This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Layer-wise Gated Adaptation of SAM for Endoscopic Surgical Instrument Segmentation

Xuguang Zhu[1[0009−0003−9716−3593]], Xiankang Yu[1[0009−0009−1886−5372]], Changfei Wang[1[0009−0007−7941−2591]], Zhaochen Zhang[1], Yuichiro Hayashi[1], Masahiro Oda[2,1[0000−0001−7714−422X]], and Kensaku Mori[1,2,3[0000−0002−0100−4797]⋆]

[1] Graduate School of Informatics, Nagoya University, Nagoya, Japan
kensaku@is.nagoya-u.ac.jp
[2] Information Technology Center, Nagoya University, Nagoya, Japan
[3] Research Center for Medical Big Data, National Institute of Informatics, Tokyo, Japan

Abstract. Precise instrument segmentation underpins advanced surgical navigation. Nevertheless, deploying a pre-trained natural image vision backbone for this task typically leads to degraded segmentation owing to severe cross-domain shifts. While adopting a self-supervised vision backbone pre-trained on endoscopic data can alleviate this issue, effectively adapting such a model to dense prediction tasks remains a formidable challenge. Visual prompt tuning is widely celebrated for cross-domain transfer learning. However, its potential to facilitate intra-domain task generalization, specifically shifting a model from general scene understanding to precise semantic delineation, is largely unexplored. Motivated by this, we introduce a parameter-efficient framework that integrates an endoscopic-native Vision Transformer with a novel Gated Visual Prompt Tuning (GVPT) mechanism. GVPT dynamically regulates the injection of task-specific cues, allowing the model to gradually incorporate segmentation information while reducing perturbations to the pretrained surgical representation. Experiments on the EndoVis2018 dataset show that our method achieves 78.85% IoU, 60.17% mcIoU, and 82.56% mIoU with only 4.55M trainable parameters, highlighting the potential of prompt tuning for in-domain task adaptation in surgical image analysis.

Keywords: Segment Anything Model · Visual Prompt Tuning · Surgical Instrument Segmentation.

1 Introduction

Surgical instrument segmentation (SIS) serves as a cornerstone for computer-assisted surgery, enabling the precise localization and tracking of instruments and anatomical structures. This capability is pivotal for downstream applications

⋆ Corresponding author

such as intraoperative guidance, skill assessment, and robotic automation, which ultimately enhance surgical precision and patient outcomes [1, 15].

Prior to the era of foundation models, convolutional neural networks (CNNs) and Transformer-based variants dominated the field of SIS [21, 26, 6]. For instance, MF-TAPNet [14] utilized optical flow to capture motion priors, yet its performance often degrades when smoke or rapid instrument movements compromise accurate flow estimation. Similarly, ISINet [11] introduced a temporal consistency module to handle occlusions, but this heavy reliance on temporal continuity restricts its robustness against abrupt camera shifts or cuts. Transitioning to Transformer architectures, TraSeTR [27] proposed a track-to-segment paradigm leveraging contrastive queries. However, its data-hungry nature makes it prone to overfitting on limited surgical datasets. Lately, MATIS [5] achieved fine-grained segmentation via masked attention mechanisms, albeit at the cost of excessive computational complexity that challenges real-time clinical deployment. While effective within specific datasets, these models suffer from limited generalization capabilities. Moreover, their dependence on temporal continuity designs often restricts their robustness in surgical scenarios.

Recently, the emergence of foundation models, represented by the Segment Anything Model (SAM) [16] and its successors [19, 8], has revolutionized the field. Trained on massive datasets like SA-1B, the SAM family demonstrates remarkable zero-shot generalization and promptable segmentation capabilities, establishing a new paradigm for vision tasks [4]. Despite their success in natural images, directly applying SAM-based models to the surgical domain remains problematic due to the substantial domain shift. This discrepancy primarily stems from the unique visual characteristics of surgical scenes, which alter the underlying statistical distribution of the data. Visually, surgical videos are plagued by complex degradations rarely seen in natural environments—such as smoke, blood occlusions, and erratic illumination [18]. Mathematically, these visual anomalies result in a feature distribution that deviates significantly from the natural image statistics found in SAM's pre-training data.

To address this, we proposed gated prompt SAM (GP-SAM), which leveraged gated visual prompt tuning for the downstream surgical segmentation task. Instead of using the original encoder from SAM or SAM 2, we retrain the MAE [12] on large private surgical datasets from scratch, then use its encoder as our vision encoder to avoid the domain gap between natural and surgical images of SAM. To further fine-tune it on the surgical segmentation task, we adopted gated visual prompts to improve the performance without any adjustments of pre-trained encoder parameters. Different from LoRA, gated visual prompt tuning will not modify the architecture and data flow of the original vision encoder, avoiding degradation on small datasets. By updating only a marginal number of parameters, this parameter-efficient strategy boosts performance on the downstream segmentation task. The main contributions are listed as follows:

- We advance the SurgicalSAM by replacing its generic vision backbone with a self-supervised Vision Transformer (ViT) pre-trained on endoscopic data, which provides a robust feature representation capability.

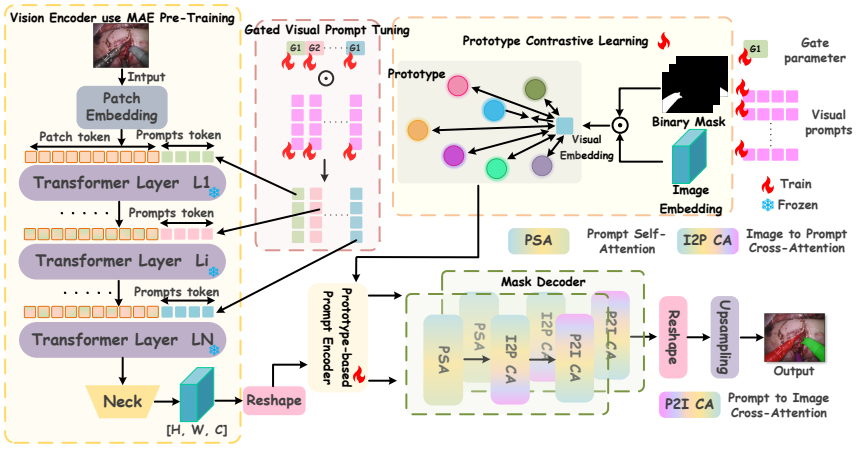


Fig. 1. The overall architecture of GP-SAM. An input endoscopic image is encoded by an MAE-pretrained Vision Transformer with frozen backbone layers. GVPT adapts the encoder by injecting layer-specific visual prompts into each Transformer block. The resulting image embedding is processed by a prototype-based prompt encoder, which uses learned class prototypes to generate dense prompts and sparse prompts. During training, prototype contrastive learning refines the prototypes. During inference, the learned prototypes are directly used by the prompt encoder to generate class-specific dense and sparse prompts. A mask decoder then fuses the prompts with the image embedding and outputs the segmentation mask.

- We propose Gated Visual Prompt Tuning (GVPT) to transition the in-domain ViT from endoscopic scene reconstruction to the surgical instrument segmentation task.
- Experiments on the EndoVis2018 dataset show that our approach achieves competitive segmentation performance, with the best IoU and mIoU scores among the compared methods.

2 Method

2.1 Overview of GP-SAM

We proposed gated prompt SAM (GP-SAM) for endoscopic scenarios illustrated in Fig. 1. Our GP-SAM leveraged an MAE encoder pretrained on large private surgical datasets and a mask decoder from SAM. Based on the module from SurgicalSAM, we further proposed gated visual prompt tuning to improve the performance on the downstream surgical instrument segmentation task.

2.2 Gated Visual Prompt Tuning

Although standard Visual Prompt Tuning (VPT) has shown remarkable success in adapting frozen backbones to downstream tasks, directly injecting randomly

initialized prompt tokens into the input sequence poses a significant challenge. The introduction of these learnable noise vectors can disrupt the feature distribution of the pre-trained encoder at the onset of training, leading to optimization instability or convergence to local minima. This is particularly detrimental in surgical scenarios where data is scarce and the domain gap is large.

To address this, we propose GVPT. Instead of forcibly inserting prompts, we introduce a learnable Adaptive Gating Mechanism that controls the injection intensity of prompts at each transformer block. By initializing this gate to zero, the task-specific prompt content is suppressed at the beginning of training, so the adaptation starts from a minimally perturbed frozen encoder rather than from strongly injected randomly initialized prompts; during optimization, the gate gradually learns how much prompt information should be incorporated at each layer. This strategy effectively adapts the pre-trained surgical representations to the segmentation task.

Given a pre-trained ViT backbone with L layers, let the input patch embeddings of the i-th layer be denoted as $\mathbf{Z}_{i-1} \in \mathbb{R}^{N\times D}$, where N is the sequence length and D is the embedding dimension. In the standard Deep-VPT, a set of learnable prompt tokens $\mathbf{P}_{i-1} \in \mathbb{R}^{M\times D}$ is concatenated with the input patch embeddings, where M is the number of prompts.

To mitigate perturbations from randomly initialized prompts, each Transformer block is equipped with a layer-specific prompt matrix $\mathbf{P}_{i-1}$ and a learnable scalar gate $\alpha_i \in \mathbb{R}$ that controls the prompt injection strength. Specifically, for the i-th Transformer layer L_i, the gated prompts $\hat{\mathbf{P}}_{i-1} \in \mathbb{R}^{M\times D}$ are formulated as

$$\hat{\mathbf{P}}_{i-1} = \alpha_i \mathbf{P}_{i-1}, \tag{1}$$

where α_i scales all prompt tokens in the i-th block. We explicitly initialize $\alpha_i = 0$ while initializing $\mathbf{P}_{i-1}$ using Xavier Uniform distribution. Although $\mathbf{P}_{i-1}$ receives no direct gradient when $\alpha_i = 0$ because $\partial\hat{\mathbf{P}}_{i-1}/\partial\mathbf{P}_{i-1} = \alpha_i$, the gate α_i can still be updated since $\partial\hat{\mathbf{P}}_{i-1}/\partial\alpha_i = \mathbf{P}_{i-1}$. After α_i moves away from zero, $\mathbf{P}_{i-1}$ is optimized normally, yielding a short warm-up behavior for layer-wise prompt injection.

Subsequently, the input sequence $\mathbf{X}_{i-1} \in \mathbb{R}^{(1+M+N)\times D}$ fed into L_i is constructed by concatenating the [CLS] token $\mathbf{x}_{i-1}^{[cls]} \in \mathbb{R}^{1\times D}$, the gated prompts $\hat{\mathbf{P}}_{i-1}$, and the input patch embeddings $\mathbf{Z}_{i-1}$ as

$$\mathbf{X}_{i-1} = \left[\mathbf{x}_{i-1}^{[cls]}, \hat{\mathbf{P}}_{i-1}, \mathbf{Z}_{i-1}\right]. \tag{2}$$

The Transformer block then processes this input sequence. Crucially, in our Deep-VPT setting, the prompts from the previous layer are discarded, and new gated prompts are injected at the current layer. The output of the i-th layer is thus computed as

$$\left[\mathbf{x}_{i}^{[cls]}, \hat{\mathbf{P}}_{i}, \mathbf{Z}_{i}\right] = L_i\left(\mathbf{X}_{i-1}\right), \tag{3}$$

where the output prompt tokens $\hat{\mathbf{P}}_i \in \mathbb{R}^{M\times D}$ are discarded before the next layer's injection. Following Deep-VPT, these prompts serve as layer-local condi-

tioning tokens rather than persistent visual tokens. This prevents the sequence length from accumulating across layers.

It allows the model to adaptively determine the importance of prompt tuning for different semantic levels. A larger $|\alpha_i|$ implies a strong reliance on domain-specific prompts, while a near-zero α_i suggests preserving the original pre-trained features. Finally, the network output is derived solely from the processed image tokens and [CLS] token.

2.3 Prototype-based Prompt Encoder and Mask Decoder

Since our main contribution lies in the gated adaptation of the image encoder, we follow SurgicalSAM [24] to use a prototype-based prompt encoder and a SAM-style mask decoder for class-specific mask prediction. Given the image embedding $\mathbf{Z}_I \in \mathbb{R}^{HW\times D}$ and the class prototype $\mathbf{e}_k \in \mathbb{R}^{1\times D}$, the prompt encoder $\mathcal{P}_E$ generates dense prompts $\mathbf{V}_d \in \mathbb{R}^{HW\times D}$ and sparse prompts $\mathbf{V}_s \in \mathbb{R}^{x\times D}$, which are given by

$$(\mathbf{V}_d, \mathbf{V}_s) = \mathcal{P}_E(\mathbf{Z}_I, \mathbf{e}_k). \tag{4}$$

The prompt encoder computes prototype-image similarity and projects the resulting cues into dense and sparse prompt embeddings, allowing each instrument class to be decoded with class-specific spatial guidance.

The mask decoder takes the sparse prompts $\mathbf{V}_s$ as token-level prompts and adds the dense prompts $\mathbf{V}_d$ to the image embedding $\mathbf{Z}_I$. A two-layer Two-Way Transformer $\mathcal{T}$ then updates the prompt tokens and image embedding as

$$(\mathbf{Q}', \mathbf{Z}_I') = \mathcal{T}(\mathbf{Q}, \mathbf{Z}_I + \mathbf{V}_d), \tag{5}$$

where $\mathbf{Q}$ denotes the concatenation of learnable output tokens and sparse prompts, $\mathbf{Q}'$ is the updated token embedding used for mask generation, and $\mathbf{Z}_I'$ is the updated image embedding after bidirectional interaction with the prompt tokens.

Following the deterministic setting of surgical instrument segmentation, we use the first mask token $\mathbf{t}_m \in \mathbb{R}^{1\times D}$ to generate a single segmentation mask $\mathbf{M}$, which is computed by

$$\mathbf{M} = \mathrm{MLP}(\mathbf{t}_m) \otimes \Phi_{\mathrm{up}}(\mathbf{Z}_I'), \tag{6}$$

where Φ_{up} denotes the upsampling module and $\otimes$ denotes the pixel-wise dot product.

The model is optimized with the Dice loss $\mathcal{L}_{\mathrm{Dice}}$ and the prototype contrastive learning loss $\mathcal{L}_{\mathrm{PCL}}$ inherited from SurgicalSAM [24]. The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{Dice}} + \mathcal{L}_{\mathrm{PCL}}. \tag{7}$$

3 Experimental Results

3.1 Dataset and Evaluation

We validated our method on EndoVis2018 [2], which contains 15 sequences from in vivo porcine nephrectomy procedures recorded by the da Vinci Xi surgical

Table 1. Comparative results on the EndoVis2018 dataset. Bold indicates the best results. "**T. Params**" means trainable parameters. Methods with "⋆" refer to zero-shot evaluation.

Method	IoU	mcIoU	mIoU	Instrument Categories							T. Params
				BF	PF	LND	MCS	UP	SI	CA	
TernausNet[13]	39.87	14.19	46.22	44.20	4.67	0.00	50.44	0.00	0.00	0.00	32.20M
MF-TAPNet[14]	39.14	24.68	67.87	69.23	6.10	11.68	70.24	0.57	14.00	0.91	37.73M
Dual-MF[26]	-	35.09	-	74.10	6.80	46.00	80.90	0.10	30.10	7.60	203.80M
ISINet[11]	70.94	40.21	73.03	73.83	48.61	30.98	88.16	2.16	37.68	0.00	162.52M
TraSeTr[27]	-	47.71	76.20	76.30	53.30	46.50	86.20	17.15	40.60	13.90	-
S3Net[6]	74.02	42.58	-	77.22	50.87	19.83	92.12	7.44	50.59	0.00	68.41M
MATIS[5]	77.01	48.65	51.58	**83.35**	38.82	40.19	**93.18**	16.17	**64.49**	4.32	68.72M
PerSAM⋆[25]	48.65	30.69	31.40	46.20	8.21	30.58	62.30	26.26	20.56	20.74	2
SurgicalSAM[24]	77.71	54.22	75.36	82.92	58.04	50.56	86.29	29.41	51.29	21.03	4.30M
TrackAnything⋆[22]	38.38	20.62	-	30.20	12.87	24.46	55.03	12.41	9.17	0.19	-
SAM2⋆[20]	60.04	58.40	-	54.44	35.11	**70.91**	70.59	74.32	27.58	35.24	-
MA-SAM2⋆[23]	64.40	**62.13**	-	56.93	44.91	66.73	75.35	**74.51**	36.82	**37.65**	-
Ours	**78.85**	60.17	**82.56**	78.67	**65.40**	69.31	89.83	35.18	52.73	30.03	4.55M

system [10] with the resolution of 1024×1280 pixels. It includes seven different tool classes: bipolar forceps (BF), prograsp forceps (PF), large needle drivers (LND), monopolar curved scissors (MCS), ultrasound probes (UP), suction instruments (SI), and clip appliers (CA). During training, we employed a data augmentation pipeline that included random scaling and cropping, horizontal and vertical flipping, rotation, and color jitter. We split 11 sequences for training and 4 sequences for validation and utilized three metrics named Intersection over Union (IoU) [3], multiclass IoU (mcIoU) [11], and mean IoU (mIoU) [9] to evaluate the performance.

3.2 Experimental setting

The image encoder used in GP-SAM was initialized from an EndoViT-style[7] self-supervised pretrained model and further pretrained on an additional private endoscopic dataset, NCC-East499. NCC-East499 contains 499 laparoscopic colon surgery videos, from which 4,233,210 frames were extracted at 1 FPS. All frames used in this additional pretraining stage were unlabeled and resized to 224×224 pixels. The private dataset has not been publicly released because official data-sharing permission has not yet been obtained.

Our proposed method was implemented using the PyTorch framework. All procedures were conducted on a workstation equipped with NVIDIA RTX 6000 Ada Generation. Moreover, to reduce GPU memory usage, we first proportionally downsample the images to 512×410 pixels using area interpolation. Subsequently, zero-padding is applied to the bottom to achieve a final resolution of 512×512 pixels. During the inference phase, we maintained the same input resolution as in training. The segmentation maps were upsampled to the original image resolution of 1024×1280 pixels for evaluation against the ground truth. We employed the AdamW [17] optimizer to train the model end-to-end. The initial learning rate was set to 2.5×10^{-4}, accompanied by a weight decay of 1×10^{-4} to mitigate overfitting. The batch size was fixed at 16, and the train-

ing process spanned 500 epochs for the EndoVis 2018 dataset. Moreover, the temperature parameter τ in the InfoNCE loss was set to 0.07.

3.3 Comparison Analysis

As detailed in Table 1, our comparative study includes seven algorithms tailored for the SIS task alongside five segmentation models adapted from SAM. Analyzing the results reveals that our approach attains the highest performance on two comprehensive metrics, yielding an IoU of 78.85% and an mIoU of 82.56%. Regarding mcIoU, our method is highly competitive, trailing the top-performing MA-SAM2 by a marginal 1.96%. While MATIS secures the highest precision for the BF, MCS, and SI instruments, it suffers severely from the inherent class imbalance of the dataset, plummeting to an IoU of merely 4.32% in the CA class. Conversely, although our framework achieves the absolute highest accuracy exclusively on the PF instrument among the seven categories, it consistently maintains top-tier segmentation precision across all remaining classes. Consequently, this balanced and robust capability translates to a strong second-place finish in overall mcIoU. The superior performance and stability of our approach can be largely attributed to the proposed GVPT. By dynamically regulating the injection of prompt information through an adaptive gating mechanism, the network smoothly transitions from general surgical scene understanding to precise instrument delineation. This strategy effectively prevents the model from overfitting to dominant classes and maintains highly generalized segmentation across diverse instrument types. The visualized segmentation results are illustrated in Fig. 2. GP-SAM demonstrates significant performance improvements in the first three scenarios, whereas its segmentation quality degrades in the fourth scenario. This degradation can primarily be attributed to the severe specular reflections and occlusions.

3.4 Ablation Studies

In this subsection, we investigated the correlation between the number of injected prompts and the overall model performance. As illustrated in Table 2, six variants (denoted as I to VI) were evaluated against the baseline model. The baseline employed a frozen ViT without any fine-tuning. Meanwhile, variants I through VI were configured with 5, 10, 20, 30, 50, and 100 prompts, respectively. As the prompt count scales up, GPU memory consumption surges drastically due to the quadratic complexity of the self-attention mechanism, accompanied by a gradual increase in the T. Ratio. Nevertheless, even at the maximum configuration of 100 prompts, the fine-tuned parameters account for a mere 1.037% of the entire ViT backbone. Regarding segmentation efficacy, the architecture achieved its optimal performance when utilizing 30 prompts, yielding an IoU of 78.85% and an mcIoU of 60.17%. Prior to this peak, the segmentation accuracy demonstrates a steady upward trend with the addition of prompts. However, performance begins to degrade beyond this threshold. These findings indicated that

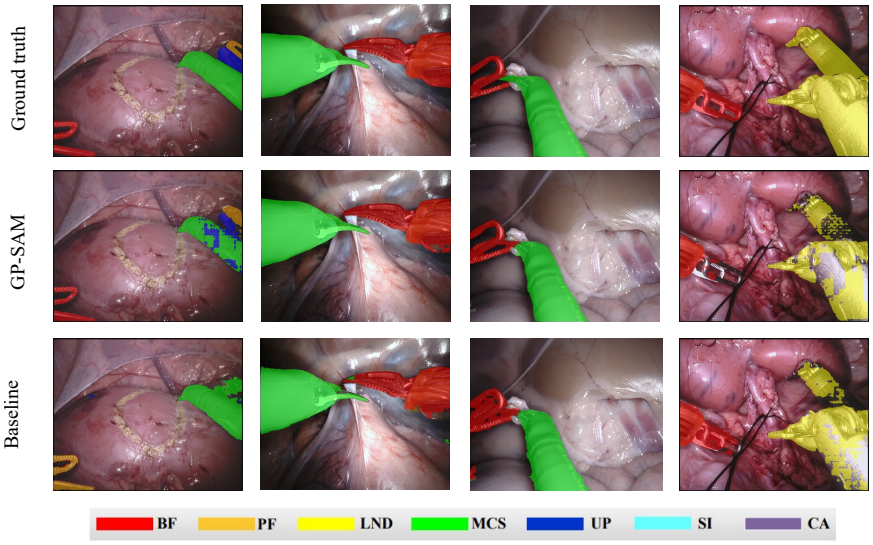


Fig. 2. Visualized segmentation results on the EndoVis2018 dataset.

Table 2. Ablation studies of different numbers of injected prompts on the EndoVis2018 dataset. Bold indicates the best results. "**T. Ratio**" is defined as the proportion of trainable prompt and gate parameters out of the total parameters in the ViT backbone.

Method	IoU	mcIoU	mIoU	Instrument Categories							T. Ratio
				BF	PF	LND	MCS	UP	SI	CA	
Baseline	67.83	45.00	69.93	71.16	39.01	55.34	80.86	21.69	26.74	20.18	0%
I	73.98	52.49	77.10	77.47	45.59	63.63	86.36	**41.28**	28.36	24.71	0.052%
II	77.23	57.39	79.64	80.85	57.69	65.36	86.37	33.07	52.73	25.63	0.105%
III	76.32	54.08	78.71	78.28	59.80	55.59	87.44	29.26	56.38	11.81	0.209%
IV	**78.85**	**60.17**	**82.56**	78.67	**65.40**	**69.31**	**89.83**	35.18	52.73	30.03	0.313%
V	76.65	54.34	77.97	81.49	45.44	55.50	86.79	32.11	44.25	34.77	0.521%
VI	78.63	58.86	81.13	**81.68**	59.34	54.12	88.18	26.34	**59.16**	**43.18**	1.037%

unconditionally expanding the prompt capacity does not necessarily translate to continuous performance gains.

4 Conclusions

In this work, we addressed the challenge of intra-domain task generalization by proposing an efficient adaptation strategy. Our results suggest that GVPT helps the network adapt pretrained endoscopic representations to fine-grained instrument segmentation on EndoVis2018. Furthermore, our prompt-number ablation indicates that increasing the number of prompt tokens does not always lead to continuous performance gains. Despite these promising results, our current framework neglects the rich temporal consistency present in surgical workflows. Additionally, as our method still operates within a fully supervised fine-tuning paradigm, it remains constrained by the severe scarcity of densely annotated sur-

gical data. Therefore, our future work will focus on integrating temporal modeling mechanisms to exploit inter-frame continuity and explicitly exploring zero-shot or few-shot learning strategies to further minimize the annotation burden in surgical scene segmentation. From an experimental perspective, future studies will also include broader cross-dataset validation, direct comparisons with alternative parameter-efficient adaptation methods, and analyzes of computational complexity and inference efficiency.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24H00720, JST Moonshot R&D Grant Number JPMJMS2214, and JST BOOST Grant Number JPMJBS2422.

Disclosure of Interest

The authors have no potential conflicts of interest to declare.

References

1. Alabi, O., Vercauteren, T., Shi, M.: Multitask learning in minimally invasive surgical vision: A review. Medical Image Analysis p. 103480 (2025)
2. 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)
3. 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)
4. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation models defining a new era in vision: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
5. Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: MATIS: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
6. 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)
7. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: EndoViT: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024)
8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)

10. Gao, S., Wang, Y., Zhou, H., Yang, K., Jiang, Y., Lu, L., Wang, S., Ma, X., Nephew, B.C., Fichera, L., Fischer, G.S., Zhang, H.K.: Laparoscopic photoacoustic imaging system integrated with the da vinci surgical system. In: Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling. SPIE Proceedings, vol. 12466. SPIE (2023)
11. González, C., Bravo-Sánchez, L., Arbelaez, P.: ISINet: An instance-based approach for surgical instrument segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 595–605. Springer (2020)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
13. Iglovikov, V., Shvets, A.: TernausNet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
14. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–448. Springer (2019)
15. Khan, U., Nawaz, U., Qayyum, A., Ashraf, S., Xie, Y., Khan, M.H., Bilal, M., Qadir, J.: Surgical scene understanding in the era of foundation ai models: A comprehensive review. arXiv preprint arXiv:2502.14886 (2025)
16. 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. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations. OpenReview.net (2019)
18. Ni, Z.L., Zhou, X.H., Wang, G.A., Yue, W.Q., Li, Z., Bian, G.B., Hou, Z.G.: SurgiNet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. Medical Image Analysis **76**, 102310 (2022)
19. Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations. OpenReview.net (2025)
20. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
21. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 624–628. IEEE (2018)
22. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
23. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented SAM2 for training-free surgical video segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–337. Springer (2025)
24. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: SurgicalSAM: Efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6890–6898 (2024)

25. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
26. Zhao, Z., Jin, Y., Gao, X., Dou, Q., Heng, P.A.: Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2020)
27. 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)