SAM Meets Robotic Surgery: An Empirical Study on Generalization, Robustness and Adaptation

An Wang[1] ⋆, Mobarakol Islam[2] ⋆, Mengya Xu[3], Yang Zhang[4], and Hongliang Ren[1,3] ⋆⋆

[1] Dept. of Electronic Engineering, Shun Hing Institute of Advanced Engineering (SHIAE), The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Dept. of Medical Physics and Biomedical Engineering, Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK
[3] Dept. of Biomedical Engineering, National University of Singapore, Singapore
[4] School of Mechanical Engineering, Hubei University of Technology, Wuhan, China
wa09@link.cuhk.edu.hk, mobarakol.islam@ucl.ac.uk, mengya@u.nus.edu, yzhangcst@hbut.edu.cn, hlren@ee.cuhk.edu.hk

Abstract. The Segment Anything Model (SAM) serves as a fundamental model for semantic segmentation and demonstrates remarkable generalization capabilities across a wide range of downstream scenarios. In this empirical study, we examine SAM's robustness and zero-shot generalizability in the field of robotic surgery. We comprehensively explore different scenarios, including prompted and unprompted situations, bounding box and points-based prompt approaches, as well as the ability to generalize under corruptions and perturbations at five severity levels. Additionally, we compare the performance of SAM with state-of-the-art supervised models. We conduct all the experiments with two well-known robotic instrument segmentation datasets from MICCAI EndoVis 2017 and 2018 challenges. Our extensive evaluation results reveal that although SAM shows remarkable zero-shot generalization ability with bounding box prompts, it struggles to segment the whole instrument with point-based prompts and unprompted settings. Furthermore, our qualitative figures demonstrate that the model either failed to predict certain parts of the instrument mask (e.g., jaws, wrist) or predicted parts of the instrument as wrong classes in the scenario of overlapping instruments within the same bounding box or with the point-based prompt. In fact, SAM struggles to identify instruments in complex surgical scenarios characterized by the presence of blood, reflection, blur, and shade. Additionally, SAM is insufficiently robust to maintain high performance when subjected to various forms of data corruption. We also attempt to fine-tune SAM using Low-rank Adaptation (LoRA) and propose Surgical-SAM, which shows the capability in class-wise mask prediction without prompt. Therefore, we can argue that, without further domain-specific fine-tuning, SAM is not ready for downstream surgical tasks.

⋆ An Wang and Mobarakol Islam are co-first authors.
⋆⋆ Corresponding author.

1 Introduction

Segmenting surgical instruments and tissue poses a significant challenge in robotic surgery, as it plays a vital role in instrument tracking and position estimation within surgical scenes. Nonetheless, current deep learning models often have limited generalization capacity as they are tailored to specific surgical sites. Consequently, it is crucial to develop generalist models that can effectively adapt to various surgical scenes and segmentation objectives to advance the field of robotic surgery [18]. Recently, segmentation foundation models have made great progress in the field of natural image segmentation. The segment anything model (SAM) [14], which has been trained on more than one billion masks, exhibits remarkable proficiency in generating precise object masks using various prompts such as bounding boxes and points. SAM stands as the pioneering and most renowned foundation model for segmentation. Whereas, several works have revealed that SAM can fail on common medical image segmentation tasks [4,8,6,16]. This is not surprising or unexpected since SAM's training dataset primarily comprises natural image datasets. Consequently, it raises the question of enhancing SAM's strong feature extraction capability for medical image tasks. Med SAM Adapter [22] utilizes medical-specific domain knowledge to improve the segmentation model through a simple yet effective adaptation technique. SAMed [23] has applied a low-rank-based finetuning strategy to the SAM image encoder, as well as prompt encoder and mask decoder on the medical image segmentation dataset.

However, evaluating the performance of SAM in the context of surgical scenes remains an insufficiently explored area that has the potential for further investigation. This study uses two publicly available robotic surgery datasets to assess SAM's generalizability under different settings, such as bounding box and point-prompted. Moreover, we have examined the possibility of fine-tuning SAM through Low-rank Adaptation (LoRA) to examine its capability to predict masks for different classes without prompts. Additionally, we have analyzed SAM's robustness by assessing its performance on synthetic surgery datasets, which contain various levels of corruption and perturbations.

2 Experimental Settings

Datasets. We have employed two classical datasets in endoscopic surgical instrument segmentation, i.e., EndoVis17 [2] and EndoVis18 [1]. For the EndoVis17 dataset, unlike previous works [20,5,13] which conduct 4-fold cross-validation for training and testing on the 8×225-frame released training data, we report SAM's performance directly on all eight sequences (1-8). For the EndoVis18 dataset, we follow the dataset split in ISINet [5], where sequences 2, 5, 9, and 15 are utilized for evaluation.

Prompts. The original EndoVis datasets [2,1] do not have bounding boxes or point annotations. We have labeled the datasets with bounding boxes for each instrument, associated with corresponding class information. Additionally, regarding the single-point prompt, we obtain the center of each instrument mask

Table 1. Quantitative comparison of binary and instrument segmentation on EndoVis17 and EndoVis18 datasets. The best and runner-up results are shown in bold and underlined.

Type	Method	Pub/Year(20-)	Arch.	EndoVis17		EndoVis18	
				Binary IoU	Instrument IoU	Binary IoU	Instrument IoU
Single-Task	Vanilla UNet	MICCAI15	UNet	75.44	15.80	<u>68.89</u>	-
	TernausNet	ICMLA18	UNet	83.60	35.27	-	46.22
	MF-TAPNet	MICCAI19	UNet	87.56	37.35	-	67.87
	Islam et al.	RA-L19	-	84.50	-	-	-
	ISINet	MICCAI21	Res50	-	55.62	-	73.03
	Wang et al.	MICCAI22	UNet	-	-	58.12	-
Multi-Task	ST-MTL	MedIA21	-	83.49	-	-	-
	AP-MTL	ICRA20	-	<u>88.75</u>	-	-	-
	S-MTL	RA-L22	-	-	-	-	43.54
	TraSeTR	ICRA22	Res50 + Trfm	-	60.40	-	<u>76.20</u>
	S3Net	WACV23	Res50	-	<u>72.54</u>	-	75.81
Prompt-based	SAM 1 Point	arxiv23	ViT_h	53.88	55.96*	57.12	54.30*
	SAM Box	arxiv23	ViT_h	**89.19**	**88.20***	**89.35**	**81.09***

* Categorical information directly inherits from associated prompts.

by simply computing the moments of the mask contour. Since SAM [14] only predicts binary segmentation masks, for instrument-wise segmentation, the output instrument labels are assigned inherited from the input prompts.

Metrics. The IoU and Dice metrics from the EndoVis17 [2] challenge[5] is used. Specifically, only the classes presented in a frame are considered in the calculation for instrument segmentation.

Comparison methods. We have involved several classical and recent methods, including the vanilla UNet [17], TernausNet [20], MF-TAPNet [13], Islam et al. [10], Wang et al. [21], ST-MTL [11], S-MTL [19], AP-MTL [12], ISINet [5], TraSeTR [24], and S3Net [3] for surgical binary and instrument-wise segmentation. The ViT-H-based SAM [14] is employed in all our investigations except for the finetuning experiments. Note that we cannot provide an absolutely fair comparison because existing methods do not need prompts during inference.

3 Surgical Instruments Segmentation with Prompts

Implementation With bounding boxes and single points as prompts, we input the images to SAM [14] to get the predicted binary masks for the target objects. Because SAM [14] can not provide consistent categorical information. We compromise to use the class information from the bounding boxes directly. In this way, we derive instrument-wise segmentation while bypassing the possible errors from misclassifications, an essential factor affecting instrument-wise segmentation accuracy.

Results and Analysis As shown in Table 1, with bounding boxes as prompts, SAM [14] outperforms previous unprompted supervised methods in binary and

[5] https://github.com/ternaus/robot-surgery-segmentation

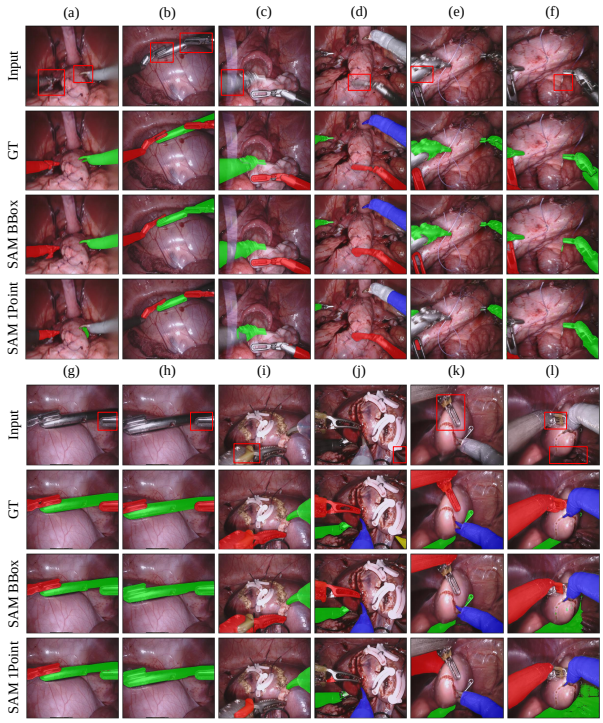

Fig. 1. Qualitative results of SAM on various challenging frames. Red rectangles highlight the typical challenging regions which cause unsatisfactory predictions.

instrument-wise segmentation on both datasets. However, with single points as prompts, SAM [14] degrades a lot in performance, indicating its limited ability to segment surgical instruments from weak prompts. This reveals the performance of the SAM closely relies on prompt quality. For complicated surgical scenes, SAM [14] still struggles to produce accurate segmentation results, as shown in columns (a) to (l) of Fig. 1. Typical challenges, including shadows (a), motion blur (d), occlusion (b, g, h), light reflection (c), insufficient light (j, l), over brightness (e), ambiguous suturing thread (f), instrument wrist (i), and irregular instrument pose (k), all lead to unsatisfied segmentation performance.

4 Robustness under Data Corruption

Implementation Referring to the robustness evaluation benchmark [7], we have evaluated SAM [14] under 18 types of data corruptions at 5 severity levels following the official implementations[6] with box prompts. Note that the *Elastic Transformation* has been omitted to avoid inconsistency between the input image

[6] https://github.com/hendrycks/robustness

Table 2. Quantitative results on various corrupted EndoVis18 validation data.

Task	Severity	Noise				Blur					Weather					Digital			
		Gaussian	Shot	Impulse	Speckle	Defocus	Glass	Motion	Zoom	Gaussian	Snow	Frost	Fog	Bright	Spatter	Contrast	Pixel	JPEG	Saturate
Binary	0	89.35																	
	1	77.69	80.18	80.43	83.28	82.01	80.53	82.99	80.30	85.40	84.08	83.12	85.38	87.43	86.69	85.76	81.12	58.77	86.64
	2	73.92	76.07	76.15	81.65	80.21	79.20	80.22	77.55	81.69	80.69	80.34	84.65	87.27	84.21	84.90	79.32	56.04	84.85
	3	69.21	71.74	73.02	77.74	76.96	72.64	75.50	75.27	78.31	79.58	78.90	83.62	87.23	82.50	83.36	73.81	56.25	86.84
	4	63.80	65.41	67.29	75.28	73.79	72.38	69.60	73.22	75.23	76.33	78.38	82.28	87.06	83.12	77.12	70.82	57.59	83.21
	5	57.07	60.61	61.61	71.83	69.85	69.59	66.25	71.58	66.96	77.66	76.82	78.84	86.43	79.62	66.58	68.55	56.77	81.26
Instrument	0	81.09																	
	1	69.51	71.83	72.25	74.82	73.64	72.13	74.33	71.41	76.79	75.40	74.42	76.82	79.16	78.24	77.17	72.94	54.86	78.27
	2	66.06	68.09	68.53	73.19	71.74	71.02	71.46	68.85	73.15	72.13	71.65	76.14	79.00	75.54	76.22	71.55	52.23	76.61
	3	62.01	64.44	65.89	69.75	68.74	64.97	67.13	67.12	70.08	70.97	70.21	75.01	78.90	73.70	74.67	66.83	51.63	78.39
	4	57.28	59.12	61.03	67.82	65.87	64.87	62.15	65.18	67.23	68.43	69.79	73.73	78.73	74.24	69.48	63.99	51.88	74.91
	5	51.56	55.16	55.86	64.76	62.43	62.23	59.26	63.96	60.60	69.33	68.32	70.45	78.19	70.72	61.14	61.79	51.01	73.35

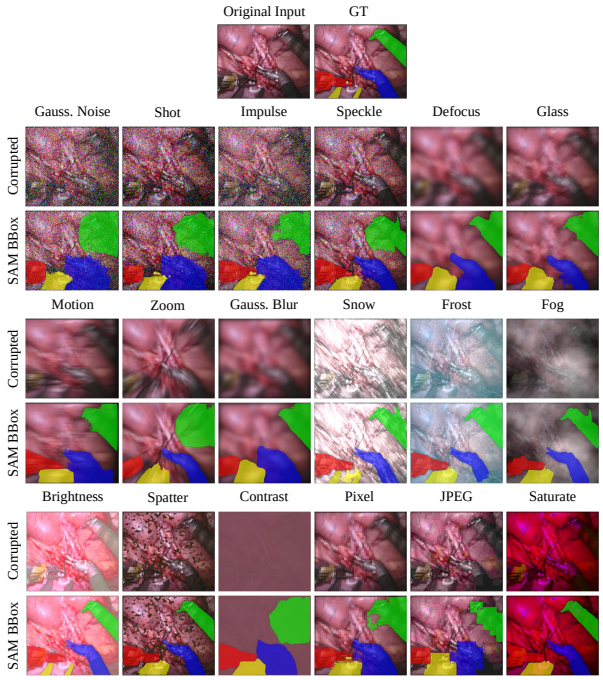

Fig. 2. Qualitative results of SAM under 18 data corruptions of level-5 severity.

and associated masks. The adopted data corruption can be allocated into four distinct categories of *Noise*, *Blue*, *Weather*, and *Digital*.

Results and Analysis The severity of data corruption is directly proportional to the degree of performance degradation in SAM [14], as depicted in Table 2. The robustness of SAM [14] may be influenced differently depending on the nature of the corruption present. However, in most scenarios, SAM's performance diminishes significantly. Notably, *JPEG Compression* and *Gaussian Noise* have the greatest impact on segmentation performance, whereas *Brightness* has a neg-

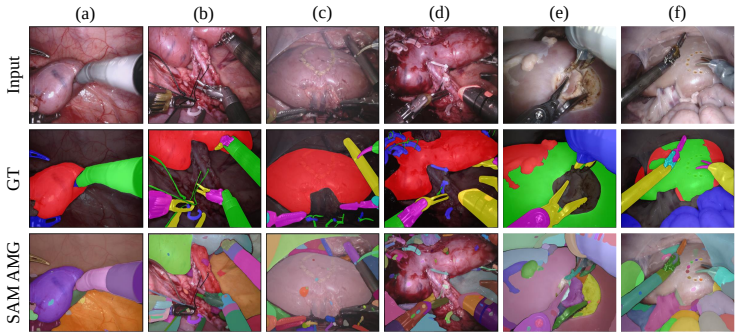

Fig. 3. Unprompted automatic mask generation for surgical scene segmentation.

ligible effect. Figure 2 presents one exemplar frame in its original state alongside various corrupted versions at a severity level of 5. We can observe that SAM [14] suffers significant performance degradation in most cases.

5 Automatic Surgical Scene Segmentation

Implementation Without prompts, SAM [14] can also facilitate automatic mask generation (AMG) for the entire image. For naive investigation of the automatic surgical scene segmentation results, we use the default parameters from the official implementation[7] without further tuning. The colors of each segmented mask are randomly assigned because SAM [14] only generates binary masks for each object.

Results and Analysis As shown in Fig. 3, in surgical scene segmentation of EndoVis18 [1] data, SAM [14] can produce promising results on simple scenes like columns (a) and (f). But it encounters difficulties when applied to more complicated scenes, as it struggles to differentiate between the entirety of instrument articulating parts accurately and to identify discrete tissue structures as interconnected units. As a foundation model, SAM [14] still lacks comprehensive awareness of objects' semantics, especially in downstream domains like surgical scenes.

6 Parameter-efficient Finetuning with Low-rank Adaptation

With the rapid emergence of foundational and large AI models, utilizing the pretrained models effectively and efficiently for downstream tasks has attracted increasing research interest. Although SAM [14] has shown decent segmentation

[7] https://github.com/facebookresearch/segment-anything

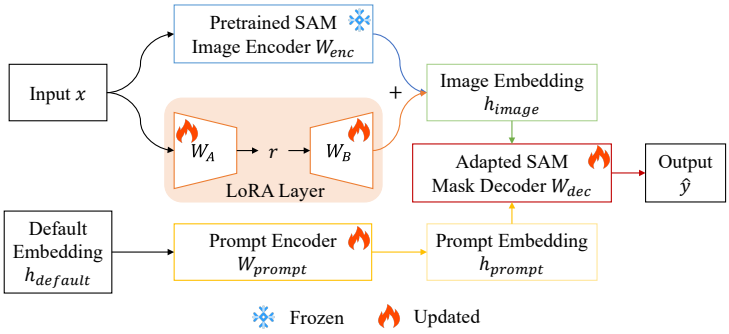

Fig. 4. Overall architecture of our SurgicalSAM.

performance with prompts and can cluster objects in surgical scenes, we seek to finetune and adapt it to make it capable of traditional unprompted multi-class segmentation pipeline - take one image as input only, and predict its segmentation mask with categorical labels.

Implementation To efficiently finetune SAM [14] and enable it to support multi-class segmentation without relying on prompts, we consider utilizing the strategy of Low-rank Adaptation (LoRA) [9] and also adapting the original mask decoder to output categorical labels. Taking inspiration from SAMed [23], we implement a modified architecture as shown in Fig. 4, whereby the pretrained SAM image encoder maintains its frozen weights W_{enc} during finetuning while additional light-weight LoRA layers are incorporated for updating purposes. In this way, we can not only leverage the exceptional feature extraction ability of the original SAM encoder, but also gradually capture the surgical data representations and store the domain-specific knowledge in the LoRA layers parameter-efficiently. We denote this modified architecture as "SurgicalSAM". With an input image x, we can derive the image embedding h_{image} following

$$h_{image} = W_{enc}x + \Delta W x, \tag{1}$$

where ΔW is the weight update matrix of LoRA layers. Then we can decompose ΔW into two smaller matrices: $\Delta W = W_A W_B$, where W_A and W_B are $A \times r$ and $r \times B$ dimensional matrices, respectively. r is a hyper-parameter that specifies the rank of the low-rank adaptation matrices. To maintain a balance between model complexity, adaptability, and the potential for underfitting or overfitting, we empirically set the rank r of W_A and W_B in the LoRA layers to 4.

During the unprompted automatic mask generation (AMG), the original SAM uses fixed default embeddings $h_{default}$ for the prompt encoder with weights W_{prompt}. We adopted this strategy and updated the lightweight prompt encoder during finetuning, as shown in Fig. 4. In addition, we modified the segmentation head of the mask decoder W_{dec} to allow for the production of predictions for

Table 3. Quantitative evaluation of SurgicalSAM under data corruption.

Severity	Noise				Blur					Weather					Digital			
	Gaussian	Shot	Impulse	Speckle	Defocus	Glass	Motion	Zoom	Gaussian	Snow	Frost	Fog	Bright	Spatter	Contrast	Pixel	JPEG	Saturate
0	71.38																	
1	24.31	30.68	28.88	45.53	59.50	60.21	61.29	56.32	64.67	57.84	54.80	54.95	66.67	65.74	57.56	64.81	54.30	60.01
2	12.19	15.43	12.77	36.92	53.85	56.48	55.72	52.81	55.54	29.68	36.33	51.32	63.73	62.59	50.89	64.00	49.56	28.92
3	5.84	6.30	7.34	17.26	45.56	43.71	50.97	49.55	47.24	42.20	26.31	44.17	62.22	60.65	36.90	54.99	46.24	64.85
4	4.26	4.15	4.63	10.19	39.23	39.64	43.27	46.38	39.65	30.21	25.80	38.28	60.90	51.22	16.42	40.64	36.69	60.36
5	3.79	3.79	3.92	6.37	32.49	38.05	38.16	43.99	26.67	13.97	20.60	20.92	59.64	40.51	4.95	34.00	24.03	50.50

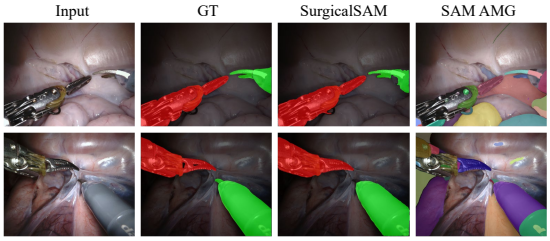

Fig. 5. Qualitative comparison of our SurgicalSAM with the original SAM.

each semantic class. In contrast to the binary ambiguity prediction of the original mask decoder of SAM, the modified decoder predicts each semantic class of $\hat{y}$ in a deterministic manner. In other words, it is capable of semantic segmentation beyond binary segmentation.

We adopt the training split of the Endo18 dataset for finetuning and test with the validation split, as other works reported in Table 1. Following SAMed [23], we adopt the combination of the Cross Entropy loss L_{CE} and Dice loss L_{Dice} which can be expressed as

$$L = \lambda L_{Dice} + (1 - \lambda) L_{CE}, \tag{2}$$

where λ is a weighting coefficient balancing the effects of the two losses. We empirically set λ as 0.8 in our experiments. Due to resource constraints, we utilize the ViT_b version of SAM and finetuning on two RTX3090 GPUs. The maximum epochs are 160, with a batch size 12 and an initial learning rate of 0.001. To stabilize the finetuning process, we apply warmup for the first 250 iterations, followed by exponential learning rate decay. Random flip, rotation, and crop are applied to augment the training images and avoid overfitting. The images are resized to 512×512 as model inputs. Besides, we use AdamW [15] optimizer with a weight decay of 0.1 to update model parameters.

Results and Analysis After naively finetuning, the SurgicalSAM model can manage the instrument-wise segmentation without reliance on prompts. With further tuning of hyper-parameters like the learning rate, the batch size, and the optimizer, SurgicalSAM can achieve **71.38%** mIoU score on the validation split of the Endo18 dataset, which is on par with the state-of-the-art models in Table 1. Since other methods in Table 1 are utilizing temporal and optical flow

information as supplement [5], or conducting multi-task optimization [24,3], the results of our image-only and single-task architecture SurgicalSAM are promising. Besides, the encoder backbone we finetuned is the smallest ViT_b due to limited computational resources. We believe the largest ViT_h backbone can yield much better performance. Compared with the original SAM, our new architecture is of great practical significance as it can achieve semantic-level automatic segmentation. Moreover, the additionally trained parameters are only **18.28MB**, suggesting the efficiency of our finetuning strategy.

Furthermore, we have evaluated the robustness of SurgicalSAM in the face of data corruption using the EndoVis18 validation dataset. As shown in Table 3, the model's performance exhibits a significant degradation when subjected to various forms of data corruption, particularly in the case of *Blur* corruption.

7 Conclusion

In this study, we explore the robustness and zero-shot generalizability of the SAM [14] in the field of robotic surgery on two robotic instrument segmentation datasets of MICCAI EndoVis 2017 and 2018 challenges, respectively. Extensive empirical results suggest that SAM [14] is deficient in segmenting the entire instrument with point-based prompts and unprompted settings, as clearly shown in Fig. 1 and Fig. 3. This implies that SAM [14] can not capture the surgical scenes precisely despite yielding surprising zero-shot generalization ability. Besides, it exhibits challenges in accurately predicting certain parts of the instrument mask when there are overlapping instruments or only with a point-based prompt. It also fails to identify instruments in complex surgical scenarios, such as blood, reflection, blur, and shade. Moreover, we extensively evaluate the robustness of SAM [14] with a wide range of data corruptions. As indicated by Table 2 and Fig. 2, SAM [14] encounters significant performance degradation in many scenarios. To shed light on adapting SAM for surgical tasks, we fine-tuned the SAM using LoRA. Our fine-tuned SAM, i.e., SurgicalSAM, demonstrates the capability of class-wise mask prediction without any prompt.

As a foundational segmentation model, SAM [14] shows remarkable generalization capability in robotic surgical segmentation, yet it still suffers performance degradation due to downstream domain shift, data corruptions, perturbations, and complex scenes. To further improve its generalization capability and robustness, a broad spectrum of evaluations and extensions remains to be explored and developed.

Acknowledgements. This work was supported by Hong Kong Research Grants Council (RGC) Collaborative Research Fund (CRF C4063-18G and CRF C4026-21GF), Shun Hing Institute of Advanced Engineering (SHIAE project BME-p1-21) at the Chinese University of Hong Kong (CUHK), General Research Fund (GRF 14203323), Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGDX20210823103535014 (202108233000303), and (GRS) #3110167.

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Baby, B., Thapar, D., Chasmai, M., Banerjee, T., Dargan, K., Suri, A., Banerjee, S., Arora, C.: From forks to forceps: A new framework for instance segmentation of surgical instruments. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6191–6201 (2023)
4. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al.: Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
5. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 595–605. Springer (2020)
6. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)
7. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
8. Hu, C., Li, X.: When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. arXiv preprint arXiv:2304.08506 (2023)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
10. Islam, M., Atputharuban, D.A., Ramesh, R., Ren, H.: Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. IEEE Robotics and Automation Letters **4**(2), 2188–2195 (2019)
11. Islam, M., Vibashan, V., Lim, C.M., Ren, H.: St-mtl: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. Medical Image Analysis **67**, 101837 (2021)
12. Islam, M., Vibashan, V., Ren, H.: Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 8433–8439. IEEE (2020)
13. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 440–448. Springer (2019)
14. 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. arXiv preprint arXiv:2304.02643 (2023)

15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Ma, J., Wang, B.: Segment anything in medical images (2023)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
18. Seenivasan, L., Islam, M., Kannan, G., Ren, H.: Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery. arXiv preprint arXiv:2304.09974 (2023)
19. Seenivasan, L., Mitheran, S., Islam, M., Ren, H.: Global-reasoned multi-task learning model for surgical scene understanding. IEEE Robotics and Automation Letters **7**(2), 3858–3865 (2022)
20. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 624–628 (2018)
21. Wang, A., Islam, M., Xu, M., Ren, H.: Rethinking surgical instrument segmentation: A background image can be all you need. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 355–364. Springer (2022)
22. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
23. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
24. Zhao, Z., Jin, Y., Heng, P.A.: Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 11186–11193. IEEE (2022)