

Augmenting Efficient Real-time Surgical Instrument Segmentation in Video with Point Tracking and Segment Anything

Zijian Wu¹, Adam Schmidt¹, Peter Kazanzides² and Septimiu E. Salcudean¹

¹Robotics and Control Laboratory, Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada

E-mail: zijianwu@ece.ubc.ca

²Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

The Segment Anything Model (SAM) is a powerful vision foundation model that is revolutionizing the traditional paradigm of segmentation. Despite this, a reliance on prompting each frame and large computational cost limit its usage in robotically assisted surgery. Applications, such as augmented reality guidance, require little user intervention along with efficient inference to be usable clinically. In this study, we address these limitations by adopting lightweight SAM variants to meet the efficiency requirement and employing fine-tuning techniques to enhance their generalization in surgical scenes. Recent advancements in Tracking Any Point (TAP) have shown promising results in both accuracy and efficiency, particularly when points are occluded or leave the field of view. Inspired by this progress, we present a novel framework that combines an online point tracker with a lightweight SAM model that is fine-tuned for surgical instrument segmentation. Sparse points within the region of interest are tracked and used to prompt SAM throughout the video sequence, providing temporal consistency. The quantitative results surpass the state-of-the-art semi-supervised video object segmentation method XMem on the EndoVis 2015 dataset with 84.8 IoU and 91.0 Dice. Our method achieves promising performance that is comparable to XMem and transformer-based fully supervised segmentation methods on *ex vivo* UCL dVRK and *in vivo* CholecSeg8k datasets. In addition, the proposed method shows promising zero-shot generalization ability on the label-free STIR dataset. In terms of efficiency, we tested our method on a single GeForce RTX 4060/4090 GPU respectively, achieving an over 25/90 FPS inference speed. Code is available at: <https://github.com/wuzijian1997/SIS-PT-SAM>

1. Introduction: Surgical instrument segmentation (SIS) is a fundamental task that provides essential visual cues for various downstream applications of robotic surgery, including augmented reality [1] and surgical scene understanding [2, 3]. Segmenting surgical tools from the tissue background is challenging due to occlusion, blood, smoke, motion artifacts, and changing illumination. While deep learning-based segmentation methods have made significant strides in recent years, achieving high accuracy relies on training using large-scale datasets with annotated images. In surgical computer vision, high-quality annotation is particularly scarce due to the time-consuming, labor-intensive, and expertise-demanding process of labeling.

Recently, the Segment Anything Model (SAM) [4], the first promptable foundation model for image segmentation, has attracted widespread attention. SAM's demonstrated impressive zero-shot generalization capability along with its flexible prompting framework make it especially useful for enabling downstream applications. In surgical scenarios, however, the application of SAM faces two challenges. Firstly, the huge computational cost of its heavyweight image encoder architecture, especially when processing high-resolution images, hinders its real-time inference capabilities [5]. Furthermore, numerous studies have reported significant performance degradation of SAM on medical images [20, 21], including images in surgical scenes [19]. In this study, we adopt the lightweight SAM variant to facilitate inference efficiency. Furthermore, we investigate the point prompt-based fine-tuning strategy for MobileSAM [7] to mitigate the performance degradation associated with the lightweight network architecture.

Despite SAM's strong automatic mask generation ability, achieving expected segmentation results in practice often requires appropriate prompts. Providing specific points or descriptive text of the target object can significantly improve the segmentation accuracy. Leveraging the long-term tracking capabilities of the Tracking Any Point (TAP) models, we employ an online point tracker, CoTracker [9], to provide sparse point prompts for SAM. Similar to SAM-PT [8] and DEVA [24], our pipeline decouples video object segmentation (VOS) into image-level segmentation, which can be task-specific, and a universal temporal

propagation. Compared to end-to-end VOS, our "tracking-by-detection" framework can take full advantage of smaller image-level datasets via fine-tuning a task-specific image segmentation model and using it in tandem with a point tracker to maintain temporal consistency.

In summary, our contribution is twofold: (1) we present a real-time video surgical instrument segmentation framework that achieves superior segmentation performance and is suitable for clinical usage due to its good efficiency; (2) we investigate the point prompt-based fine-tuning strategy (will open source) for lightweight SAM using surgical datasets, and the model fine-tuned on only two datasets shows promising generalization on unseen surgical videos.

2. Related work: The goal of Tracking Any Point (TAP) is to estimate the motion of arbitrary physical points throughout a video. TAP-Vid [12] first formalized this task alongside a benchmark dataset and baseline method for TAP. Recent work has showcased the promising online TAP capability and exhibited great robustness to occlusion and exit from the field of view. PIPs++ [13] and TAPIR [14] demonstrate substantial robustness under occlusion and achieve real-time inference speed on high-resolution video. Notably, CoTracker [9] achieves state-of-the-art tracking performance by jointly tracking a set of query points. CoTracker is an online algorithm that processes video sequentially through a sliding window. Optical flow [10, 11] can be used for TAP but tends to accumulate errors over time and faces challenges in handling occlusions, which are common occurrences in surgical scenarios.

SAM is the first vision foundation model for image segmentation, trained over the SA-1B dataset consisting of 1 billion high-quality annotated images [4]. SAM demonstrates impressive zero-shot inference capability on natural images and supports flexible prompts. Nevertheless, SAM's performance often declines in specific fields [19, 20, 21, 23], which can be attributed to a substantial domain gap. Much research has been dedicated to adapting SAM to medical images [15, 18] including surgical images [16, 17]. SurgicalSAM [16] and AdaptiveSAM [17] are adapted to the surgical domain by providing class and text prompts.

However, neither has real-time inference speed. The computational cost of SAM stems from its heavy image encoder, with some research [5, 6, 7, 35] aiming to accelerate SAM's inference and reduce the demand for computation resources.

3. Method: Our proposed framework consists of two key components: a point tracker and a point-based segmentation model. Both components can be flexibly replaced with state-of-the-art models. As shown in Fig. 1, the pipeline can be described as follows. To begin with, the first frame mask of the video sequence is generated to indicate the region of interest (ROI) where query points are initialized. Subsequently, a set of query points is selected within the ROI based on a sampling strategy. After this pre-processing, we employ a point tracker to track these query points and utilize them as prompts at each frame for the segmentation model.

In Section 3.1, we illustrate the pre-processing that is used to initialize query points. In Section 3.2, we formalize and clarify the proposed TAP + SAM framework. In Section 3.3, we introduce the SAM fine-tuning strategy employed in this study.

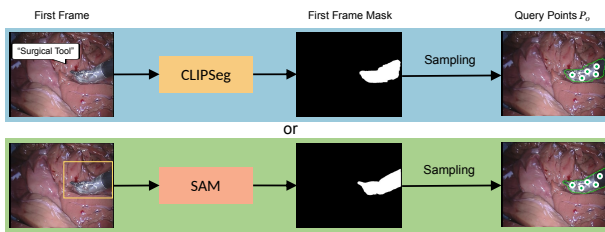


Figure 1 The pre-processing workflow to generate the query points. The segmentation model can be semi-automatic, i.e., SAM + bounding box prompt (bottom row), or fully automatic, i.e., CLIPSeg + fixed text prompt “surgical tool” (top row). Note that, without the initial mask, just manually picking query points is also feasible.

3.1. Pre-processing: We use the SAM model to generate the initial mask by simply inputting a few points or bounding boxes. Inspired by recent advancements in vision-language models, several zero-shot segmentation models based on text prompts have been developed [25, 26]. To achieve the “fully” automatic pipeline, we incorporate a text-promptable segmentation model CLIPSeg [25] to automatically generate the initial mask by setting a text prompt, “surgical tool”. While CLIPSeg can only provide a coarse initial mask, it remains feasible for query point selection as long as its output roughly covers the region of the target instrument.

We initialize query points using the first frame mask. We investigate various query point sampling strategies, including random sampling, uniform sampling on a grid, SIFT keypoints, Shi-Tomasi corner points, and K-Medoids [36] clustering centers. We selected K-Medoids clustering centers because they ensure even partitioning of the entire cluster. The number of medoids assigned to each instance ranges from 1 to 9, and we choose 5 in the experiments. Let

$$P_0 = \{(p_i, t_0)\}, \quad p_i = (x_i, y_i), \quad i = 1, \dots, N \quad (1)$$

be the initial set of query point locations p_i at time t_0 .

3.2. Tracking Any Point + Segment Anything: We describe our framework with reference to Fig. 2. Given a video $V = \{I_t\}$, in which $I_t \in R^2$ is the image at time t , along with a set of initial query points P_0 in I_0 , we use TAP to predict the query points $P_t = \{(p_i, t)\}$, at time t

$$P_t = TAP(V, P_0), \quad (2)$$

and we use P_t and the current image I_t in a segmentation model

$$M_t = Seg(I_t, P_t). \quad (3)$$

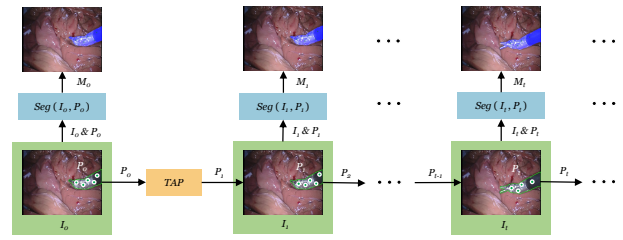


Figure 2 The overview of our video surgical instrument segmentation pipeline that combines a segmentation model $Seg(\cdot)$ and a point tracker $TAP(\cdot)$.

to produce the mask M_t .

We incorporate the state-of-the-art model, CoTracker [9], as the off-the-shelf online point tracker to propagate the initial query points throughout the video sequence. CoTracker takes a short video clip consisting of several frames as the input. It processes video frames in a serial fashion via a 4-frame sliding window, making it suitable for online applications. We also integrated PIPs++ [13] and TAPIR [14] into our software, but only CoTracker is used for experiments in this paper. For the segmentation model Seg , we adopt the fine-tuned MobileSAM to enable real-time processing throughout the entire pipeline, while achieving accurate segmentation.

3.3. Fine-tuning the Segment Anything Model: As for the segmentation model, we first tested two state-of-the-art lightweight SAM variants, MobileSAM and Light HQ-SAM [6]. However, we observed that both methods perform poorly in situations where specularly, blood, or weak lighting is present, as shown in Fig. 3. Driven by this limitation, enhancing the generalization of the lightweight SAM for surgical instrument segmentation becomes crucial. The state-of-the-art MedSAM [15] has demonstrated that fully fine-tuning SAM for medical images can yield promising results. Fully fine-tuning refers to freezing the prompt encoder and updating both the image encoder and mask decoder. Compared with strategies that only update the mask decoder or introduce an adapter layer, fully fine-tuning achieves superior performance. In

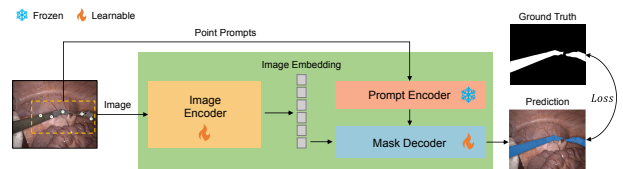


Figure 3 The pipeline of SAM fine-tuning using points. The input for the SAM model consists of images paired with points within the target area. The green rectangle represents the MobileSAM architecture.

this study, we investigate the fully fine-tuning strategy for the most widely used lightweight SAM variant, MobileSAM, to optimize its capability for surgical instrument segmentation. As depicted in Fig. 3, the network architecture of MobileSAM is consistent with the original SAM. Unlike using bounding box prompts in MedSAM, we utilize point prompts during training to maintain consistency with the prompt type used during inference. Both the image encoder and mask decoder are learnable and updated during the fine-tuning. Due to the small number of learnable parameters of MobileSAM (10.13M), the cost of computation is significantly reduced. We train our model on 4 V100 GPUs for 50 epochs. The training is based on the pre-trained MobileSAM weight. In datasets with instance-level labels, we randomly sample 5 points as prompts within the area of each instance. For datasets with binary labels, we randomly sample 5 points as prompts within the segmented region. The loss L utilizes an unweighted combination of binary

cross entropy loss and Dice loss [28], represented as

$$L = L_{BCE} + L_{Dice} \quad (4)$$

We used the AdamW [33] optimizer for training, with a batch size of 32. The initial learning rate is set to $1e-5$ and follows a cosine decay schedule. All images are resized to 1024 by 1024 and undergo random up-down and left-right flip data augmentation. All images are Min-Max normalized and standardized.

4. Results:

4.1. Datasets: We conduct quantitative comparisons on the EndoVis 2015 [31], UCL dVRK [29] and CholecSeg8k [38] datasets to validate the feasibility and performance of our proposed framework. We finetune the SAM model using the training set of the ROBUST-MIS 2019 [39] dataset and provide qualitative results because there is no video-level annotation for quantitative evaluation. To test the zero-shot generalization, we test our method on the unlabeled STIR [30] dataset and display qualitative results.

The EndoVis 2015 dataset provides 25 FPS videos and corresponding articulated da Vinci robotic instruments in *ex vivo* background. It consists of four 45-second training videos and six testing videos (four 15-second and two 60-second videos). Note that there is only one type of instrument (needle driver) in the training set while two types (needle driver and scissor) in the testing set. We keep the original data split for fine-tuning and testing.

The *ex vivo* UCL dVRK dataset consists of 14 videos of 300 frames with corresponding binary segmentation masks. The dataset is split into training (Video 1-8), validation (Video 9 and 10) and testing (Video 11-14) sets. The videos are collected and annotated at 6.7 FPS.

Based on the Cholec80 dataset [40], the CholecSeg8k dataset consists of 8080 frames (from 17 *in vivo* cholecystectomy video clips) with segmentation ground truth. We choose 8 consecutive clips, which are 905 frames in total, as the testing set. The rest of the data is split into the training and validation set as 80% and 20% ratio.

We select two extra datasets for qualitative evaluation. The ROBUST-MIS 2019 data comprises 10,040 annotated images, of which 5,983 are in the training set. The testing set is divided into three stages with increasing levels of difficulty. The dataset provides one labeled frame per clip. The STIR dataset consists of high-resolution label-free videos collected by da Vinci Xi.

4.2. Quantitative Results: We perform comparisons between the state-of-the-art semi-supervised VOS method XMem [32], and fully supervised image segmentation methods TransUNet [22] and SwinUNet [37]. The ablation study of different SAM variants (MobileSAM, HQ-SAM Light, and the default SAM) demonstrates the significant performance improvement of finetuning. All these SAMs use the CoTracker as the online point tracker. We adopt the widely-used segmentation metrics, *IoU* and *Dice*, for quantitative comparison and ablation study. Table 1 displays the quantitative results on three datasets, in which the first 3 rows are comparison results, the next 3 rows are the ablation results, and the bottom row is ours.

Table 1 Quantitative Results of Different Datasets

Methods	EndoVis 2015 [31]		UCL dVRK [13]		CholecSeg8k [38]	
	IoU	Dice	IoU	Dice	IoU	Dice
TransUNet [22]	57.7	71.8	79.1	88.0	71.7	82.8
SwinUNet [37]	59.5	73.0	81.1	89.3	81.5	89.5
XMem [32]	82.6	89.3	91.9	95.4	81.6	87.9
PT+MobileSAM [7]	70.1	80.6	45.0	57.1	49.7	57.4
PT+HQ-SAM Light [6]	69.0	80.2	55.9	68.1	61.3	70.4
PT+SAM (ViT-H) [4]	79.6	88.3	74.3	83.3	64.5	70.6
Ours	84.4	91.0	89.4	93.8	81.9	88.6

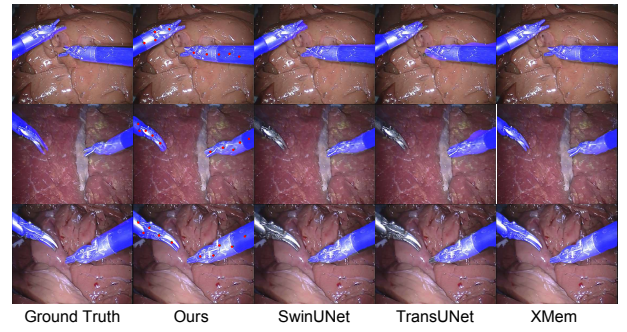


Figure 4 Visualization of segmentation results from several methods on the EndoVis 2015 dataset, in which the images are acquired from the testing Video 1, 5, and 6, respectively (from top row to bottom row). Note that red dots in Fig. 4 - 8 are the point prompts tracked by CoTracker.

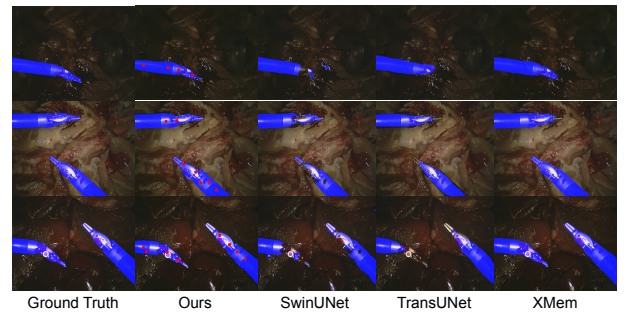


Figure 5 Visualization of segmentation results from several methods on the UCL dVRK dataset, in which the images are acquired from the testing Video 1, 3, and 4, respectively (from top row to bottom row).

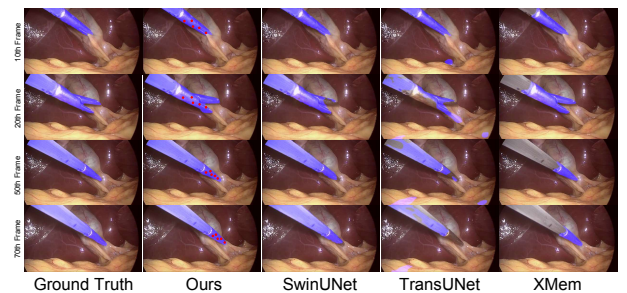


Figure 6 Segmentation results of different methods on the CholecSeg8k dataset. The y-axis is along the frame order (10th, 20th, 50th, and 70th) in one clip.

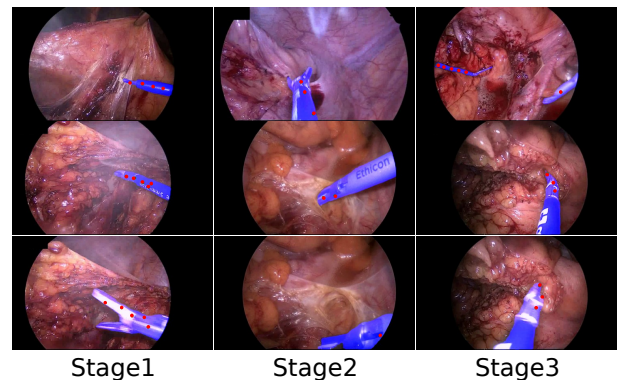


Figure 7 Qualitative results from our proposed methods on the ROBUST-MIS 2019 dataset, in which columns from left to right are the samples from stages 1, 2, and 3, respectively.

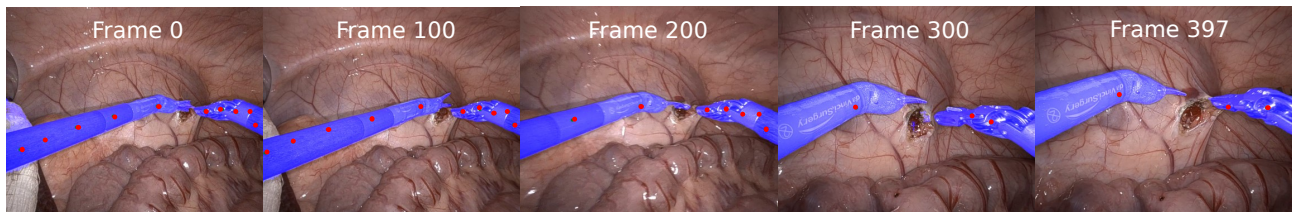


Figure 8. Visualization of qualitative results from our proposed methods on STIR dataset.

4.3. **Qualitative Results:** Fig. 4 and Fig. 5 display a few predicted masks from various methods on EndoVis 2015 and UCL dVRK datasets. Fig. 6 is the segmentation results of the frame 10, 20, 50, and 70, which are sampled on a video clip from the testing set of the CholecSeg8k dataset. Fig. 7 is the visualization of segmentation results on different difficulty testing stages of the ROBUST-MIS 2019 dataset. To evaluate the generalization of our method, we test our framework, which is fine-tuned using the EndoVis 15 and UCL dVRK dataset, on some videos from the STIR dataset. Fig. 8 shows some segmentation results of our method, in which frame 397 is the end frame. The segmentation performance of our method is robust in this video.

4.4. **Efficiency:** Table 2 displays the metrics of efficiency of different methods. The inference latency, inference memory, and learnable parameters represent the time efficiency, computation efficiency for inference, and computation efficiency for training. We run the inference procedure of each method on a laptop with RTX 4060 GPU and a desktop with RTX 4090 GPU. The inference latency of our method achieves 38 ms (26 FPS) and 11 ms (90 FPS), which are real-time and over real-time inference speed, on these two machines. The CoTracker's inference speed fluctuates in the 50-60 FPS range. In terms of computational efficiency, our proposed method only requires 2.8G inference memory. Our method is also training efficient with a small 10.1M parameters. We only list learnable parameters of partial methods, because the other methods do not involve training in this study.

Table 2 Efficiency Metrics with RTX 4090 and 4060 GPUs

Methods	Inference Latency (ms)		Inference Memory (G)	Learnable Param. (M)
	4090	4060		
XMem [32]	4	5	2.0	-
TransUNet [22]	7	8	2.8	105.3
SwinUNet [37]	4	7	1.9	27.2
PT+MobileSAM [7]	11	38	2.8	-
PT+Light HQ-SAM [6]	12	40	2.8	-
PT+ViT-H SAM [4]	220	1300	8.2	-
Ours	11	38	2.8	10.1

5. Discussion: Overall, our method outperforms the state-of-the-art semi-supervised VOS model, XMem, on the EndoVis 2015 dataset. Notably, in the one-minute videos, Video 5 and 6, our proposed method exhibits obvious improvement compared to XMem. This shows that our framework can robustly leverage the temporal information provided by universal TAP methods. The TransUNet and SwinUNet cannot recognize the scissors in testing videos because of the overfitting to the instrument type in the training videos.

As for the UCL dVRK dataset, our proposed method achieves promising performance with slightly weaker quantitative results than XMem. All four UCL dVRK testing videos are captured under weak illumination conditions, which is unrealistic in real surgery. The dark scenes make it challenging for SAM to distinguish the boundary from the tissue background, especially when there are no

such weak lighting conditions in the training set. In contrast, the mask propagation-based XMem takes the full first frame ground truth for. The low video frame rate (6.7 FPS) of the UCL dVRK dataset also poses an obstacle for the TAP, thereby making it hard to provide effective point prompts throughout the video sequence.

The quantitative results on the *in vivo* CholecSeg8k dataset achieve the best *IoU* and second best *Dice*. The segmentation performance of our method is comparable to the state-of-the-art fully supervised image segmentation model SwinUNet.

The ablation study demonstrates the significance of fine-tuning by the substantial improvement compared to non-fine-tuned SAMs, even that of the powerful SAM with ViT-H backbone. Note that the inference speed of the ViT-H SAM is far away from real-time.

Compared to the propagation-based model such as XMem, our method has other advantages. XMem requires an accurate first frame ground truth for inference, while our pipeline only needs a known text prompt "surgical tool". Furthermore, XMem cannot recognize new objects during the video, while our method can tackle this by enabling users to easily pick a few new query points, instead of providing a high-quality mask.

6. Limitations and Future Work: In general, our method achieves satisfied segmentation performance along with good efficiency in extensive SIS video datasets. **However, when processing more challenging datasets like SAR-RARP [41], the segmentation performance of our method is not satisfactory, as depicted in Fig. 9. The videos of the SAR-RARP dataset are recorded during real robot-assisted laparoscopic prostatectomy with a cluttered scene, significant blood, specular reflection, wide-range rapid instrument motion, and frequent camera focus changes.** These natures cause huge challenges for both point tracking and segment anything model.

To enhance the performance in more challenging surgical scenarios, we plan to design a novel SAM prompt method to implicitly leverage the spatio-temporal information from point tracking instead of the naive combination. Kinematics data is a strong prior for instrument identification. However, this is only available for robotic surgical instruments. Other modality information such as text description of surgical instruments can be leveraged as a prior knowledge.

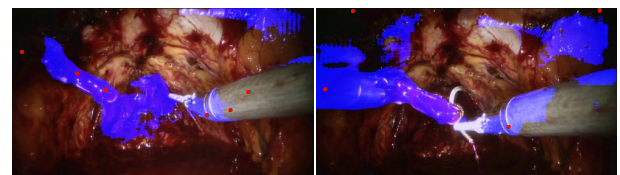


Figure 9. Failure cases on the SAR-RARP dataset.

7. Conclusions: In this study, we present a novel framework using a universal TAP and a fine-tuned lightweight SAM for real-time surgical instrument segmentation in video. Its commendable efficiency and accuracy make it suitable for applications in clinical settings. We investigate the availability of fine-tuning MobileSAM

using point prompts and demonstrate the importance of fine-tuning for SIS. Extensive experiments validate the advancement of our proposed pipeline. Furthermore, our SAM + TAP pipeline demonstrates the potential to serve as a strong VOS baseline by integrating other image segmentation models.

8 References

- [1] Kalia, M., Mathur, P., Tsang, K., Black, P., Navab, N. and Salcudean, S.: Evaluation of a marker-less, intra-operative, augmented reality guidance system for robot-assisted laparoscopic radical prostatectomy. *International Journal of Computer Assisted Radiology and Surgery* **15**(7), 1225–1233 (2020)
- [2] Yip, M., Salcudean, S., Goldberg, K., Althoefer, K., Menciassi, A., Opfermann, J.D., Krieger, A., Swaminathan, K., Walsh, C.J., Huang, H. and Lee, I.C.: Artificial intelligence meets medical robotics. *Science* **381**(6654), 141–146 (2023)
- [3] Ding, X. and Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. *IEEE Transactions on Medical Imaging* **41**(11), 3309–3319 (2022)
- [4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P.: Segment anything. *arXiv preprint arXiv:2304.02643*. (2023)
- [5] Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M. and Wang, J.: Fast Segment Anything. *arXiv preprint arXiv:2306.12156*. (2023)
- [6] Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K. and Yu, F.: Segment anything in high quality. In: *Advances in Neural Information Processing Systems* (2023)
- [7] Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S. and Hong, C.S.: Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289*. (2023)
- [8] Rajič, F., Ke, L., Tai, Y.W., Tang, C.K., Danelljan, M. and Yu, F.: Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*. (2023)
- [9] Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A. and Rupprecht, C.: Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*. (2023)
- [10] Teed, Z. and Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: *Computer Vision–ECCV 2020: 16th European Conference* (2020)
- [11] Teed, Z. and Deng, J.: RAFT-3D: Scene Flow using Rigid-Motion Embeddings. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021)
- [12] Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A. and Yang, Y.: TAP-Vid: A Benchmark for Tracking Any Point in a Video. In: *Advances in Neural Information Processing Systems* (2022)
- [13] Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G. and Guibas, L.J.: PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023)
- [14] Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J. and Zisserman, A.: TAPIR: Tracking Any Point with per-frame Initialization and temporal Refinement. *arXiv preprint arXiv:2306.08637*. (2023)
- [15] Ma, J., He, Y., Li, F., Han, L., You, C. and Wang, B.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024)
- [16] Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J. and Wang, Z.: SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2023)
- [17] Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S. and Patel, V.M.: AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation. *arXiv preprint arXiv:2308.03726*. (2023)
- [18] Zhang K. and Liu D.: Customized Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv:2304.13785*. (2023)
- [19] Wang, A., Islam, M., Xu, M., Zhang, Y. and Ren, H.: SAM Meets Robotic Surgery: An Empirical Study in Robustness Perspective. *arXiv preprint arXiv:2304.14674*. (2023)
- [20] Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N. and Zhang, Y.: Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis* **89** (2023)
- [21] Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C. and Liu, S.: Segment anything model for medical images? *Medical Image Analysis* **92** (2024)
- [22] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*. (2021)
- [23] Wang, L., Ye, X., Zhu, L., Wu, W., Zhang, J., Xing, H. and Hu, C.: When SAM Meets Sonar Images. *arXiv preprint arXiv:2306.14109*. (2023)
- [24] Cheng, H.K., Oh, S.W., Price, B., Schwing, A. and Lee, J.Y.: Tracking Anything with Decoupled Video Segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023)
- [25] Lüddecke, T. and Ecker, A.: Image Segmentation Using Text and Image Prompts. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
- [26] Zhou, Z., Alabi, O., Wei, M., Vercauteren, T. and Shi, M.: Text Promptable Surgical Instrument Segmentation with Vision-Language Models. In: *Advances in Neural Information Processing Systems* (2023)
- [27] Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y.: Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv:2304.12620*. (2023)
- [28] Milletari, F., Navab, N. and Ahmadi, S.A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *Fourth International Conference on 3D Vision (3DV)* (2016)
- [29] Colleoni, E., Edwards, P. and Stoyanov, D.: Synthetic and Real Inputs for Tool Segmentation in Robotic Surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020)
- [30] Schmidt, A., Mohareri, O., DiMaio, S. and Salcudean, S.E.: STIR: Surgical Tattoos in Infrared. *arXiv preprint arXiv:2309.16782*. (2023)
- [31] Bodenstedt, S., Allan, M., Agustinis, A., Du, X., Garcia-Peraza-Herrera, L., Kennigott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D. and Sznitman, R.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*. (2018)
- [32] Cheng, H.K. and Schwing, A.G.: XMem: Long-Term Video

Object Segmentation with an Atkinson-Shiffrin Memory Model. In: European Conference on Computer Vision (2022)

- [33] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. (2017)
- [34] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M. and Sorkine-Hornung, A.: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
- [35] Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F. and Krishnamoorthi, R.: EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything. arXiv preprint arXiv:2312.00863. (2023)
- [36] Mannor, S., Jin, X., Han, J. and Zhang, X.: K-medoids clustering. Encyclopedia of machine learning. (2011)
- [37] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In European Conference on Computer Vision. (2022)
- [38] Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L. and Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453. (2020)
- [39] Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N. and Bruno, P.: Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. Medical image analysis, 70, p.101920. (2021)
- [40] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M. and Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging, 36(1), pp.86-97. (2016)
- [41] Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y. and Boels, M.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496. (2023)