BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Interactive Polyp Segmentation in Images and Videos

Bc. Eva Mičánková*

Abstract

Colorectal cancer is one of the most common malignancies worldwide, with early detection being critical for successful treatment. This work focuses on the development of interactive polyp segmentation and tracking models for both images and videos to assist real-time colonoscopy procedures and efficient data annotation.

The proposed solution involves training classical segmentation models (U-Net, U-Net++) and fine-tuning the recently released foundational model SAM2 on proprietary medical data. For image segmentation, different image encoders were evaluated, while for video tracking, a specialized fine-tuning strategy was designed to handle sparse ground-truth annotations. Performance is assessed using standard segmentation metrics such as IoU and F-score.

The results show that classical models and SAM2 achieve comparable performance on public datasets, reaching an F-score of 95% and an IoU of approximately 92%. Fine-tuning SAM2 demonstrates its flexibility to adapt from open-world segmentation to specialized medical applications.

This work highlights the potential of adapting strong foundational models like SAM2 for the medical imaging domain, inspiring further research. Moreover, the developed models are planned to be integrated into MAIA Labs annotation tools and can assist in future downstream tasks.

*xmican10@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Colorectal cancer is the second most common malignancy in Europe, the third worldwide, and also the third most frequent cancer in the Czech Republic. Early detection significantly improves the chances of successful treatment, which is why a nationwide colorectal cancer screening program has been implemented in the Czech Republic since 2014. Colonoscopy remains the most reliable method for early detection, as it enables both identification and removal of precancerous polyps. To support real-time assistance during colonoscopy and efficient annotation of new data, interactive polyp segmentation and tracking models are being developed, enabling also downstream tasks such as polyp size measurement and long-term tracking and classification.

This work addresses the problem of interactive colorectal polyp segmentation in both static images and video sequences. The goal is to develop methods that can accurately delineate the shape and location of polyps in both images and video sequences from the colon and rectum. A reliable segmentation model should be robust across different imaging conditions, generalize well to unseen data, and provide precise boundaries around the polyps. The performance of the solution will be evaluated using standard metrics such as Intersection over Union (IoU) and the Dice coefficient.

The proposed solution involves training various segmentation models for interactive image segmentation and video tracking using data provided by MAIA Labs[1]. The work compares a classical approach, based on models like U-Net and U-Net++, with the fine-tuning of the recently released foundation model SAM2, originally developed for open-world segmentation tasks.

In this work, models were trained on proprietary data, focusing on interactive polyp segmentation in images and videos. The recently released SAM2 model was fine-tuned for these specialized tasks, demonstrating its potential as a strong foundational model that can be adapted far beyond open-world segmentation. The results highlight the flexibility and robustness of the SAM2 architecture, suggesting that it can

[1] https://maia-labs.com/

be effectively fine-tuned for medical imaging applications as well, potentially inspiring further research in this direction. Moreover, the developed models are planned to be integrated into MAIA Labs' annotation tools and can support various downstream tasks, such as automated dataset creation and model-assisted labeling.

2. Proposed method

The proposed solution consists of two parts: interactive polyp segmentation in images and interactive polyp segmentation and tracking in video sequences.

For interactive image segmentation, a classical approach based on U-Net [1] and U-Net++ [2] architectures from Segmentation Models PyTorch [3] was first employed. To maximize data utilization, randomized online augmentations of both the dataset and prompts were applied during training. The training dataset contains 19,303 images. The models take a 4-channel input — the RGB image concatenated with a binary mask representing the user-provided bounding box. Various image encoders were experimented with during development; however, the MiT-B2 [4] encoder for U-Net and the ResNet34 [5] encoder for U-Net++ achieved the best performance. Both encoders are pretrained on ImageNet.

Subsequently, the Segment Anything Model 2 (SAM2) [6] was fine-tuned under the same conditions. SAM2 takes the RGB image as input, along with bounding box coordinates provided to the prompt encoder. For fine-tuning, the `sam2.1_hiera_tiny.pt` checkpoint was used. Various fine-tuning configurations were experimented with, and the best results were achieved when all model components — the image encoder, prompt encoder, and mask decoder — were fine-tuned together.

For interactive segmentation across videos, a traditional approach combining polyp detection (or a user-provided bounding box), a tracking algorithm, and the previously developed interactive segmentation models is compared to fine-tuning SAM2 for interactive video polyp segmentation and tracking. The SAM2 model incorporates memory modules that handle object tracking across frames. For fine-tuning, the `sam2.1_hiera_tiny.pt` checkpoint was used. The fine-tuning process is specifically tailored to the provided dataset, which contains sequences with sparse polygon mask annotations. During training, the model is given an initial bounding box, and the entire video sequence is processed. The loss is computed only on the last frame of the sequence, where the corresponding polyp segmentation mask is available.

3. Results

For the image models, the results were very close to each other; all models achieved an F-score of 95% and an IoU of approximately 92%, which suggests that the models are limited by the precision of the training dataset. The models were evaluated on the public Kvasir-SEG [7] and CVC-ColonDB datasets [8]. The best performance was achieved by U-Net, closely followed by SAM2.

For video evaluation, I am currently creating a testing dataset by manually annotating video sequences from colonoscopy recordings. The classical approach will be compared to the SAM2 video model by computing the IoU and F-score metrics over time, in order to evaluate how well the methods are able to track and segment polyps.

4. Conclusions and future work

This work focuses on developing and evaluating interactive polyp segmentation models for both images and videos.

Currently, the evaluation of video segmentation models is in progress, based on a manually annotated testing dataset. Future work will include a detailed comparison between the classical approach, combining detection, tracking, and segmentation, and the fine-tuned SAM2 model for video tracking and segmentation. The performance will be assessed by analyzing IoU and F-score metrics over time to evaluate tracking stability and segmentation accuracy.

The developed models are planned to be integrated into the MAIA Labs annotation platform, where they will assist in interactive labeling, automated dataset creation, and other downstream tasks, ultimately contributing to improved efficiency and quality in the development of diagnostic tools.

Acknowledgements

I would like to thank my supervisor, Ing. Michal Hradiš, Ph.D., for his guidance and support, as well as MAIA Labs for their collaboration and assistance.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, 2020.

[3] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel-org/segmentation_models.pytorch, 2020. Accessed: 2025-04-26.

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[6] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[7] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[8] Joan Bernal, Javier Sánchez, and Francesc Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.