SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation

Wenxi Yue[1], Jing Zhang[1], Kun Hu[1], Yong Xia[2], Jiebo Luo[3], Zhiyong Wang[1]
[1]School of Computer Science, The University of Sydney
[2]School of Computer Science, Northwestern Polytechnical University
[3]Department of Computer Science, University of Rochester
{wenxi.yue, jing.zhang1, kun.hu, zhiyong.wang}@sydney.edu.au, yxia@nwpu.edu.cn, jluo@cs.rochester.edu

Abstract

The Segment Anything Model (SAM) is a powerful foundation model that has revolutionised image segmentation. To apply SAM to surgical instrument segmentation, a common approach is to locate precise points or boxes of instruments and then use them as prompts for SAM in a zero-shot manner. However, we observe two problems with this naive pipeline: (1) the domain gap between natural objects and surgical instruments leads to inferior generalisation of SAM; and (2) SAM relies on precise point or box locations for accurate segmentation, requiring either extensive manual guidance or a well-performing specialist detector for prompt preparation, which leads to a complex multi-stage pipeline. To address these problems, we introduce SurgicalSAM, a novel end-to-end efficient-tuning approach for SAM to effectively integrate surgical-specific information with SAM's pre-trained knowledge for improved generalisation. Specifically, we propose a lightweight *prototype-based class prompt encoder* for tuning, which directly generates prompt embeddings from class prototypes and eliminates the use of explicit prompts for improved robustness and a simpler pipeline. In addition, to address the low inter-class variance among surgical instrument categories, we propose *contrastive prototype learning*, further enhancing the discrimination of the class prototypes for more accurate class prompting. The results of extensive experiments on both EndoVis2018 and EndoVis2017 datasets demonstrate that SurgicalSAM achieves state-of-the-art performance while only requiring a small number of tunable parameters. The source code is available at *https://github.com/wenxi-yue/SurgicalSAM*.

Introduction

Surgical instrument segmentation (SIS) is a crucial task in surgical vision, aimed at precisely delineating surgical instruments in operative scenes. It provides vital assistance to surgeons and facilitates the development of advanced computer-assisted operation systems (Shademan et al. 2016; Jin et al. 2021; Liu et al. 2021; Jian et al. 2020; Yue et al. 2023; Zhang and Tao 2020). Existing deep learning methods for SIS have achieved impressive results through the design and training of specialist models featuring task-specific components. Nevertheless, these methods usually require training the complete set of model parameters (*i.e.*, full training) using SIS datasets, resulting in inefficiency. In addition, due to the limited scale of the SIS datasets, the trained models tend to exhibit subpar generalisation performance.

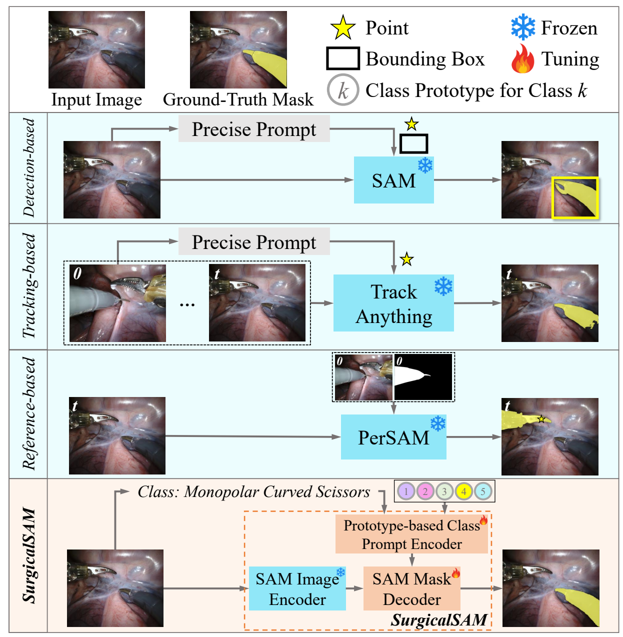
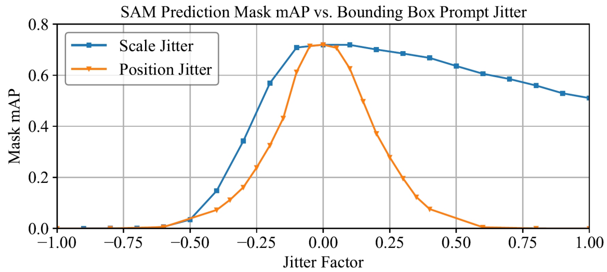
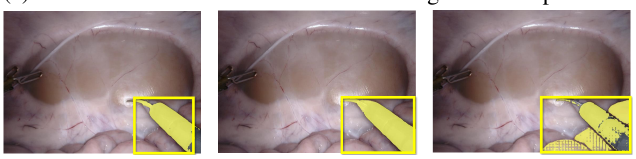


Figure 1: Comparison of our SurgicalSAM against existing detection-based, tracking-based, and reference-based zero-shot SAM frameworks for surgical instrument segmentation.

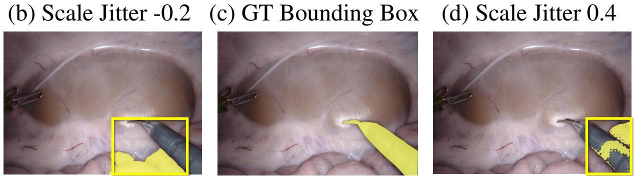
The Segment Anything Model (SAM) (Kirillov et al. 2023) has recently gained significant attention as a pioneering foundation model for promptable segmentation. Utilising SAM for downstream medical tasks holds great promise for enhancing training efficiency and leveraging strong pre-trained knowledge. Current research predominantly employs SAM in a zero-shot manner for medical image segmentation. However, the lack of sufficient medical data in SAM pre-training and the substantial domain gap between natural objects and medical targets hinders the direct generalisation of SAM towards medical tasks. Many studies have reported subpar performance of SAM in zero-shot medical image segmentation (Deng et al. 2023; He et al. 2023; Wald et al.



(a) SAM Prediction Mask mAP vs. Bounding Box Prompt Jitter



(b) Scale Jitter -0.2 (c) GT Bounding Box (d) Scale Jitter 0.4



(e) Position Jitter -0.2 (f) GT Mask (g) Position Jitter 0.4

Figure 2: Prompt robustness study of SAM against bounding box jitter in terms of scale and position for surgical instrument segmentation. A jitter factor of 0 represents the ground-truth bounding box with no jitter; a higher absolute value of the jitter factor indicates larger prompt noises.

2023; Mazurowski et al. 2023; Huang et al. 2023; Cheng et al. 2023; Wang et al. 2023a,b).

Specifically, surgical instruments differ significantly from natural objects in terms of specialised appearance, complex anatomical background, and high inter-category similarity. We evaluate three essential zero-shot SAM strategies on SIS: (1) MT-RCNN (MaskTrack-RCNN) (Yang, Fan, and Xu 2019) or Mask2Former (Cheng et al. 2022) as a bounding box detector followed by SAM, (2) Track Anything (Yang et al. 2023), and (3) PerSAM (Zhang et al. 2023), representing detection-based, tracking-based, and reference-based frameworks, respectively. As shown in Fig. 1, these methods demonstrate inferior results, where detection-based and tracking-based methods depict incorrect contours and the reference-based method misidentifies the instrument class. This further highlights the challenge of bridging the natural-surgical domain gap and emphasises the necessity of SAM tuning.

In addition, the performance of SAM relies on the precise locations of explicit prompts (Cheng et al. 2023; Wald et al. 2023). We confirm this through a prompt robustness study on SIS by introducing various scale and position jitters to the ground-truth bounding box as a prompt for SAM and recording the prediction mAP. As shown in Fig. 2, our study demonstrates SAM's sensitivity to prompt jitters: even minor deviations in the provided bounding box prompts can significantly impair segmentation accuracy. As a result, existing zero-shot SAM frameworks often involve complex multi-stage pipelines, requiring either precise manual guidance or a well-performing specialist detector to provide accurate points or bounding boxes for accurate prompting. This complexity further restricts the direct application of SAM in the surgical domain.

To address the above challenges, we propose SurgicalSAM, an end-to-end approach that effectively mitigates the surgical-natural domain gap through efficient tuning of SAM. A comparison of SurgicalSAM against existing pipelines is shown in Fig. 1. We propose a lightweight prototype-based class prompt encoder, which takes an instrument class as a prompt and learns the class prototypes by interacting with the image embedding to directly generate prompt embeddings for the mask decoder. By tuning the prototype-based class prompt encoder and the mask decoder, surgical knowledge is integrated with SAM's pre-trained knowledge, effectively mitigating the domain gap. Moreover, our strategy of directly generating latent prompt embeddings from class prompts and eliminating the use of explicit points and bounding boxes further addresses the poor robustness associated with explicit prompts as well as maintains an end-to-end pipeline.

In SurgicalSAM, the class prototypes play a vital role in effectively prompting the instrument of interest from an image. However, different surgical instrument categories often exhibit high similarity and low inter-class differences, thus posing a big challenge. To address this, we further propose contrastive prototype learning, utilising contrastive loss to acquire discriminative learned class prototypes. This method enhances the distinction between fine-grained instrument categories, resulting in more accurate class prompting and improved segmentation outcomes.

In summary, the contributions of this paper are threefold:

- We introduce SurgicalSAM to integrate surgical instrument knowledge with the pre-trained knowledge in SAM through efficient tuning for class promptable surgical instrument segmentation. It outperforms both specialist models and complex multi-stage solutions.
- We propose a prototype-based class prompt encoder that eliminates the use of explicit prompts and facilitates direct learning of latent prompt embeddings from class prompts for an end-to-end pipeline. We also propose contrastive prototype learning to enhance the discrimination of the prototypes of fine-grained instrument categories for more accurate class prompting.
- We conduct extensive experiments on the challenging EndoVis2018 and EndoVis2017 datasets, achieving state-of-the-art (SOTA) performance while significantly improving training efficiency.

Related Work

Surgical Instrument Segmentation

Current research addresses SIS by training customised specialist models. Early research employs a pixel classification paradigm to predict pixel-wise class probabilities in a frame. Notably, TernausNet pioneers this direction using a U-Net-based encoder-decoder network (Shvets et al. 2018). This has been later extended with feature pyramid attention (Ni et al. 2020) and flow-based temporal priors (Jin et al. 2019; Zhao et al. 2020). Nevertheless, these approaches encounter spatial class inconsistency, where one instrument may be assigned multiple instrument types.

An alternative paradigm is mask classification, which aims to predict a set of masks and associate each mask with a class label, inherently reducing spatial class inconsistency. ISINet introduces mask classification to instrument segmentation with Mask-RCNN (González, Bravo-Sánchez, and Arbelaez 2020; He et al. 2017). Later, Baby et al. (2023) improve its classification performance by designing a specialised classification module. In addition, TraSeTR integrates tracking cues with a track-to-segment transformer (Zhao, Jin, and Heng 2022) and MATIS incorporates temporal consistency with Mask2Former (Ayobi et al. 2023; Cheng et al. 2022). Although various methods have been proposed for surgical instrument segmentation, they primarily rely on designing specialist models and training the complete set of model parameters, which is inefficient. Particularly with the small datasets in the surgical domain, these models may exhibit subpar generalisation performance.

Segment Anything Model

SAM is recognised as a pioneering foundation model for image segmentation. The large-scale pre-training equips it with excellent zero-shot generalisation capabilities, driving various downstream applications (Wang et al. 2023c; Li et al. 2023; Yan et al. 2023). However, SAM has been shown to struggle with zero-shot generalisation to medical scenarios (Deng et al. 2023; He et al. 2023; Mazurowski et al. 2023; Huang et al. 2023; Cheng et al. 2023) due to the substantial domain gap between natural objects and medical subjects. Moreover, SAM relies on explicit points and bounding boxes at precise locations for accurate segmentation (Cheng et al. 2023; Wald et al. 2023). As a result, extensive manual guidance or a specialist detector is often required, leading to a complex multi-stage pipeline (Wang et al. 2023a).

To bridge the natural-medical domain gap, some studies seek to adapt SAM through domain-specific fine-tuning. However, they either require accurate point or bounding box prompts (Ma et al. 2023; Wu et al. 2023) or employ universal prompt embeddings for all classes which lack discrimination for fine-grained surgical instrument categories (Zhang and Liu 2023; Chen et al. 2023; Wang et al. 2023b). In contrast, our approach introduces a novel efficient-tuning approach for SAM with a prototype-based prompt encoder, which generates prompt embeddings from contrastively-learned class prototypes. This enhances the discrimination of fine-grained classes while simplifying the pipeline by eliminating the need for explicit prompts.

Methodology

Overview

In this work, we address the task of surgical instrument segmentation in a class promptable manner through efficient tuning of SAM. Specifically, given a surgical image $I \in \mathbb{R}^{H\times W\times 3}$ with spatial resolution $H \times W$ and the class of an instrument in the image c as prompt, our goal is to predict the class c mask of the image, denoted as $M^{(c)}$:

$$M^{(c)} = SurgicalSAM(I, c). \tag{1}$$

SurgicalSAM is composed of three core components as shown in Fig. 3(a): an image encoder, a prototype-based class prompt encoder, and a mask decoder. Similar to SAM, the image encoder E_I first extracts the embedding of the input image as $F_I \in \mathbb{R}^{h\times w\times d}$, with $h \times w$ denoting the shape of the image embedding and d representing the number of embedding channels. Then, our prototype-based class prompt encoder E_{CP} utilises the class prototypes B to activate the image embedding and leverages the obtained activated feature conditioned on the prompt class c to generate prompt embeddings, including dense prompt embeddings $T_D^{(c)}$ and sparse prompt embeddings $T_S^{(c)}$. Finally, the image embedding and prompt embeddings are used to predict the mask $M^{(c)}$ by the mask decoder D_M. The above process can be expressed as:

$$F_I = E_I(I), \tag{2}$$

$$T_D^{(c)}, T_S^{(c)} = E_{CP}(F_I, B, c), \tag{3}$$

$$M^{(c)} = D_M(F_I, [T_D^{(c)}, T_S^{(c)}, T_O]), \tag{4}$$

where T_O denotes the learnable output tokens in SAM.

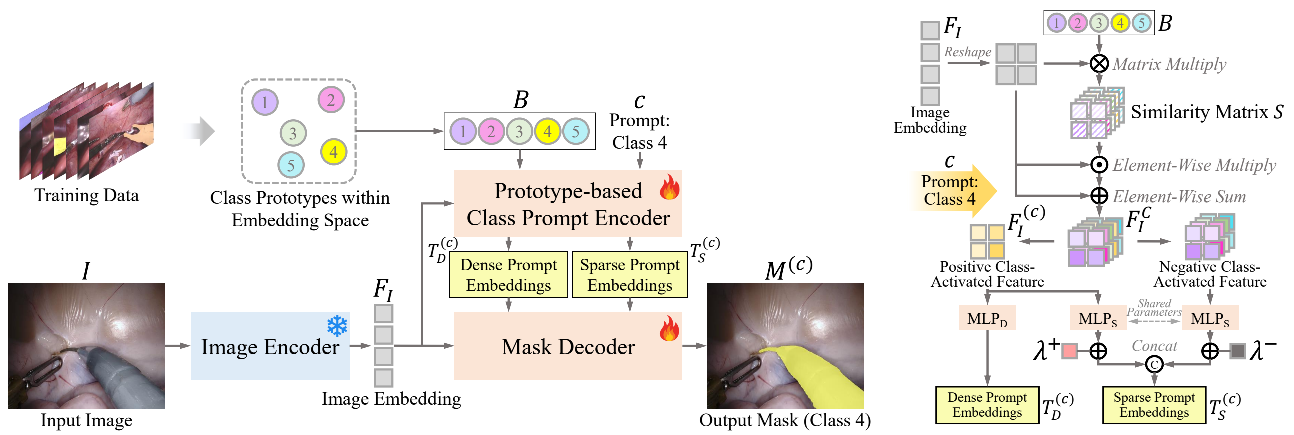
Prototype-based Class Prompt Encoder

The prototype-based class prompt encoder exploits the similarity between the image and class prototypes to create prompt embeddings. Specifically, as shown in Fig. 3(b), the spatial-wise similarity between the image embedding and the class prototype is computed to activate class-specific regions within the image, resulting in a class-activated feature to generate prompt embeddings for the mask decoder. Furthermore, inspired by the utilisation of both foreground and background point prompts in SAM, we propose to not only employ the prototype of the prompted class but integrate all class prototypes to incorporate both positive and negative cues. Such a strategy provides more robust priors for the model to effectively distinguish between instrument classes with high similarity.

Specifically, the prototype-based class prompt encoder E_{CP} is built upon a prototype bank $B = concat(\{B^{(k)}\}_{k\in\{1,2,...,C\}}) \in \mathbb{R}^{C\times d}$ consisting of a representative prototype for each class, where C is the total number of classes. Given an image I with image embedding F_I, we construct a similarity matrix $S = concat(\{S^{(k)}\}_{k\in\{1,2,...,C\}}) \in \mathbb{R}^{C\times h\times w}$ to represent the spatial-wise similarity of the image with the prototypes of all classes. It is generated by computing the dot product between the image embedding at every spatial location and each class prototype:

$$S^{(k)} = F_I \times B^{(k)}, \text{for } k \in \{1, 2, ..., C\}. \tag{5}$$

The similarity matrix is then employed as spatial attention to activate the class-specific regions, resulting in class-activated feature for all classes $F_I^C =$



(a) Overview of SurgicalSAM (b) Prototype-based Class Prompt Encoder

Figure 3: SurgicalSAM for class promptable surgical instrument segmentation through efficient tuning of SAM.

$concat(\{F_I^{(k)}\}_{k\in\{1,2,...,C\}}) \in \mathbb{R}^{C\times h\times w\times d}$:

$$F_I^{(k)} = F_I \circ S^{(k)} + F_I, \text{for } k \in \{1, 2, ..., C\}, \quad (6)$$

where $\circ$ and $+$ represents element-wise multiplication and addition, respectively, and $F_I^{(k)} \in \mathbb{R}^{h\times w\times d}$ represents the class-activated feature for class k.

Finally, the class-activated feature is used to formulate dense and sparse prompt embeddings. In SAM, dense prompt embeddings are derived from foreground masks, providing *positive* cues for segmenting the object. Imitating this, we leverage the class-activated feature of the *positive* class, *i.e.*, the prompted class c, for encoding dense prompt embeddings $T_D^{(c)} \in \mathbb{R}^{h\times w\times d}$. This is achieved through a two-layer Multilayer Perceptron (MLP):

$$T_D^{(c)} = g_D(ReLU(f_D(F_I^{(c)}))), \quad (7)$$

where f_D and g_D are two linear projection functions with intermediate dimension r_D. On the other hand, the sparse prompt embeddings in SAM are encoded from both *positive* information (foreground points and bounding boxes) and *negative* information (background points). Inspired by this, we generate sparse prompt embeddings using the class-activated feature of all classes that include both *positive*, prompted class and *negative*, non-prompted classes. The positive and negative classes are then distinguished through a pair of positive and negative embeddings. Specifically, F_I^C is first fed into a two-layer MLP to obtain positivity-agnostic sparse prompt embeddings $\hat{T}_S^C = concat(\{\hat{T}_S^{(k)}\}_{k\in\{1,2,...,C\}}) \in \mathbb{R}^{C\times n\times d}$:

$$\hat{T}_S^C = g_S(ReLU(f_S(F_I^C))), \quad (8)$$

where f_S and g_S are two linear projection functions with intermediate dimension r_S, n indicates the number of sparse tokens per class, and $\hat{T}_S^{(k)} \in \mathbb{R}^{n\times d}$ represents the positivity-agnostic sparse prompt embedding activated by class k. Then, a pair of positive and negative embeddings, $\lambda^+ \in \mathbb{R}^d$ and $\lambda^- \in \mathbb{R}^d$, are respectively added to the embeddings corresponding to positive class (class c) and negative classes (classes other than c), resulting in the final sparse prompt embeddings $T_S^{(c)} \in \mathbb{R}^{C\times n\times d}$ that are positivity-aware:

$$T_S^{(c)} = concat(\{\hat{T}_S^{(k)} + \mathbb{1}(k=c)\lambda^+ + (1-\mathbb{1}(k=c))\lambda^-\}), \text{for } k \in \{1, 2, ..., C\}. \quad (9)$$

$T_S^{(c)}$ is then reshaped to $Cn \times d$ and is fed with $T_D^{(c)}$ into the mask decoder for mask prediction.

Contrastive Prototype Learning

Our method relies on discriminative class prototypes for precise instrument category identification and accurate class region activation. However, obtaining accurate class prototypes in surgical scenarios with highly similar instrument appearances is challenging. To enhance prototype discriminativeness for more accurate class prompting, we propose contrastive prototype learning to acquire the optimised class prototypes during tuning of the framework, as illustrated in Fig. 4. Specifically, we propose prototype contrastive loss motivated by infoNCE loss (van den Oord, Li, and Vinyals 2019; Poole et al. 2019), where the class prototypes are considered as anchors and the SAM-based class embeddings in training images are regarded as samples. Given image embedding F_I, the ground-truth binary mask of class c, $G^{(c)}$, is processed to resolution $h \times w$ and used to extract the SAM-based class embedding $v^{(c)} \in \mathbb{R}^d$ for class c by averaging the foreground features:

$$v^{(c)} = \frac{\sum_i^{hw}(F_I \circ G^{(c)})}{\sum_i^{hw} G^{(c)}}. \quad (10)$$

To this end, the prototype contrastive loss is expressed as:

$$\mathcal{L}_{PCL} = -\frac{1}{C}\sum_{k=1}^{C} log\frac{exp(B^{(k)} \cdot v^{(k)}/\tau)}{\sum_{q=1}^{C} exp(B^{(k)} \cdot v^{(q)}/\tau)}, \quad (11)$$

where τ refers to the temperature parameter for modulating the similarities and $B^{(k)}$ is the prototype of class k. It can

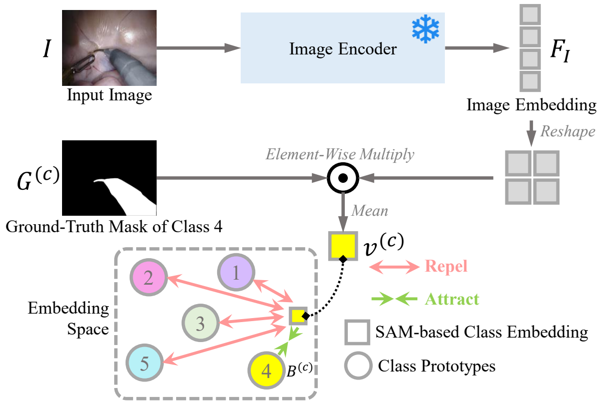


Figure 4: Contrastive Prototype Learning.

be seen that $\mathcal{L}_{PCL}$ strengthens the similarity between the prototype of class k (anchor) and the SAM-based class embeddings of k (positive samples), simultaneously suppressing the similarity between the prototype of class k (anchor) with the SAM-based class embeddings of the classes other than k (negative samples). This results in more discriminative prototype representations and enhanced surgical domain knowledge infusion through SAM tuning.

Efficient Tuning

SurgicalSAM is of high training efficiency. During tuning, the large image encoder is frozen and only the parameters of the lightweight prototype-based prompt encoder and mask decoder are updated. The tuning is end-to-end, supervised by a loss function consisting of two terms: dice loss for segmentation (Milletari, Navab, and Ahmadi 2016) and prototype contrastive loss for prototype learning:

$$\mathcal{L} = \mathcal{L}_{DICE} + \mathcal{L}_{PCL}, \tag{12}$$

$$\mathcal{L}_{DICE} = \frac{2\sum_i^{HW} m_i g_i}{\sum_i^{HW} m_i^2 + \sum_i^{HW} g_i^2}, \tag{13}$$

where m_i and g_i are the predicted logit and the ground-truth binary value at pixel i of the image, respectively.

Experiments and Discussion

Datasets and Evaluation

We use the EndoVis2018 (Allan et al. 2020) and EndoVis2017 (Allan et al. 2019) datasets and adhere to the standard protocols defined by Shvets et al. (2018) and González, Bravo-Sánchez, and Arbelaez (2020). EndoVis2017 consists of eight videos, each with 255 frames, for which we perform 4-fold cross-validation following Shvets et al. (2018). EndoVis2018 offers 11 training videos and four validation videos with each consisting of 149 frames. Both datasets provide seven instrument categories.

For evaluation, we follow prior research and adopt three segmentation metrics: Challenge IoU (Allan et al. 2019), IoU, and mean class IoU (mc IoU) (González, Bravo-Sánchez, and Arbelaez 2020; Baby et al. 2023; Ayobi et al. 2023). The efficiency of our method is evaluated in terms of training speed, training GPU usage, and inference speed.

Implementation Details

The data from EndoVis2017 and EndoVis2018 are preprocessed following Shvets et al. (2018). For the prototype-based prompt encoder, the intermediate dimensions r_D and r_S are both set to 128 and the number of tokens per class n is set to 2 and 4 for EndoVis2018 and EndoVis2017, respectively. For prototype contrastive loss, a temperature τ of 0.07 is used. In terms of training, we initialise the image encoder, the mask decoder, and the positive and negative embeddings (λ^+ and λ^-) of SurgicalSAM with SAM's pre-trained weight of the ViT-H version (Dosovitskiy et al. 2020). The image encoder and the positive and negative embeddings of our model remain frozen while the weights of the prompt encoder and mask decoder are updated. We employ an Adam optimiser with a learning rate of 0.001 and 0.0001 for EndoVis2018 and EndoVis2017, respectively. To reduce computational load, we adopt pre-computed image embeddings in training, employing a batch size of 32. Our model is implemented using PyTorch and trained and evaluated on an Nvidia Tesla V100 16GB GPU.

Main Results

The comparison of SurgicalSAM with existing methods on EndoVis2018 and EndoVis2017 are presented in Table 1 and Table 2, respectively. A visual comparison of the predictions is shown in Fig. 5. The evaluated instrument categories include *Bipolar Forceps (BF)*, *Prograsp Forceps (PF)*, *Large Needle Driver (LND)*, *Suction Instrument (SI)*, *Vessel Sealer (VS)*, *Clip Applier (CA)*, *Grasping Retractor (GR)*, *Monopolar Curved Scissors (MCS)*, and *Ultrasound Probe (UP)*. In our comparison, we categorise existing strategies into specialist models and SAM-based models. Remarkably, SurgicalSAM surpasses existing SAM-based models, matching or even exceeding the performance of SOTA specialist models, while using only a few tunable parameters.

In terms of SAM-based models, the three zero-shot SAM baselines: MT-RCNN or Mask2Former with SAM (Yang, Fan, and Xu 2019; Cheng et al. 2022) (detection-based), Track Anything (Yang et al. 2023) (tracking-based), and PerSAM (Zhang et al. 2023) (reference-based), all exhibit inferior performance. In particular, PerSAM is notably unsuitable for the task due to its reliance on a single instance for visual reference and a simple two-point prompting mechanism. Given the substantial intra-class variance and low inter-class variance among surgical instruments, a single instance lacks the necessary information for accurately referencing an instrument, resulting in missing instances in prediction, as shown in Fig. 5(b) and (d). Additionally, the use of just one foreground point and one background point fails to effectively prompt SAM for zero-shot instrument segmentation due to SAM's lack of surgical domain knowledge, leading to an incorrect interpretation of the instrument contours (Fig. 5(a), (b), and (c)). While Track Anything exhibits improved performance compared to PerSAM, its efficacy heavily relies on the quality of prompts, as shown by the large gap between the results obtained from prompting with one point versus five points. Furthermore, the significant motion of instruments often causes Track Anything to lose

Method Category	Method	Challenge IoU	IoU	mc IoU	Instrument Categories							#Params
					BF	PF	LND	SI	CA	MCS	UP	
	TernausNet	46.22	39.87	14.19	44.20	4.67	0.00	0.00	0.00	50.44	0.00	32.20M
	MF-TAPNet	67.87	39.14	24.68	69.23	6.10	11.68	14.00	0.91	70.24	0.57	37.73M
	Dual-MF	70.40	-	35.09	74.10	6.80	46.00	30.10	7.60	80.90	0.10	203.80M
Specialist Model	ISINet	73.03	70.94	40.21	73.83	48.61	30.98	37.68	0.00	88.16	2.16	162.52M
	TraSeTr	76.20	-	47.71	76.30	53.30	46.50	40.60	13.90	86.20	17.15	-
	S3Net	75.81	74.02	42.58	77.22	50.87	19.83	50.59	0.00	92.12	7.44	68.41M
	MATIS Frame	82.37	77.01	48.65	83.35	38.82	40.19	64.49	4.32	93.18	16.17	68.72M
	MT-RCNN + SAM	78.49	78.49	56.07	79.83	74.86	43.12	**62.88**	16.74	91.62	23.45	57.67M
	Mask2Former + SAM	78.72	78.72	52.50	**85.95**	**82.31**	44.08	0.00	**49.80**	**92.17**	13.18	68.72M
	TrackAnything (1 Point)	40.36	38.38	20.62	30.20	12.87	24.46	9.17	0.19	55.03	12.41	-
	TrackAnything (5 Points)	65.72	60.88	38.60	72.90	31.07	**64.73**	10.24	12.28	61.05	17.93	-
SAM-based Model	PerSAM	49.21	49.21	34.55	51.26	34.40	46.75	16.45	15.07	52.28	**25.62**	-
	PerSAM (Fine-Tune)	52.21	52.21	37.24	57.19	36.13	53.86	14.34	25.94	54.66	18.57	2
	SurgicalSAM (Ours)	**80.33**	**80.33**	**58.87**	83.66	65.63	58.75	54.48	39.78	88.56	21.23	4.65M
	GT Centroid + SAM	60.26	60.26	63.34	44.35	65.92	30.99	87.14	69.69	80.04	65.26	-
	GT Bbox + SAM	88.04	88.04	84.23	87.10	86.81	72.23	91.21	75.91	93.08	83.24	-

Table 1: Comparative Results on the EndoVis2018 Dataset. #Params represents number of tunable parameters.

Method Category	Method	Challenge IoU	IoU	mc IoU	Instrument Categories						
					BF	PF	LND	VS	GR	MCS	UP
	TernausNet	35.27	12.67	10.17	13.45	12.39	20.51	5.97	1.08	1.00	16.76
	MF-TAPNet	37.25	13.49	10.77	16.39	14.11	19.01	8.11	0.31	4.09	13.40
	Dual-MF	45.80	-	26.40	34.40	21.50	64.30	24.10	0.80	17.90	21.80
Specialist Model	ISINet	55.62	52.20	28.96	38.70	38.50	50.09	27.43	2.10	28.72	12.56
	TraSeTr	60.40	-	32.56	45.20	56.70	55.80	38.90	11.40	31.30	18.20
	S3Net	72.54	71.99	46.55	75.08	54.32	61.84	35.50	27.47	43.23	28.38
	MATIS Frame	68.79	62.74	37.30	66.18	50.99	52.23	32.84	15.71	19.27	23.90
	Mask2Former + SAM	66.21	66.21	55.26	66.84	**55.36**	**83.29**	73.52	26.24	36.26	45.34
	TrackAnything (1 Point)	54.90	52.46	55.35	47.59	28.71	43.27	82.75	**63.10**	66.46	55.54
	TrackAnything (5 Points)	67.41	64.50	62.97	55.42	44.46	62.43	**83.68**	62.59	67.03	**65.17**
SAM-based Model	PerSAM	42.47	42.47	41.80	53.99	25.89	50.17	52.87	24.24	47.33	38.16
	PerSAM (Fine-Tune)	41.90	41.90	39.78	46.21	28.22	53.12	57.98	12.76	41.19	38.99
	SurgicalSAM (Ours)	**69.94**	**69.94**	**67.03**	**68.30**	51.77	75.52	68.24	57.63	**86.95**	60.80
	GT Centroid + SAM	44.42	44.42	54.41	63.42	36.03	22.57	54.21	75.18	70.17	59.25
	GT Bbox + SAM	76.31	76.31	81.18	89.36	73.44	67.67	90.04	87.79	94.03	65.91

Table 2: Comparative Results on the EndoVis2017 Dataset.

track or confuse between instruments with similar appearances (Fig. 5(b), (c), and (d)). Detection-based SAM shows the most promising performance among the three zero-shot SAM baselines. However, its effectiveness relies on a well-trained detector model which requires significant training effort. Also, without SAM tuning, the lack of domain knowledge can result in incomplete masks or misidentification of instrument categories (Fig. 5(a), (b), and (c)).

SurgicalSAM outperforms *all* three zero-shot SAM baselines. Different from these solutions, SurgicalSAM integrates surgical domain knowledge with SAM's pre-trained general knowledge, enhancing its expertise with surgical instruments and resulting in more accurate segmentation (Fig. 5). Meanwhile, the tuning of SurgicalSAM is highly efficient, requiring significantly fewer tunable parameters than the detection-based model (4.65M for SurgicalSAM vs. 57.67M for MT-RCNN + SAM). Furthermore, SurgicalSAM utilises learned prototypes as references, which are more general and descriptive than the single instance reference in PerSAM, and eliminates the use of explicit prompts for a pipeline much simpler than the multi-stage detection-based pipeline.

We also establish two oracle scenarios by employing ground-truth centroids or ground-truth bounding boxes as prompts for SAM. As shown in Table 1 and Table 2, SurgicalSAM demonstrates substantial superiority over the utilisation of ground-truth centroids, achieving an improvement of 20.07% and 25.52% in Challenge IoU for EndoVis2018 and EndoVis2017, respectively. These promising results show that SurgicalSAM already attains superior results compared to employing basic manual guidance.

Moreover, SurgicalSAM achieves SOTA performance competitive with the specialist models while requiring substantially fewer tunable parameters (4.65M for SurgicalSAM vs. 68.72M for MATIS Frame). Particularly, significant improvements can be observed in mean class IoU, indicating that the general knowledge in foundation models serves as extra priors that help to diminish the class imbalance problem in small datasets. In summary, our method achieves promising performance with high efficiency.

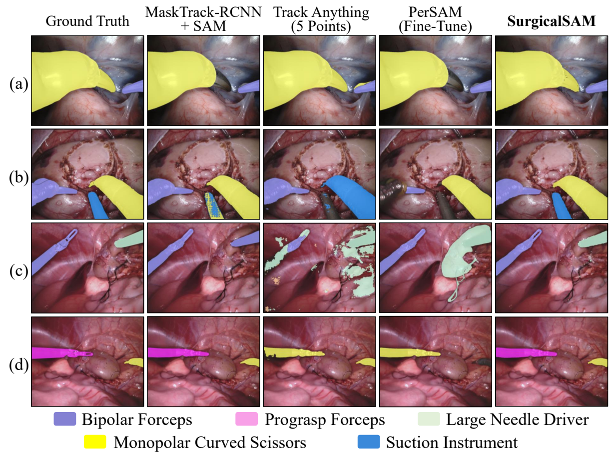


Figure 5: Visualisation of Predicted Masks.

	Challenge IoU	mc IoU	Challenge IoU	mc IoU
$n \setminus \mathcal{L}_{PCL}$	✗		✓	
2	76.38	53.95	**80.33**	**58.87**
4	78.26	56.54	79.46	58.40
6	77.28	53.71	79.67	56.97
8	76.98	53.94	80.10	58.30

Table 3: Ablation Study on SurgicalSAM.

Ablation Study

We conduct an ablation study on EndoVis2018 for contrastive prototype learning and the number of tokens n. Specifically, we remove the contrastive prototype learning module and use fixed class prototypes computed by taking the average of the class embeddings across all training samples. The results, as depicted in Table 3, show a significant difference. Without the contrastive learning process, the pre-computed fixed prototypes tend to be overly similar across different instrument categories due to their highly similar appearance. Contrastive prototype learning helps the model to learn more discriminative class prototypes and accurately identify the instrument classes. Moreover, the efficacy of contrastive prototype learning remains consistent across different numbers of tokens. Regarding the impact of different numbers of tokens on our complete model, as shown in Table 3, no notable changes can be observed. In contrast to the original SAM which is sensitive to the number of points provided (Cheng et al. 2023), the use of class prompt in our work demonstrates enhanced robustness.

Cross-Dataset Generalisation

We verify the cross-dataset generalisability of SurgicalSAM by training it on one dataset and evaluating it on another. The results are shown in Table 4, where only the instrument classes shared by both datasets are considered. Compared to the SOTA specialist model MATIS Frame, our method consistently performs better in both ways (EndoVis2018 to EndoVis2017 and EndoVis2017 to EndoVis2018). Notably, when trained on EndoVis2018 and evaluated on EndoVis2017, we achieve a large improvement of 11.43% in the IoU averaged over all classes. This underscores the advantage of SurgicalSAM over dedicated specialist models in terms of its ability to effectively generalise to new data distributions, owing to its integration of both foundation general knowledge and surgical domain expertise.

T	V	Method	Instrument Categories (IoU) BF	PF	LND	MCS	Mean IoU
18	*17*	MATIS Frame	45.57	32.62	44.98	58.84	45.50
		SurgicalSAM	**70.95**	**35.21**	**45.46**	**76.08**	**56.93**
17	*18*	MATIS Frame	**65.55**	13.89	38.25	**65.58**	45.81
		SurgicalSAM	44.50	**27.17**	**50.76**	62.94	**46.34**

Table 4: Cross-Dataset Generalisation. T: training dataset; V: validation dataset; *18*: EndoVis2018; *17*: EndoVis2017.

Method	Speed$^{\mathcal{T}}$ (fps)			Memory$^{\mathcal{T}}$ (GB)		
	bz=2	bz=16	bz=32	bz=2	bz=16	bz=32
MATIS Frame	3.1	-	-	13.1	-	-
MT-RCNN+SAM	8.2	12.8	-	3.2	13.9	-
SurgicalSAM	**40.1**	**57.4**	**59.8**	**1.9**	**5.9**	**9.6**

Method	Speed$^{\mathcal{I}}$ (fps) Online Feature	Speed$^{\mathcal{I}}$ (fps) Offline Feature
MT-RCNN+SAM	1.6	14.3
SurgicalSAM	**1.7**	**91.7**

Table 5: Complexity Analysis. $\mathcal{T}$: Training; $\mathcal{I}$: Inference.

Complexity Analysis

We conduct a complexity analysis of SurgicalSAM against the best-performing zero-shot SAM baseline (MT-RCNN + SAM) and the SOTA specialist model MATIS Frame (Ayobi et al. 2023). Their comparison regarding training efficiency across three batch sizes (bz) and inference efficiency is depicted in Table 5. In training, our method demonstrates considerably improved efficiency with notably faster speed and lower GPU memory consumption. Owing to the small number of tunable parameters, SurgicalSAM utilises less than 1/6 of the GPU memory of MATIS Frame with the same batch size, while achieving training over 10 times faster. In inference, the end-to-end pipeline of SurgicalSAM allows it to run faster than the complex multi-stage SAM baseline.

Conclusion

In this paper, we present SurgicalSAM, a novel method to efficiently tune SAM for surgical instrument segmentation. SurgicalSAM introduces a prototype-based class prompt encoder, which generates prompt embeddings directly from class prototypes. This eliminates the need for explicit points or boxes from manual guidance or specialist detectors, enabling an end-to-end pipeline and enhancing prompt robustness. We also introduce contrastive prototype learning to enhance the discriminative capability of class prototypes, improving differentiation among fine-grained instrument categories. Our method achieves state-of-the-art performance on both EndoVis2018 and EndoVis2017 with remarkable training and inference efficiency. It shows great promise for adapting SAM for surgical instrument segmentation.

Acknowledgements

This study was partially supported by Australian Research Council (ARC) grant DP210102674.

References

Allan, M.; Kondo, S.; Bodenstedt, S.; Leger, S.; Kadkhodamohammadi, R.; Luengo, I.; Fuentes, F.; Flouty, E.; Mohammed, A.; Pedersen, M.; Kori, A.; Alex, V.; Krishnamurthi, G.; Rauber, D.; Mendel, R.; Palm, C.; Bano, S.; Saibro, G.; Shih, C.-S.; Chiang, H.-A.; Zhuang, J.; Yang, J.; Iglovikov, V.; Dobrenkii, A.; Reddiboina, M.; Reddy, A.; Liu, X.; Gao, C.; Unberath, M.; Kim, M.; Kim, C.; Kim, C.; Kim, H.; Lee, G.; Ullah, I.; Luna, M.; Park, S. H.; Azizian, M.; Stoyanov, D.; Maier-Hein, L.; and Speidel, S. 2020. 2018 Robotic Scene Segmentation Challenge. arXiv:2001.11190.

Allan, M.; Shvets, A.; Kurmann, T.; Zhang, Z.; Duggal, R.; Su, Y.-H.; Rieke, N.; Laina, I.; Kalavakonda, N.; Bodenstedt, S.; Herrera, L.; Li, W.; Iglovikov, V.; Luo, H.; Yang, J.; Stoyanov, D.; Maier-Hein, L.; Speidel, S.; and Azizian, M. 2019. 2017 Robotic Instrument Segmentation Challenge. arXiv:1902.06426.

Ayobi, N.; Pérez-Rondón, A.; Rodríguez, S.; and Arbeláez, P. 2023. MATIS: Masked-Attention Transformers for Surgical Instrument Segmentation. In *ISBI*, 1–5.

Baby, B.; Thapar, D.; Chasmai, M.; Banerjee, T.; Dargan, K.; Suri, A.; Banerjee, S.; and Arora, C. 2023. From Forks to Forceps: A New Framework for Instance Segmentation of Surgical Instruments. In *WACV*, 6180–6190. IEEE.

Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; and Mao, P. 2023. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. In *ICCV Workshops*, 3367–3375.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 1290–1299.

Cheng, D.; Qin, Z.; Jiang, Z.; Zhang, S.; Lao, Q.; and Li, K. 2023. SAM on Medical Images: A Comprehensive Study on Three Prompt Modes. arXiv:2305.00035.

Deng, R.; Cui, C.; Liu, Q.; Yao, T.; Remedios, L. W.; Bao, S.; Landman, B. A.; Tang, Y.; Wheless, L. E.; Coburn, L. A.; Wilson, K. T.; Wang, Y.; Fogo, A. B.; Yang, H.; and Huo, Y. 2023. Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging. In *Medical Imaging with Deep Learning, short paper track*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

González, C.; Bravo-Sánchez, L.; and Arbelaez, P. 2020. ISINet: An Instance-Based Approach for Surgical Instrument Segmentation. In *MICCAI*, 595–605. Springer.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*, 2961–2969.

He, S.; Bao, R.; Li, J.; Stout, J.; Bjornerud, A.; Grant, P. E.; and Ou, Y. 2023. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. arXiv:2304.09324.

Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. 2023. Segment Anything Model for Medical Images? *Medical Image Analysis*, 103061.

Jian, Z.; Yue, W.; Wu, Q.; Li, W.; Wang, Z.; and Lam, V. 2020. Multitask Learning for Video-based Surgical Skill Assessment. In *DICTA*, 1–8.

Jin, Y.; Cheng, K.; Dou, Q.; and Heng, P.-A. 2019. Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video. In *MICCAI*, 440–448. Springer.

Jin, Y.; Long, Y.; Chen, C.; Zhao, Z.; Dou, Q.; and Heng, P.-A. 2021. Temporal Memory Relation Network for Workflow Recognition From Surgical Video. *IEEE Transactions on Medical Imaging*, 40(7): 1911–1923.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*, 4015–4026.

Li, Y.; Zhang, J.; Teng, X.; and Lan, L. 2023. RefSAM: Efficiently Adapting Segmenting Anything Model for Referring Video Object Segmentation. arXiv:2307.00997.

Liu, D.; Li, Q.; Jiang, T.; Wang, Y.; Miao, R.; Shan, F.; and Li, Z. 2021. Towards Unified Surgical Skill Assessment. In *CVPR*, 9522–9531.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2023. Segment Anything in Medical Images. arXiv:2304.12306.

Mazurowski, M. A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; and Zhang, Y. 2023. Segment Anything Model for Medical Image Analysis: An Experimental Study. *Medical Image Analysis*, 102918.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *3DV*, 565–571. IEEE.

Ni, Z.-L.; Bian, G.-B.; Wang, G.-A.; Zhou, X.-H.; Hou, Z.-G.; Chen, H.-B.; and Xie, X.-L. 2020. Pyramid Attention Aggregation Network for Semantic Segmentation of Surgical Instruments. In *AAAI*, volume 34, 11782–11790.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *ICML*, 5171–5180. PMLR.

Shademan, A.; Decker, R. S.; Opfermann, J. D.; Leonard, S.; Krieger, A.; and Kim, P. C. 2016. Supervised Autonomous Robotic Soft Tissue Surgery. *Science Translational Medicine*, 8(337): 337ra64–337ra64.

Shvets, A. A.; Rakhlin, A.; Kalinin, A. A.; and Iglovikov, V. I. 2018. Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. In *ICMLA*, 624–628. IEEE.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wald, T.; Roy, S.; Koehler, G.; Disch, N.; Rokuss, M. R.; Holzschuh, J.; Zimmerer, D.; and Maier-Hein, K. 2023. SAM. MD: Zero-Shot Medical Image Segmentation Capabilities of the Segment Anything Model. In *Medical Imaging with Deep Learning, short paper track*.

Wang, A.; Islam, M.; Xu, M.; Zhang, Y.; and Ren, H. 2023a. SAM Meets Robotic Surgery: An Empirical Study in Robustness Perspective. arXiv:2304.14674.

Wang, A.; Islam, M.; Xu, M.; Zhang, Y.; and Ren, H. 2023b. SAM Meets Robotic Surgery: An Empirical Study on Generalization, Robustness and Adaptation. In *MICCAI Workshops*.

Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; and Zhang, L. 2023c. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. In *NeurIPS Datasets and Benchmarks Track*.

Wu, J.; Zhang, Y.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; and Jin, Y. 2023. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv:2304.12620.

Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; and Sun, X. 2023. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023. Track Anything: Segment Anything Meets Videos. arXiv:2304.11968.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video Instance Segmentation. In *ICCV*, 5188–5197.

Yue, W.; Liao, H.; Xia, Y.; Lam, V.; Luo, J.; and Wang, Z. 2023. Cascade Multi-Level Transformer Network for Surgical Workflow Analysis. *IEEE Transactions on Medical Imaging*.

Zhang, J.; and Tao, D. 2020. Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet of Things Journal*, 8(10): 7789–7817.

Zhang, K.; and Liu, D. 2023. Customized Segment Anything Model for Medical Image Segmentation. arXiv:2304.13785.

Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Ma, X.; Dong, H.; Gao, P.; and Li, H. 2023. Personalize Segment Anything Model with One Shot. arXiv:2305.03048.

Zhao, Z.; Jin, Y.; Gao, X.; Dou, Q.; and Heng, P.-A. 2020. Learning Motion Flows for Semi-supervised Instrument Segmentation from Robotic Surgical Video. In *MICCAI*, 679–689. Springer.

Zhao, Z.; Jin, Y.; and Heng, P.-A. 2022. TraSeTR: Track-to-Segment Transformer with Contrastive Query for Instance-level Instrument Segmentation in Robotic Surgery. In *ICRA*, 11186–11193. IEEE.