SCISSR: Scribble-Conditioned Interactive Surgical Segmentation and Refinement

Haonan Ping[1,2], Jian Jiang[1], Cheng Yuan[1], Qizhen Sun[1], Lv Wu[1], and Yutong Ban[1⋆]

[1] Global College, Shanghai Jiao Tong University, Shanghai, China
[2] University of Michigan, Ann Arbor, MI, USA
yban@sjtu.edu.cn

Abstract. Accurate segmentation of tissues and instruments in surgical scenes is annotation-intensive due to irregular shapes, thin structures, specularities, and frequent occlusions. While SAM models support point, box, and mask prompts, points are often too sparse and boxes too coarse to localize such challenging targets. We present SCISSR, a scribble-promptable framework for interactive surgical scene segmentation. It introduces a lightweight Scribble Encoder that converts freehand scribbles into dense prompt embeddings compatible with the mask decoder, enabling iterative refinement for a target object by drawing corrective strokes on error regions. Because all added modules (the Scribble Encoder, Spatial Gated Fusion, and LoRA adapters) interact with the backbone only through its standard embedding interfaces, the framework is not tied to a single model: we build on SAM 2 in this work, and by communicating only through standard embedding interfaces, the design is in principle compatible with other prompt-driven segmentation architectures. To preserve pre-trained capabilities, we train only these lightweight additions while keeping the remaining backbone frozen. Experiments on EndoVis 2018 demonstrate strong in-domain performance, while evaluation on the out-of-distribution CholecSeg8k further confirms robustness across surgical domains. SCISSR achieves 95.41% Dice on EndoVis 2018 with five interaction rounds and 96.30% Dice on CholecSeg8k with three interaction rounds, outperforming iterative point prompting on both benchmarks.

Keywords: Interactive segmentation · Scribble annotation · Surgical scene segmentation

1 Introduction

Pixel-level segmentation of surgical scenes is essential for intraoperative guidance and postoperative analysis, yet large-scale annotation is expensive due to cluttered, deformable, and overlapping anatomy and tools. Interactive segmentation reduces this cost: the annotator provides a coarse prompt, receives a predicted mask, and refines it through further interaction.

⋆ Corresponding author. This work was completed while Haonan Ping was an undergraduate research assistant at Shanghai Jiao Tong University.

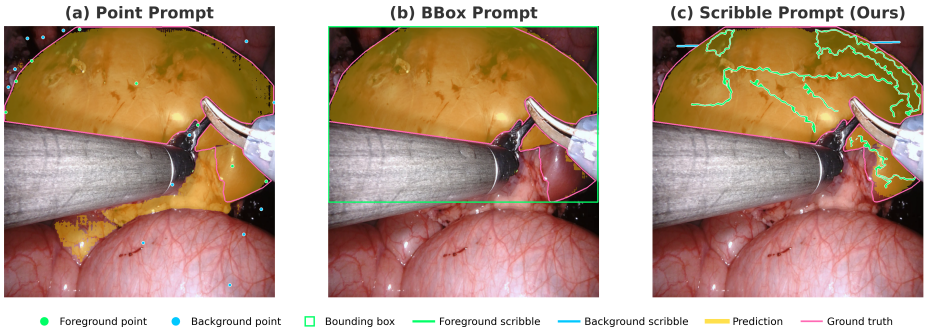

Fig. 1. Motivation for scribble prompts. Points provide sparse cues and boxes enclose large background regions, while scribbles outline the target with dense spatial coverage.

Foundation models such as SAM [9], SAM 2 [16], and SAM 3 [2] have made prompt-based segmentation practical, and medical adaptations [14,3] extend their reach. However, the supported prompt types (points, boxes, masks, and text) each have limitations in surgical scenes: bounding boxes cover large background areas around curved vessels or thin instruments, point clicks require many rounds for complex morphologies, and SAM 3's text prompts address concept-level recognition rather than instance-level delineation. As shown in Fig. 1, scribbles offer a natural middle ground: they trace the target with dense spatial coverage while requiring only slightly more effort than a click.

Prior interactive methods are predominantly click-based [26,18,19,12] (including medical variants [21,13]), and most surgical adaptations of SAM/SAM 2 retain point/box prompting [14,3,29,8,28,27]. In contrast, ScribblePrompt [23] is primarily developed for intensity-normalized (often single-channel) biomedical images, represents scribbles as per-pixel clicks rather than dense spatial embeddings, and does not leverage SAM 2's memory bank, while other works use scribbles only as weak supervision [10,20,30].

We present SCISSR, which equips SAM 2 with scribble-conditioned multi-round refinement for surgical scene segmentation. Our contributions are threefold: (1) a lightweight Scribble Encoder that maps freehand scribbles to dense prompt embeddings, enabling iterative correction via successive strokes; (2) an architecture-agnostic design where the Scribble Encoder, Spatial Gated Fusion, and LoRA [6] adapters attach through standard embedding interfaces, establishing interface-level compatibility with other prompt-driven segmentation architectures as a design principle; and (3) evaluation on both EndoVis 2018 and the out-of-distribution CholecSeg8k, confirming that scribble prompting generalizes across surgical domains (95.41% Dice on EndoVis 2018, 96.30% on CholecSeg8k).

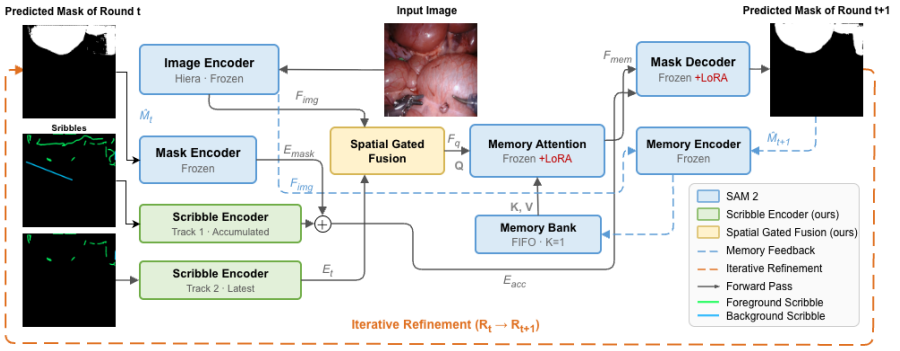

Fig. 2. Overview of SCISSR. Track 1 encodes all accumulated scribbles as dense prompt embeddings for the mask decoder; Track 2 encodes only the latest correction and injects it into the Memory Attention query via Spatial Gated Fusion. The mask is iteratively refined across rounds ($R_0 \rightarrow R_1 \rightarrow R_2 \rightarrow \cdots$). Blue: frozen SAM 2 components; green/yellow: trainable components.

2 Method

2.1 Overview

Fig. 2 illustrates SCISSR, a scribble-promptable framework for interactive refinement. Although we instantiate SCISSR on SAM 2 in this work, every added component communicates with the backbone exclusively through its standard embedding interfaces (dense prompt embeddings, memory-attention queries, and LoRA-injected projections). Given an image $I \in \mathbb{R}^{3\times H\times W}$ and a two-channel scribble map $S \in \{0,1\}^{2\times H\times W}$ (positive/foreground and negative/background), we freeze SAM 2's image encoder and introduce three lightweight components: (i) a Scribble Encoder that converts S into dense prompt embeddings, (ii) a Spatial Gated Fusion (SGF) module that injects the *latest* correction into the memory query, and (iii) a memory-driven iterative refinement loop that repurposes SAM 2's temporal memory for multi-round correction on a single image.

A key design is a **dual-track** scribble pathway. **Track 1 (Accumulated → Dense Prompt):** the union of scribbles across rounds is encoded and added to the mask decoder's dense prompt embedding, preserving the user's full intent. **Track 2 (Latest → Memory Query):** only the current round's scribble is encoded and fused into the query features of Memory Attention through SGF, encouraging attention to focus on newly corrected regions.

2.2 Scribble Encoder

The Scribble Encoder follows SAM 2's mask-prompt downscaling design. We resize S to 256×256, apply two stride-2 convolution blocks (2×2 Conv → LayerNorm2d → GELU), and a final 1×1 projection to D=256, producing a dense embedding $E_S \in \mathbb{R}^{256\times 64\times 64}$ aligned with SAM 2's image-embedding resolution. We zero the embedding when the channel input is empty.

2.3 Spatial Gated Fusion

The Spatial Gated Fusion (SGF) module injects the latest scribble into the query features of Memory Attention. A binary hard gate $g_h \in \{0, 1\}$ disables the fusion branch when no scribble is provided. When active, image features $F \in \mathbb{R}^{D \times H' \times W'}$ and scribble embedding E_S are concatenated and processed by a spatial mixing operator:

$$\texttt{SpatialMix}(\texttt{Concat}(F,\, E_S)) = \phi_{\text{dw}}\Big(\phi_{1\times1}(\texttt{Concat}(F,\, E_S))\Big),$$

where $\phi_{1\times1}$ reduces $2D$ channels back to D (with GroupNorm and GELU), and ϕ_{dw} is a 7×7 depthwise-separable convolution block that propagates the scribble signal spatially. A learnable scalar α (initialized to zero) controls the fusion strength:

$$F' = F + \alpha \cdot g_h \cdot \texttt{SpatialMix}(\texttt{Concat}(F,\, g_h \cdot E_S))\,. \tag{1}$$

Because α is initialized to zero, SGF acts as an identity mapping at the start of training and leaves the pretrained features unchanged.

2.4 Memory-Driven Iterative Refinement

SCISSR repurposes SAM 2's temporal memory for multi-round correction on a single image. Image features are computed once and cached. Track 1 accumulates all scribbles (channel-wise maximum) as a dense prompt, while Track 2 injects only the latest correction via SGF. The Memory Bank retains only the previous round's encoded features as a single memory entry for cross-attention. This suffices because Track 1 already preserves the full interaction history.

2.5 Toggleable LoRA for Scribble Adaptation

We keep SAM 2's image encoder entirely frozen. LoRA adapters [6] are inserted into the query and value projections of all multi-head attention layers in the **mask decoder** and **memory attention** module, with B zero-initialized so that $\Delta W{=}0$ at the start of training. Importantly, these adapters are toggleable: we enable them for scribble-conditioned refinement and can disable them for standard video propagation by setting the LoRA scaling to zero.

2.6 Training Pipeline

Training proceeds in two stages. **Stage 1** trains the Scribble Encoder and mask decoder LoRA without memory or SGF, using multi-round ($T{=}3$) scribble-to-mask prediction with oracle corrections. **Stage 2** jointly trains all four components (Scribble Encoder, SGF, mask decoder LoRA, Memory Attention LoRA) with the full iterative pipeline. For each training sample, we unroll $T{=}3$ rounds: initial scribbles at $t{=}0$ and error-driven corrective scribbles for $t{>}0$ (Sec. 2.7).

The loss at each round combines Focal Loss and Dice Loss with linearly increasing round weights $w_t = t + 1$:

$$\mathcal{L}_{\text{total}} = \sum_{t=0}^{T-1} \frac{w_t}{\sum_j w_j} \Big[20 \cdot \mathcal{L}_{\text{focal}}(\hat{M}_t, M^*) + \mathcal{L}_{\text{dice}}(\hat{M}_t, M^*) \Big]. \tag{2}$$

2.7 Scribble Generation

Since large-scale scribble annotations are unavailable, we synthesize scribbles from ground-truth masks. An Adaptive Scribble Generator selects one of four types per connected component based on geometric cues: *centerline* (skeleton strokes), *wave skeleton* (oscillating centerline), *contour* (boundary-following with inward offset), and *line* (negative strokes for false positives), with mild spatial perturbations to reduce the gap between synthetic and real freehand input. For correction rounds, false negatives receive positive geometry-aware scribbles and false positives receive negative cross-out strokes [23].

3 Experiments

3.1 Datasets and Metrics

We evaluate on two laparoscopic surgical datasets. **EndoVis18** [1] provides 15 training and 4 test sequences of robotic nephrectomy (2,235 / 999 frames, 10 foreground classes, 4,616 test samples). As the model is trained on this dataset, it serves as the in-distribution (ID) benchmark. **CholecSeg8k** [5] provides 8,080 cholecystectomy frames with 12 foreground classes; we hold out 1,679 frames (8,800 test samples). No CholecSeg8k data is seen during training, making it an out-of-distribution (OOD) benchmark within the endoscopic/laparoscopic surgical domain. We report IoU and Dice (%) per round as **sample-average** (*mIoU/mDice*) and **class-average** (*cIoU/cDice*); R_k denotes the prediction after k refinement rounds.

3.2 Implementation Details

We build on SAM 2 Tiny as the backbone. LoRA adapters with rank $r = 8$ and scaling factor $\alpha/r = 2$ are inserted into the query and value projections of all multi-head attention layers in the mask decoder and Memory Attention module; the image encoder remains entirely frozen. We train for 10 epochs using the AdamW optimizer with a weight decay of 0.01, a batch size of 2, and automatic mixed-precision training (AMP) on a single NVIDIA RTX 4090 (24GB), with learning rates of 1×10^{-4} for Stage-2 newly added modules and 1×10^{-5} for Stage 1 modules.

Table 1. Comparison on EndoVis 2018 (ID, N=4,616) and CholecSeg8k (OOD, N=8,800). mIoU / mDice (%). *Rk*: after k refinement rounds. N/A: not applicable.

Method	EndoVis 2018 (ID)						CholecSeg8k (OOD)			
	R0		**R2**		**R4**		**R0**		**R2**	
	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice
Point Prompt Baselines										
SAM2 Tiny (1pt/CC)	56.47	68.31	62.73	73.77	55.83	66.73	62.18	72.30	62.18	73.44
SAM2 Tiny (10pt/ch)	62.09	73.93	71.17	80.39	63.15	72.08	67.65	77.84	75.02	83.18
SAM3 (1pt/CC)	57.29	69.18	59.70	70.36	51.32	60.24	56.74	67.24	52.62	61.69
SAM3 (10pt/ch)	58.69	71.09	62.52	72.59	40.38	47.95	57.36	69.23	68.86	78.69
Bounding Box Baselines										
SAM 2 Tiny (BBox)	64.11	73.23	N/A		N/A		76.22	84.62	N/A	
SAM 3 (BBox)	69.13	77.97	N/A		N/A		77.85	85.78	N/A	
MedSAM2 (BBox)	45.87	54.54	N/A		N/A		76.74	82.69	N/A	
Supervised Baselines (trained on EndoVis 2018)										
U-Net [17]	50.70	61.50	N/A		N/A		N/A		N/A	
UPerNet [24]	58.40	66.80	N/A		N/A		N/A		N/A	
HRNet [22]	63.30	71.80	N/A		N/A		N/A		N/A	
SegFormer [25]	63.00	71.90	N/A		N/A		N/A		N/A	
SegNeXt [4]	64.30	72.50	N/A		N/A		N/A		N/A	
STSwin-CL [7]	63.60	72.00	N/A		N/A		N/A		N/A	
LSKA-Net [11]	66.20	75.30	N/A		N/A		N/A		N/A	
TAFPNet [27]	82.60	89.90	N/A		N/A		N/A		N/A	
Ours (Scribble Prompt)										
SCISSR (Adaptive)	75.72	85.18	88.32	93.49	90.18	94.61	83.42	90.24	92.30	95.82
SCISSR (Contour)	79.42	87.63	90.03	94.51	**91.60**	**95.41**	84.25	90.75	**93.22**	**96.30**
SCISSR (Wave)	73.55	83.66	88.69	93.72	90.36	94.39	82.49	89.57	92.04	95.67
SCISSR (Centerline)	74.16	84.13	85.31	91.58	86.55	92.30	83.00	90.00	90.80	94.88

3.3 Comparison with Baselines

We compare against point-based iterative methods (SAM 2 Tiny and SAM 3, each with 1 pt/CC and 10 pt/ch protocols), single bounding box baselines (SAM 2 Tiny, SAM 3, and MedSAM2 [15]), all evaluated under the same fixed-round automated protocol (Sec. 3.4). 1 pt/CC uses one click per error-region connected component (at its centroid), whereas 10 pt/ch provides up to 10 positive and 10 negative clicks per round (prioritizing connected-component centroids when selecting points).

Table 1 reports results on both datasets. On EndoVis 2018, the contour model reaches 91.60% mIoU at R4, while point-based methods often degrade with more rounds, with R4 metrics generally lower than R2. Bounding box and supervised baselines do not support iterative refinement and are therefore marked N/A for later rounds. On CholecSeg8k (OOD), our adaptive model reaches 92.30% mIoU at R2 without any CholecSeg8k training data, exceeding box baselines (76–78% mIoU) by over 14 pp. Although our model is also trained exclusively on EndoVis 2018, it generalizes well to the unseen CholecSeg8k domain, whereas supervised baselines trained on the same data are not directly applicable to CholecSeg8k due to differences in the label set.

Convergence efficiency. For interactive annotation, the practical value of a method depends on how quickly it reaches acceptable quality. Table 2 summarizes success rates and mean rounds on EndoVis 2018, showing that SCISSR

Table 2. Convergence efficiency on EndoVis 2018 (N=4,616). Success: % of masks reaching the Dice threshold within 4 rounds.

Method	Dice ≥	Success (%)	Mean Rnd	Cumulative % by round				
				R0	**R1**	**R2**	**R3**	**R4**
SAM2 Tiny (1pt/CC)	0.75	75.5	1.51	50.3	66.4	72.2	74.5	75.5
	0.85	56.4	1.76	31.2	45.3	51.6	54.7	56.4
	0.90	40.9	1.89	19.5	31.2	36.9	39.4	40.9
SAM3 (1pt/CC)	0.75	71.6	1.44	51.9	64.2	68.4	70.4	71.6
	0.85	52.9	1.75	30.1	42.2	48.4	51.2	52.9
	0.90	39.7	2.04	17.6	27.8	34.2	37.8	39.7
SCISSR (Adaptive)	0.75	**99.0**	**1.37**	70.2	92.9	97.3	98.5	99.0
	0.85	**94.7**	**1.68**	48.4	81.5	90.7	93.4	94.7
	0.90	**87.7**	**1.98**	33.3	66.9	79.2	85.2	87.7

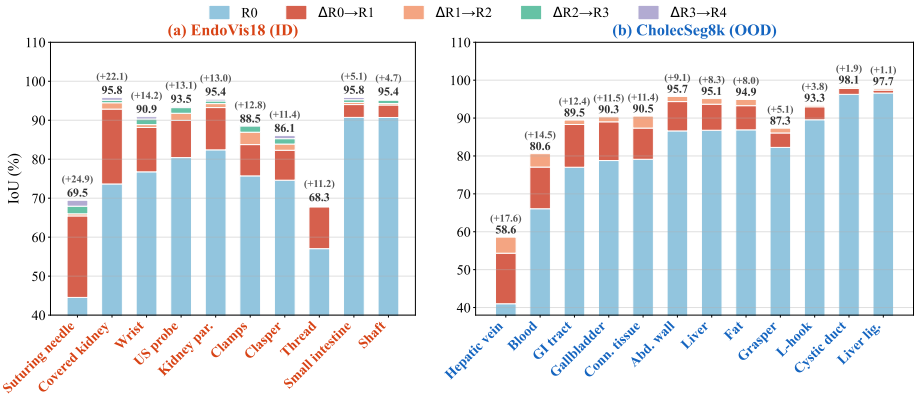

Fig. 3. Per-class IoU with incremental refinement gains. (a) EndoVis 2018 (contour, R0→R4). (b) CholecSeg8k (OOD, R0→R2). Numbers at bar tops: final IoU and total gain.

consistently converges faster and succeeds more often than point-based baselines across Dice thresholds.

Per-class analysis. Fig. 3 shows per-class results. On EndoVis 2018, classes with irregular geometry benefit most: *suturing needle* (+24.92 pp), *covered kidney* (+22.15 pp), and *wrist* (+14.16 pp). On CholecSeg8k (OOD), the model generalizes across all 12 classes without any CholecSeg8k training data.

3.4 Automated Evaluation Protocol

We use an automated protocol to simulate iterative interaction. An initial scribble is generated from the ground-truth mask (Sec. 2.7); the model predicts a mask and, if IoU $< \tau$, corrective scribbles are generated on error regions for the next round, repeating for at most T rounds. Point-click baselines follow the same protocol with clicks at the center of each error-region connected component.

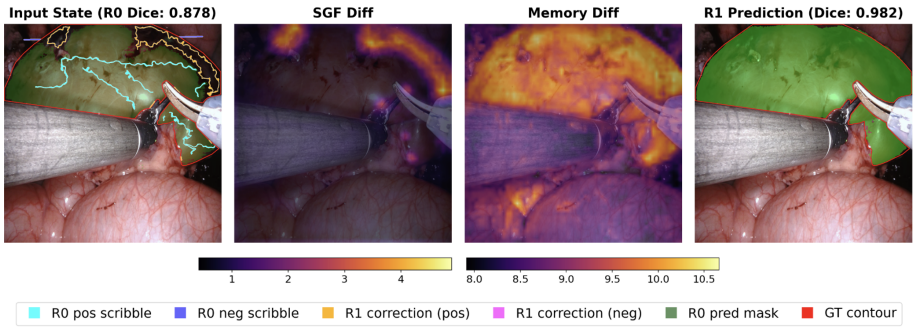

Fig. 4. Qualitative visualization of feature changes from R0→R1: SGF-induced query modification ($|F_q - F_{img}|$) and Memory-induced update ($|F_{mem} - F_{img}|$), alongside input scribbles and the refined R1 prediction.

Table 3. Component ablation on EndoVis 2018. All metrics (%) are sample-average (m) and class-average (c). Each row adds one module to the previous configuration.

Configuration	R0				R1				R2			
	mIoU	mDice	cIoU	cDice	mIoU	mDice	cIoU	cDice	mIoU	mDice	cIoU	cDice
Baseline	69.88	80.73	69.68	80.51	76.56	85.65	75.56	84.86	80.21	88.16	78.56	86.94
+ SGF	74.63	84.29	70.51	80.98	80.28	88.35	78.48	87.08	83.59	90.53	80.98	88.68
+ SGF + Memory	**75.77**	**85.06**	**71.97**	**82.09**	**85.96**	**92.03**	**82.43**	**89.47**	**88.30**	**93.49**	**84.90**	**91.18**

Limitations of point prompts. One might attribute our gains to scribbles containing more labeled pixels. However, for point prompting, both 1pt/CC and the denser 10pt/ch protocol in Table 1 can degrade as rounds increase, despite receiving more clicks. This indicates that point quantity alone is insufficient; the structured spatial layout of scribbles provides richer shape and boundary cues than isolated clicks.

3.5 Ablation Studies

Scribble Generation strategy. We evaluate four strategies (Sec. 2.7) on EndoVis 2018 and CholecSeg8k. Table 1 shows that contour-only performs best along all rounds, while centerline-only strokes lag, indicating boundary cues more effective for refinement.

Component contribution. Table 3 indicates that SGF yields consistent improvements over the baseline by injecting the latest correction into the memory-attention query, while Memory further boosts performance and its benefit becomes more evident in later rounds as correction history accumulates. Fig. 4 provides a qualitative view of these effects: SGF spatially diffuses the R1 correction scribble into the query features, while Memory Attention attends to the previous-round prediction stored in the memory bank and enhances features over relevant regions, leading to a refined mask.

4 Conclusion

We presented SCISSR, a scribble-promptable framework for interactive surgical segmentation. All added modules attach through standard embedding interfaces, supporting interface-level portability as a design principle. Experiments on EndoVis 2018 and the out-of-distribution CholecSeg8k show that scribble prompts converge faster and more accurately than point or box. Future work will extend SCISSR to video-level annotation, empirically validate portability to further SAM-family architectures, and assess usability with human annotators.

Acknowledgments. This work has been supported by the program of National Natural Science Foundation of China (No. 62503322), and the AI for Science Seed Program of Shanghai Jiao Tong University (project number 2025AI4SQY06).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

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. 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 (2025)
3. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
4. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z.N., Cheng, M.M., Hu, S.M.: SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 1140–1156 (2022)
5. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: CholecSeg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on Cholec80. arXiv preprint arXiv:2012.12453 (2020)
6. 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. In: International Conference on Learning Representations (ICLR) (2022)
7. Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.A., Stoyanov, D.: Exploring intra- and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging **41**(11), 2991–3002 (2022)
8. Kamtam, D.N., Shrager, J.B., Malla, S.D., Wang, X., Lin, N., Cardona, J.J., Yeung-Levy, S., Hu, C.: A fine-tuned foundational model SurgiSAM2 for surgical video anatomy segmentation and detection. Scientific Reports **15**, 35961 (2025)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolber, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)

10. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167 (2016)
11. Liu, M., Han, Y., Wang, J., Wang, C., Wang, Y., Meijering, E.: LSKANet: Long strip kernel attention network for robotic surgical scene segmentation. IEEE Transactions on Medical Imaging **43**(4), 1308–1320 (2024)
12. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22290–22300 (2023)
13. Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Medical Image Analysis **72**, 102102 (2021)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024)
15. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. ArXiv **abs/2504.03600** (2025), `https://api.semanticscholar.org/CorpusID:277596374`
16. Ravi, N., Gabber, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolber, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015)
18. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-brs: Rethinking backpropagating refinement for interactive segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8623–8632 (2020)
19. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: IEEE International Conference on Image Processing (ICIP). pp. 3141–3145 (2022)
20. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised CNN segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 507–522 (2018)
21. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T.: DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1559–1572 (2019)
22. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3349–3364 (2021)
23. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 207–225 (2024)
24. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)

25. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 12077–12090 (2021)
26. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep interactive object selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 373–381 (2016)
27. Yuan, C., Ban, Y.: Surgical scene segmentation by transformer with asymmetric feature enhancement. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
28. Yuan, C., Jiang, J., Yang, K., et al.: Systematic evaluation and guidelines for segment anything model in surgical video analysis. arXiv preprint arXiv:2501.00525 (2024)
29. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: SurgicalSAM: Efficient class promptable surgical instrument segmentation. In: AAAI Conference on Artificial Intelligence. pp. 6890–6898 (2024)
30. Zhang, K., Zhuang, X.: CycleMix: A holistic strategy for medical image segmentation from scribble supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11656–11665 (2022)