Surgical Instrument Localization using Language and Vision Foundation Models

Ruoxi Zhao[1,2], Neil Getty[2]
[1]University of California, Merced, Merced, CA
[2]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL

MOTIVATION

The ability of detecting and tracking surgical instruments from endoscopic videos can lead to many transformational interventions, not limited to assessing surgical performance or identifying tools used and choreography, yet it is tedious and time consuming to label the tools manually frame by frame from a wide variety of surgeries.
In this project, we are using the segment anything model (SAM) developed by Meta AI[1], which is a foundation model in image segmentation that was trained on one billion masks and 11 million images, to segment the desired surgical instrument based on the bounding box prompts. Since SAM was not specifically trained on surgery videos, it lacks the domain knowledge in identifying and segmenting the entire surgical tools. Therefore, we experimented with using Contrastive Language-Image Pre-Training (CLIP)[2] and fine-tuning SAM on customized datasets. The goal of this project is to utilize and improve the existing language and vision foundation models as an automatic image segmentation tool that can be beneficial in robotic surgery.

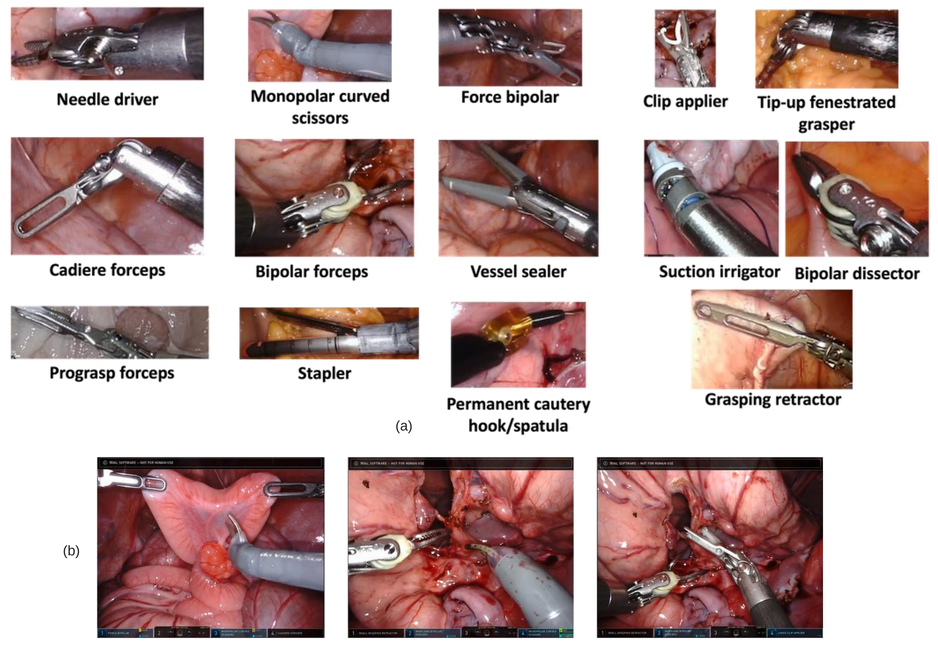


Figure 1. The 14 surgical instruments that we want to localize and identify in the MICCAI 2023 challenge[3](a); three example frames of the endoscopic videos(b).

MAJOR ACCOMPLISHMENTS

Language/Vision Joint Representation Modeling for Semantic Segmentation

CLIP is a neural network model that was trained on numerous pairs of text and images; therefore, it is capable of translating text to images and vice versa using SAM with CLIP[4].

We experimented with using the CLIP-interrogator[5] as the text decoder:

- generate a description for each surgical instruments;
- feed the keywords as text prompts to SAM;
- check if SAM segments out the corresponding tools.

Findings and Problems:

- the words “metal”, “handle”, and “knife” generate a mask for all the existing tools in the image;
- The model failed to capture the smaller parts of the tools;
- Some of the text prompts are similar and contain the same keywords. It is hard to segment out the corresponding tool by itself with its text decoder from CLIP.

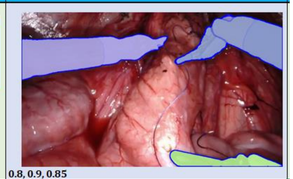
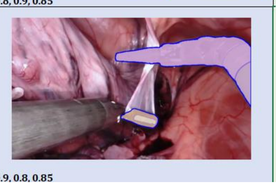
Surgical Tool Name	Segmentation using SAM	Text Decoder from CLIP (Classic Mode, CLIP model: ViT-L-14/openai, Blip-large)	Feeding Text Decoder to SAM with CLIP (With context prompt only)	Testing on surgery frames with context prompt only
Needle Driver		there is a close up of a metal object with a black background, a digital rendering inspired by Katsukawa Shunshō, polycount, cobra, holding dagger, closeup of fist, pen		 0.8, 0.9, 0.85
Cadiere Forceps		there is a metal object with a metal handle on it, a digital rendering inspired by Jozef Israëls, cg society, figuration libre, mechanical paw, neck shackle, large chain		 0.9, 0.8, 0.85

Figure 2. Two examples of utilizing CLIP-interrogator to get the text decoder, using the keywords as text prompt to generate mask from SAM, and comparing results.

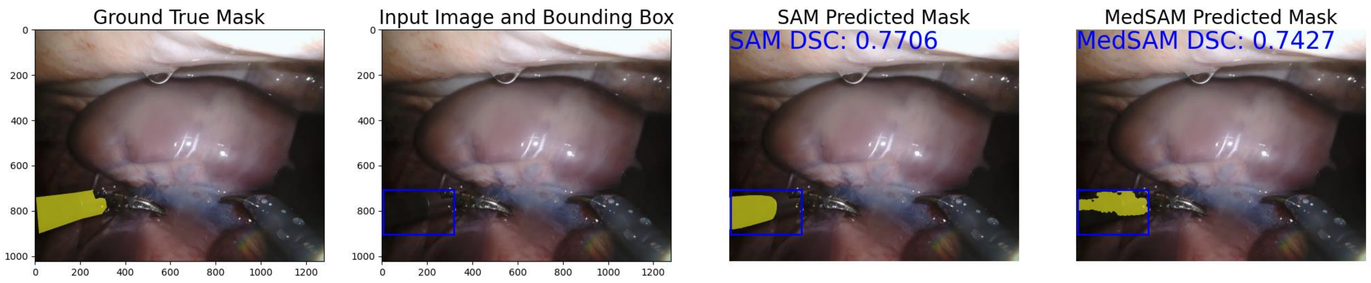


Figure 3. From left to right, an example of the ground true mask, the input image and bounding box, the predicted mask from SAM with a mean dice similarity coefficient(DSC) score, comparing with the predicted mask from the fine-tuned model MedSAM with a DSC score. A higher DSC score means that the predicted mask has more pixel-wise agreement to the ground truth mask, i.e., a higher DSC scores indicates a better predicted mask.

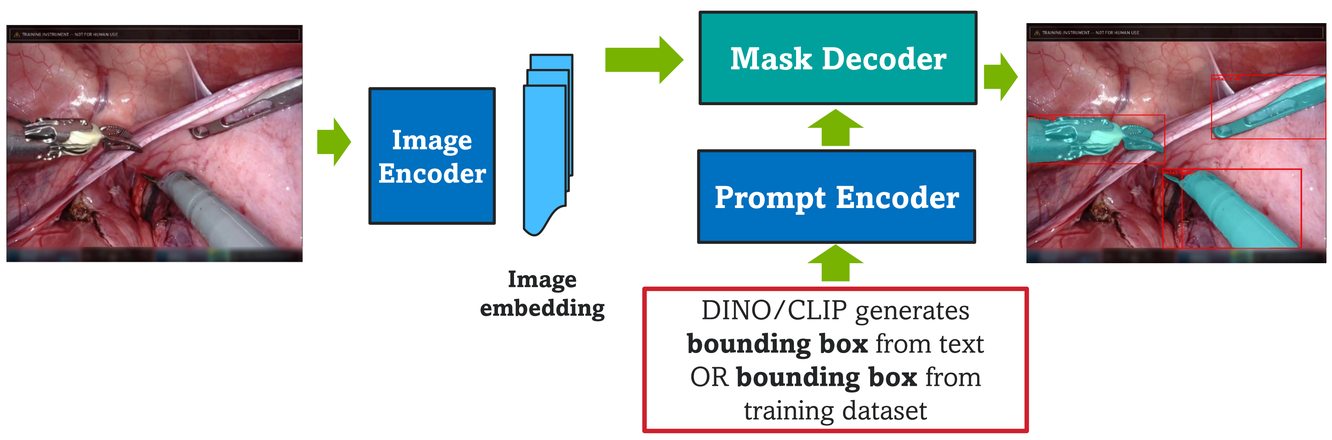


Figure 4. A general structure of the SAM model. The heavyweight image encoder outputs an image embedding, that can be efficiently queried with input prompts(bounding box) and output a mask over the desired areas according to the prompts[1]. In this case, the model DINO+SAM[7] generates bounding boxes according to the text prompt “metal”, then generates masks inside the bounding boxes.

Fine-tune Vision Foundation Model on Unseen Domain

Segment Anything in Medical Images (MedSAM)[6] is a fine-tuned model based on SAM in the medical image domain. In our project, we used the fine-tuning framework from MedSAM to train SAM on our customized dataset, the MICCAI challenges in 2017 and 2018 on segmenting surgical tools. We are using the bounding boxes as prompts, where they correspond to different parts of the surgical instruments.

The Customized Dataset and Training Results:

- 6270 training images with 21128 masks;
- 93,728,252 trainable parameters in the image encoder and mask decoder;
- The images’ original sizes were 1024*1280 and 1080*1920 pixels, resized and padded all images to be 1024*1024 during the preprocessing;
- Trained the SAM model on four Tesla V100-SXM2-32GB GPUs with a batch size of 3, learning rate of 0.0008, and sam_vit_b_01ec64.pth as the SAM checkpoint;
- Each GPU was allocated with 29 GBs CUDA space after loading the entire dataset, and each epoch took around 11 minutes.

So far, our best result was running with 67 epochs with a loss of approximately 0.15. Weak labels of the dataset from the MICCAI 2023 challenge is another obstacle that we need to overcome, as we trained the model with bounding boxes of the ground true mask, but we are only supplied with the name of surgical instruments that are presented in the video.

IMPACT

AI has the potential to improve training, augment performance, understand outcomes, or even automate surgery. Automatically detecting and localizing the different surgical instruments in endoscopic videos is a vital step towards context awareness in robotic surgery. A fully capable approach for localizing and tracking surgical tools in videos may be used to estimate the kinematic movements of the robot, which are the physical manipulations of the expert surgeon. As this technology develops, we may explore predictive and generative models that can generate kinematic data and feed it to the robot for automation.

FUTURE DIRECTIONS

- Develop a fully unsupervised approach for surgical tool localization using foundational models, taking advantage of image embeddings of detected masks and reference images.
- Explore alternative approaches of fine-tuning SAM.
- Further exploration of language models including Grounding DINO[7].
- Applying image retrieving technique on the surgical tool dataset

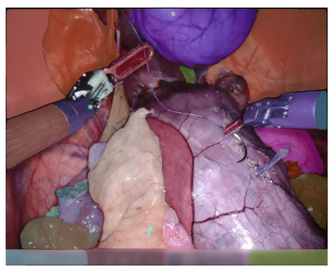


Figure5. An example of using Segment Everything Mode in SAM to generate all the possible masks with no prompt, for the purpose of unsupervised learning.

REFERENCE

[1]Kirillov, A., “Segment Anything”, arXiv e-prints, 2023. doi:10.48550/arXiv.2304.02643.
[2]Radford, A., “Learning Transferable Visual Models From Natural Language Supervision”, arXiv e-prints, 2021. doi:10.48550/arXiv.2103.00020.
[3]Surgical tool localization in endoscopic video, MICCAI 2023 challenge, https://surgtoolloc23.grand-challenge.org/surgtoolloc23/
[4]Segment-Anything-with-CLIP GitHub repo https://github.com/Curt-Park/segment-anything-with-clip.git
[5]CLIP-interrogator GitHub repo https://github.com/pharmapsychotic/clip-interrogator.git
[6]Ma, J. and Wang, B., “Segment Anything in Medical Images”, arXiv e-prints, 2023. doi:10.48550/arXiv.2304.12306.
[7]Grounding DINO GitHub repo https://github.com/IDEA-Research/GroundingDINO.git
[8]Segment-Anything GitHub repo https://github.com/facebookresearch/segment-anything.git
[9]MedSAM GitHub repo https://github.com/bowang-lab/MedSAM.git
[10]OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat
[11]Hu, L. (2023, May 30). How to fine-tune meta sam. Medium. https://pub.towardsai.net/fine-tune-meta-sam-19f7cd4331dd
[12]Glue, R. (2023, June 9). Fine-tune segment-anything model. Medium. https://medium.com/@rustemgal/fine-tune-segment-anything-model-9877993d9db9
[13]Bonnet, A. (2023, April 13). How to fine-tune segment anything. Encord. https://encord.com/blog/learn-how-to-fine-tune-the-segment-anything-model-sam/

ACKNOWLEDGMENTS

I want to express my gratitude towards Dr. Neil Getty for choosing me as his mentee and teaching me throughout this internship. I appreciate this research opportunity provided by Sustainable Horizons Institute, Sustainable Research Pathway, and Argonne National Lab. I also thank MICCAI for organizing the challenge and supplying the datasets. Finally, I want to acknowledge my mentors, friends, and family for the supports.