with debris. The lower OIS-ODS of HalF-SAM indicates a stable edge extraction from the video dataset. However, HalF-SAM yields slightly lower AP, due to the thicker edge predictions, which have little effect for downstream applications.

Table 1. Quantitative Results for Haustral Fold Edge Detection

Method	Test set	ODS (\uparrow)	OIS (\uparrow)	OIS-ODS (\downarrow)	AP (\uparrow)
Self-Supervised DexiNed [5]	C+D	0.314	0.325	0.011	0.259
DexiNed [10]	C+D	0.487	0.552	0.065	0.404
Frozen SAM2 Encoder + Decoder	C+D	0.533	0.572	0.039	0.484
HalF-SAM w/o Memory	C+D	0.539	0.581	0.042	<u>0.475</u>
HalF-SAM	C	0.883	0.883	<0.001	0.397
HalF-SAM	D	0.852	0.852	<0.001	0.374
HalF-SAM	C+D	<u>0.867</u>	<u>0.867</u>	<0.001	0.388

* **C+D**: complete test dataset with both clean videos and videos with debris; **C**: only clean videos; **D**: only videos with debris.

* **Bold** numbers indicate the best performance; Underscored numbers indicate the second best performance.

4.4 Experimental Results

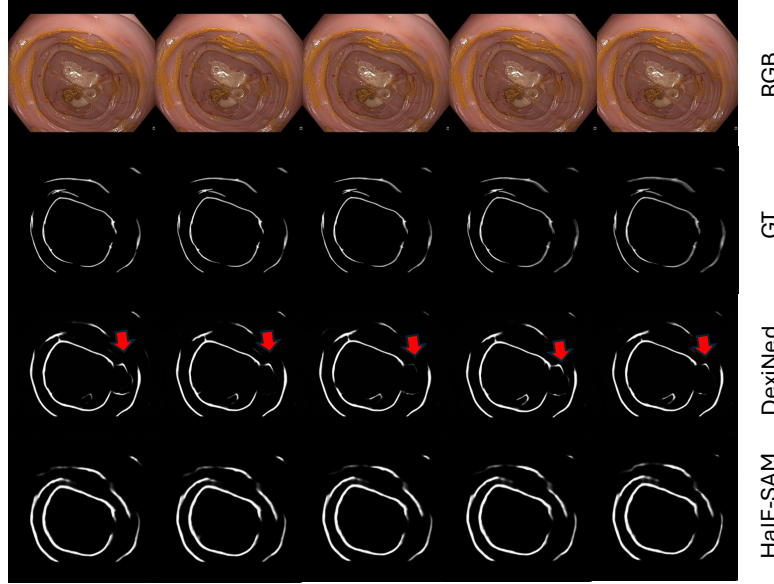


Fig. 2. Qualitative results showing HalF-SAM debris suppression and temporal consistency. Notice the edge flickering in DexiNed edges.

Ablation Study The second section of Table 1 shows the results of ablation study. The pre-trained SAM2 encoder enhances feature representation, resulting in significant improvement compared to DexiNed. The SAM adaptor enabled the SAM2 encoder to achieve domain adaptation, further boosting the accuracy. The HalF-SAM incorporates a memory mechanism to achieve better edge detection accuracy and stability.

Qualitative Comparison Figure 2 presents a qualitative comparison between the proposed method and other edge detection models. Our observations indicate that our model effectively reduces false-positive edges caused by colonic artifacts, such as debris and specular reflections. In addition to training on debris-filtered edges, a possible reason for this improvement is the enhanced semantic understanding of scenes provided by the SAM image encoder. Furthermore, we note that our model generates temporally consistent edges across frames, compared to other models, which could be attributed to the integration of a memory module.

5 Conclusion

In this paper, we introduced HalF-SAM, a novel model that incorporates the powerful Segment Anything Model 2 image encoder and multi-layer memory attention modules for haustral fold edge detection in colonoscopy videos. Our approach demonstrates robust performance in challenging conditions, including the presence of debris and dirty optics, while generating temporally consistent edges. This advancement opens new possibilities for using haustral fold edge tracking instead of point-wise feature tracking in colonoscopy 3D reconstruction and navigation. To support further research in this area, we will make available a challenging haustral fold detection dataset.

Future work will focus on adapting the model for in vivo colonoscopy data to enhance its clinical applicability. Previous work has adapted optical flow estimation models for quantitative evaluation of temporal consistency [5]. We plan to evaluate the potential for performance improvements in downstream tasks such as 3D reconstruction and depth estimation using the detected haustral fold edges and the optical flow maps collected along with the video dataset.

Acknowledgments. This work was partially supported by Catalyst Award funding from Johns Hopkins University.

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

References

1. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2021. CA: A Cancer Journal for Clinicians **71**(1), 7–33 (2021). <https://doi.org/https://doi.org/10.3322/caac.21654>, <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21654>
2. Zhao, S., Wang, S., Pan, P., Xia, T., Chang, X., Yang, X., Guo, L., Meng, Q., Yang, F., Qian, W., et al.: Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis. *Gastroenterology* **156**(6), 1661–1674 (2019)
3. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. *Medical image analysis* **72**, 102100 (2021)
4. Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J.: A deep learning framework for quality assessment and restoration in video endoscopy. *Medical image analysis* **68**, 101900 (2021)
5. Jin, W., Daher, R., Stoyanov, D., Vasconcelos, F.: A self-supervised approach for detecting the edges of haustral folds in colonoscopy video. In: *MICCAI Workshop on Data Engineering in Medical Imaging*. pp. 56–66. Springer (2023)
6. Li, Y., Ng, W.Y., Sun, Y., Huang, Y., Li, J., Chiu, P.W.Y., Li, Z.: Colon lumen center detection enables autonomous navigation of an electromagnetically actuated soft-tethered colonoscope. *IEEE Transactions on Instrumentation and Measurement* (2024)

7. Chen, R.J., Bobrow, T.L., Athey, T., Mahmood, F., Durr, N.J.: Slam endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv:1907.00283 (2019)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
9. 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)
10. Soria, X., Riba, E., Sappa, A.: Dense extreme inception network: Towards a robust cnn model for edge detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1912–1921. IEEE Computer Society, Los Alamitos, CA, USA (mar 2020). <https://doi.org/10.1109/WACV45572.2020.9093290>, <https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093290>
11. Hong, D., Tavanapong, W., Wong, J., Oh, J., De Groen, P.C.: Colon fold contour estimation for 3d visualization of colon structure from 2d colonoscopy images. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 121–124. IEEE (2011)
12. Hong, D., Tavanapong, W., Wong, J., Oh, J., De Groen, P.C.: 3d reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics* **38**(1), 22–33 (2014)
13. Mathew, S., Nadeem, S., Kaufman, A.: Foldit: Haustral folds detection and segmentation in colonoscopy videos. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 221–230. Springer (2021)
14. Ju, X., Daher, R., Caramalau, R., Huang, B., Stoyanov, D., Vasconcelos, F.: Segcol challenge: Semantic segmentation for tools and fold edges in colonoscopy data (2024), <https://arxiv.org/abs/2412.16078>
15. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3367–3375 (2023)
16. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: International conference on machine learning. pp. 29441–29454. PMLR (2023)
17. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)
18. He, J., Zhang, S., Yang, M., Shan, Y., Huang, T.: Bi-directional cascade network for perceptual edge detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3828–3837 (2019)
19. Golhar, M.V., Galeano Fretes, L.S., Ayers, L., Akshintala, V.S., Bobrow, T.L., Durr, N.J.: C3vdv2 - colonoscopy 3d video dataset with enhanced realism. arXiv preprint arXiv:2506.24074 (2025)
20. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis* **90**, 102956 (2023)