INTERACTIVE POLYP SEGMENTATION IN IMAGES AND VIDEOS

Author: Bc. Eva Mičánková
xmican10@stud.fit.vutbr.cz

Supervisor: Ing. Michal Hradiš, Ph.D.

MOTIVATION

Colorectal cancer is the second most common malignancy in Europe, the third worldwide, and the third most frequent cancer in the Czech Republic.

TECHNICAL MOTIVATION

Interactive polyp segmentation and tracking models are crucial for:

- Real-time assistance during colonoscopy
- Efficient annotation of new data

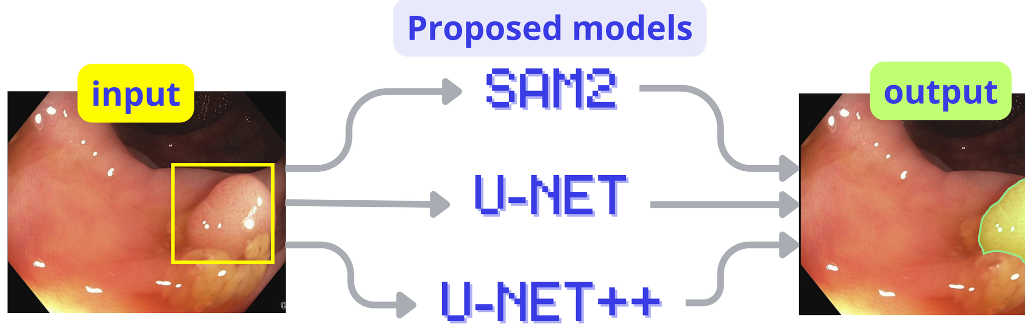
These models also enable downstream tasks such as:

- Polyp size measurement
- Long-term tracking & classification

IMAGES

Introduction

- The following models take an image and a user-provided bounding box as input, and predict a binary segmentation mask that accurately outlines the shape of the polyp.



All models were trained under the following conditions:

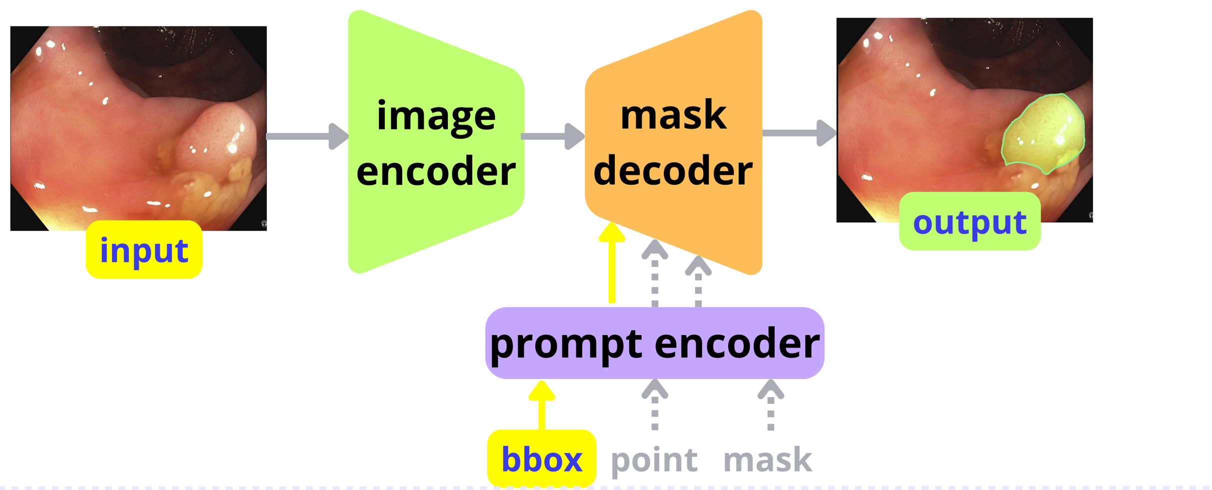
- Training dataset size: 19,303 images
- Image resolution: 512 x 512 px
- Augmentation: geometric and photometric transformations
- Bounding box augmentations were applied

SAM2 Meta

- Open-source model developed by Meta AI, released in August 2024.
- It is an open-world foundational model designed for interactive segmentation and tracking across images and videos.
- Pre-trained SAM2 checkpoint was fine-tuned for this task.

Technical Information

- Model Encoder: Hiera-t
- All model parts were fine-tuned (image encoder, mask decoder and prompt encoder)
- Trainable parameters: 38,962,498
- GPU used: NVIDIA A100 (40 GB RAM)
- Training iterations: 200,000
- Total training time: 1d 2h 52min 36s



U-Net & U-Net++

- Open-source architectures originally developed for biomedical image segmentation, first introduced in 2015.
- U-Net features a symmetric encoder-decoder structure with skip connections to enable precise localization.
- U-Net++ extends this design with nested and dense skip connections, aiming to bridge semantic gaps and improve segmentation accuracy.
- Both models use an ImageNet-pretrained encoder to enhance feature extraction.

U-Net technical Information

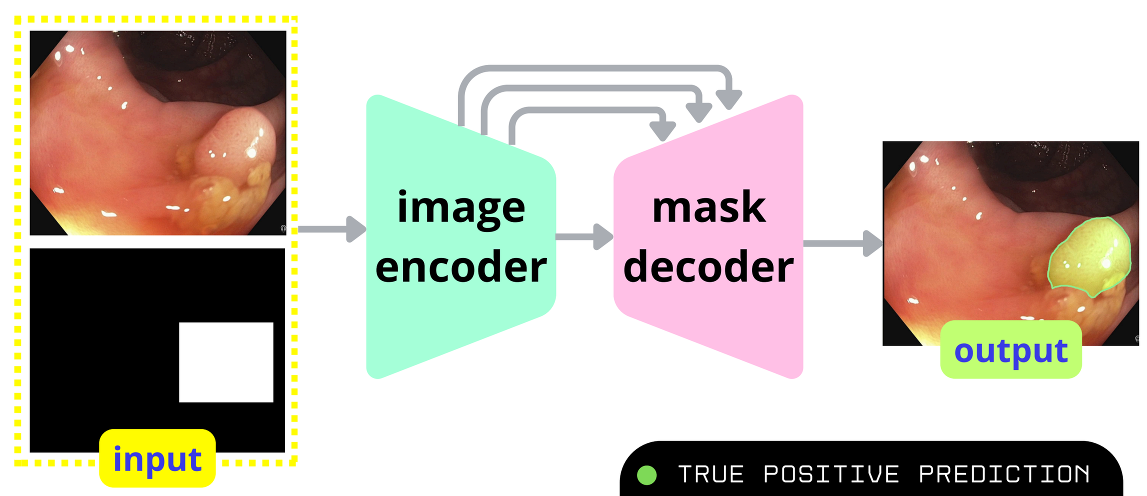
- Model Encoder: MiT-B2
- Trainable parameters: 27,605,400
- Total training time: 13h 34min 46s

U-Net++ technical Information

- Model Encoder: ResNet34
- Trainable parameters: 26,284,468
- Total training time: 22h 8min 39s

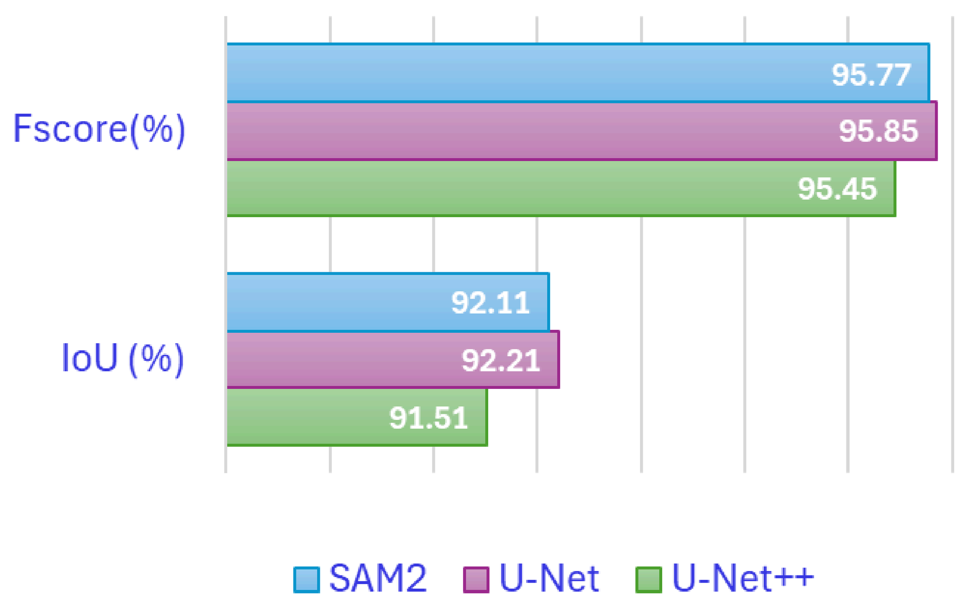
Training setup

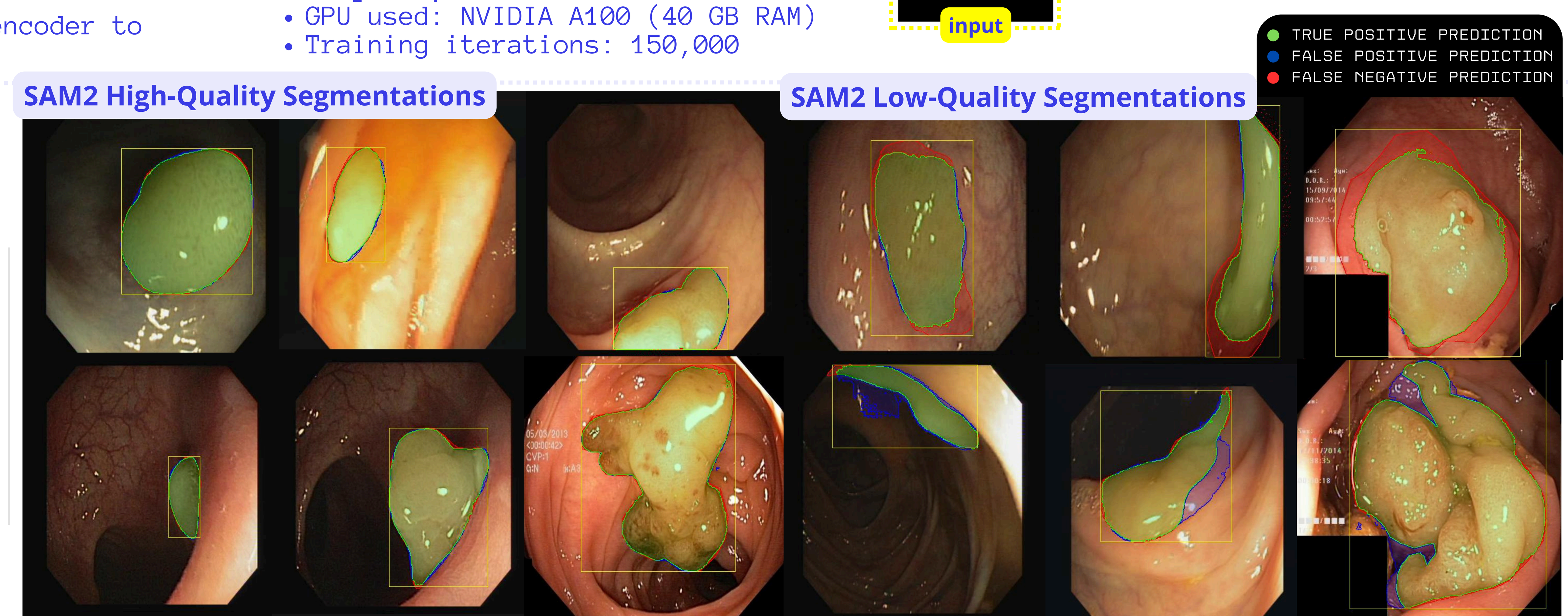
- GPU used: NVIDIA A100 (40 GB RAM)
- Training iterations: 150,000



Results

- Evaluation was performed on the Kvasir-SEG and CVC-ColonDB datasets to assess segmentation performance.





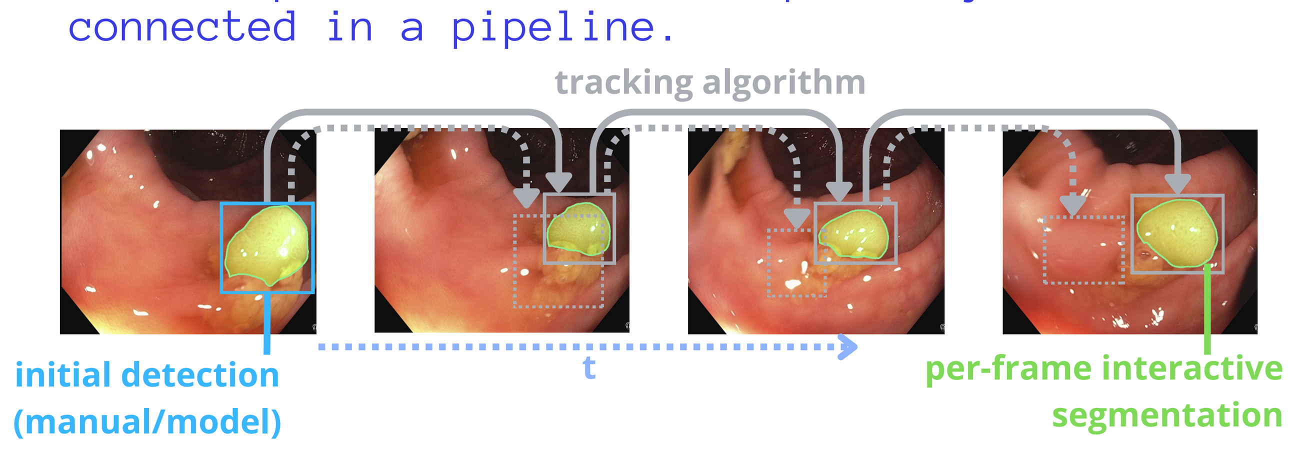
VIDEOS

Introduction

- In this section, we evaluate how SAM2 performs as an end-to-end solution compared to traditional approaches using detection, tracking, and segmentation.

Traditional method

- A multi-stage system combining a polyp detection model, a tracking algorithm and interactive image segmentation models.
- Each component is trained separately and connected in a pipeline.



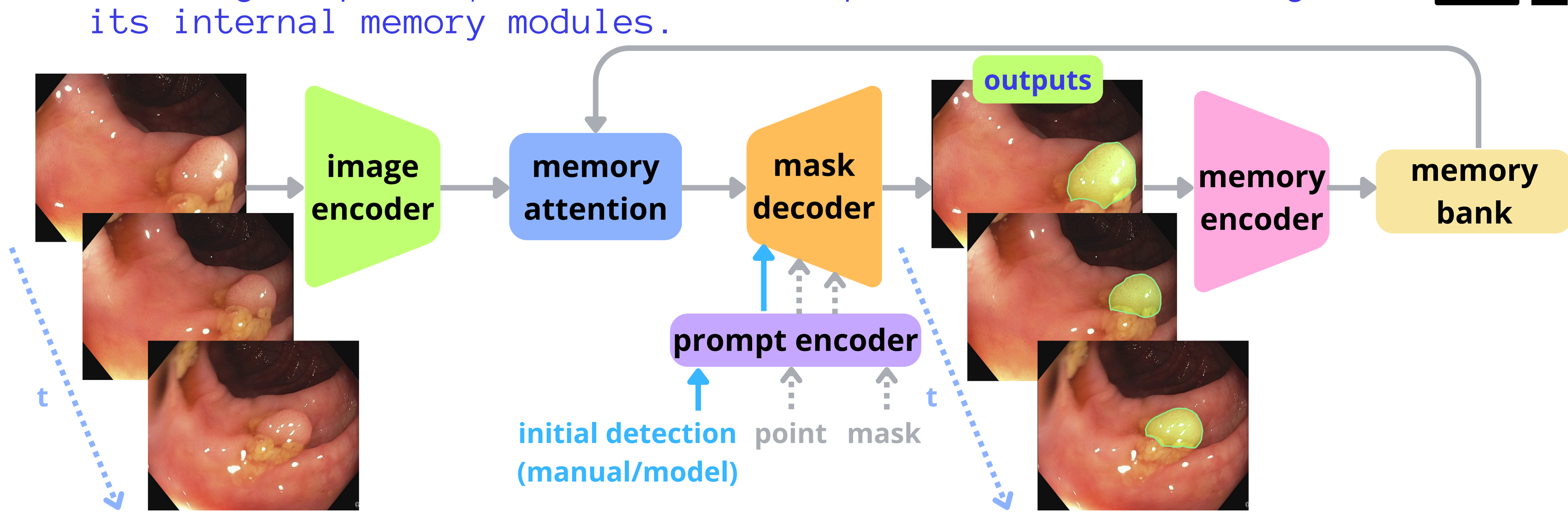
Results

- I am currently creating the testing dataset by manually annotating new video sequences.

See examples

SAM2 Meta

- Unlike traditional pipelines that require an explicit tracking component, SAM2 handles temporal coherence through its internal memory modules.



BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Maia labs *This work is based on data provided by MAIA Labs.*

Excel@FIT 2025