CSAM-HQ: A Multi-stage Refinement Framework for Surgical Instrument Segmentation based on SAM and Probabilistic Graphical Models

Xueyi Shi[1], Zhenzhen Li[3], Fucang Jia[4], Changmiao Wang[5], Tianqiao Zhang[1(✉)], and Huoling Luo[2(✉)]

[1] School of Life and Environmental Sciences, Guilin University of Electronic Technology
[2] School of Digital Media, Shenzhen University of Information Technology
[3] School of Information Engineering, Jiangxi University of Water Resources and Electric Power
[4] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[5] Shenzhen Research Institute of Big Data, Shenzhen, China
tianqiaozhang@guet.edu.cn, luohl@suit-sz.edu.cn

Abstract. Accurate surgical instrument segmentation is essential for navigation and safety in minimally invasive surgery. However, existing vision-based approaches remain susceptible to boundary ambiguity, topological fragmentation, and inter-class interference in complex surgical scenes, leading to compromised geometric precision and false positives. To address these challenges, we propose CSAM-HQ, a hierarchical refinement framework built upon the prototype-based Segment Anything Model. CSAM-HQ adopts a coarse-to-fine paradigm, where prototype-driven semantic localization first produces reliable global predictions, followed by a learnable HQ-Token with residual correction to refine boundaries and recover high-frequency geometric details, particularly for thin structures. To further mitigate inter-class conflicts, we design an IoU-aware mask quality calibration head that adaptively suppresses ambiguous predictions. In addition, a fully connected conditional random field is incorporated to enforce pixel-level spatial and appearance consistency, improving structural coherence. This dual refinement strategy enhances geometric fidelity while maintaining parameter efficiency. Extensive experiments on EndoVis2017, EndoVis2018, and CATARACTS demonstrate consistent improvements over prior methods. The proposed framework substantially improves thin instrument segmentation and reduces inter-class ambiguity across diverse surgical environments. Code is available at https://github.com/shixueyi/CSAM-HQ.

Keywords: Surgical instrument segmentation · High quality · Segment anything model · Conditional random field

1 Introduction

Minimally Invasive Surgery (MIS), particularly complex procedures such as Laparoscopic Hepatectomy (LH), has become a cornerstone of modern oncology

due to its reduced trauma and faster recovery [7, 20]. However, restricted 2D visual feedback, limited workspace, and lack of haptic sensation significantly degrade surgical perception and instrument maneuverability, increasing procedural risk [6, 12]. Therefore, accurate surgical instrument segmentation is essential for Computer-Assisted Surgery (CAS), enabling intraoperative decision-making, surgical simulation [9,21], and robotic control [1]. Recent advances in deep learning have substantially improved surgical instrument segmentation. Early Convolutional Neural Network (CNN)-based approaches explored temporal modeling and multi-task learning to address motion blur, occlusion, and appearance variation [11, 13, 14, 22, 23]. Despite their effectiveness, these methods typically rely on large-scale annotations and often exhibit limited generalization across different surgical domains. The emergence of large-scale vision foundation models, such as the Segment Anything Model (SAM) [18], provides a promising alternative through strong zero-shot generalization. Subsequent adaptations, including MFSAM [25], integrate SAM with task-specific decoders to alleviate data scarcity. To further unleash the clinical potential of foundation models, recent frameworks like SurgicalSAM [24] introduce prototype-based class prompts, while cutting-edge methods such as S3Net [5] and PartSAM [26] explore novel structural paradigms and part-level decomposition to enhance semantic discriminability. Nevertheless, despite their remarkable localization performance, these adaptations still tend to produce coarse boundaries and fragmented masks for thin or fine structures, as high-frequency geometric details are often lost during feature encoding and decoding.

To address these limitations, we propose CSAM-HQ, a Conditional Random Field (CRF)-enhanced high-quality surgical segmentation anything model. The proposed framework introduces a learnable High-Quality (HQ) token to refine mask prediction by decoupling global semantic localization from local geometric correction, while a fully connected CRF is employed as post-processing to enforce structural consistency. The main contributions of this work are threefold: (i) enhancing surgical foundation model segmentation by incorporating an HQ-Based residual refinement mechanism for boundary-aware prediction; (ii) introducing a hybrid framework that integrates foundation model semantics with CRF-based structural optimization; and (iii) extensive experiments on EndoVis2017, EndoVis2018, and CATARACTS demonstrate consistent improvements in segmentation accuracy and robustness, particularly for thin and complex surgical instruments, while maintaining high parameter efficiency.

2 Methodology

The overall architecture of CSAM-HQ is illustrated in Fig. 1. Designed for precise instrument segmentation, the framework takes a laparoscopic or endoscopic image $I \in \mathbb{R}^{H\times W\times 3}$ as input and consists of three stages: (1) feature encoding, where a pre-trained SAM image encoder extracts visual embeddings, and a prototype-based prompt encoder generates class-specific prompts; (2) HQ-Mask decoding, which integrates a learnable HQ-Token for boundary refinement; and (3) structure-aware optimization, where a fully connected CRF module executes

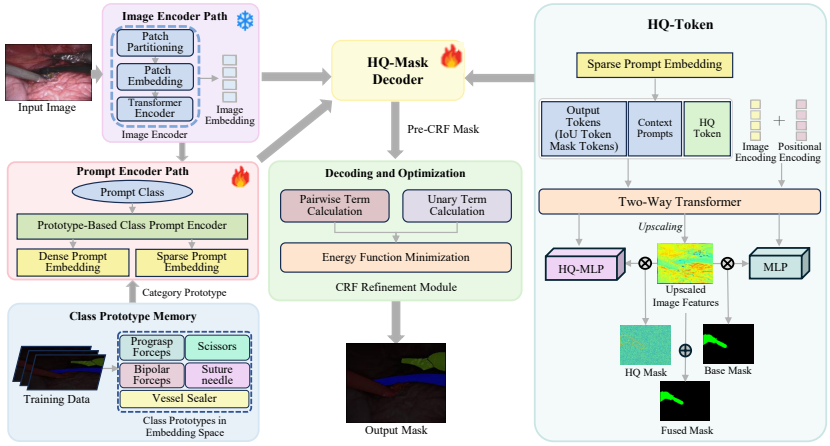

Fig. 1: Overview of CSAM-HQ. comprising an image encoder, a prototype-based prompt encoder, and an HQ-Mask decoder (detailed internal structure expanded on the right) utilizing a learnable HQ-Token and parallel HQ-MLP branch.

joint probabilistic inference to structurally integrate foundation model semantics with low-level pixel regularizations for topological consistency.

2.1 Image and Prompt Encoding

Following the standard SAM architecture, the image encoder first maps the input image I to feature embeddings $P_I \in \mathbb{R}^{h \times w \times d}$. To enable automated segmentation, we adopt the prototype-based class prompt encoder from SurgicalSAM [24]. Specifically, utilizing a learnable prototype memory bank B, the encoder computes a similarity matrix between image features and class prototypes. This similarity map serves as spatial attention to activate class-specific regions, which are subsequently projected via Multi-Layer Perceptrons (MLP) into dense embeddings $T_D^{(c)}$ and sparse embeddings $T_S^{(c)}$. These prompt embeddings, which encapsulate category-specific geometric cues, are then fed into the decoder along with P_I for mask prediction.

2.2 High-Quality Mask Refinement

To mitigate the boundary ambiguity and resolution bottlenecks inherent in the standard SurgicalSAM [24] decoder, we adapt the high-frequency feature learning mechanism from HQ-SAM [17]. Unlike its original class-agnostic design, we propose a prototype-guided HQ strategy to suit the multi-class surgical context. Specifically, we introduce a learnable HQ-Token, $T_{hq} \in \mathbb{R}^{1 \times d}$ (implemented as a learnable embedding initialized alongside standard tokens), which is explicitly conditioned on the class-specific prototype embeddings $T_S^{(c)}$ within SAM's standard two-way transformer [18]. This bidirectional transformer infrastructure executes mutual cross-attention between the image embedding and the token sequence, ensuring that the high-frequency refinement is semantically aligned with

the target instrument. The transformer input sequence is formulated as:

$$T_{in} = [T_{iou}, T_{Mask}, T_{hq}, T_S^{(c)}], \tag{1}$$

where T_{Mask} is the mask token, and T_{iou} feeds into an HQ-MLP calibration head to predict mask quality for ambiguity suppression.

We adopt a residual learning paradigm that decouples global semantic localization from local geometric refinement. While the base branch provides robust class anchoring ($M_{base}^{(c)}$), the HQ branch utilizes the updated token T'_{hq} to regress a high-frequency correction term. Formally, the final refined mask $M_{final}^{(c)}$ is obtained by fusing the coarse prediction with this refinement layer:

$$M_{final}^{(c)} = M_{base}^{(c)} + \underbrace{\sigma(\psi(T'_{hq}) \cdot \Phi_{upsampled})}_{M_{refine}^{(c)}}, \tag{2}$$

where $M_{refine}^{(c)}$ denotes the high-frequency residual refinement layer, $\Phi_{upsampled}$ represents the upsampled high-resolution features from the image encoder, $\psi(\cdot)$ is an MLP that projects the updated token T'_{hq} into dynamic channel weights to compute a token-to-feature dot product, and $\sigma(\cdot)$ denotes the Sigmoid activation function. This residual offset effectively recovers sharp contours and rectifies topological inconsistencies, particularly for slender surgical instruments.

2.3 Structure-Aware Optimization via Energy Minimization

Although the HQ decoder improves local accuracy, independent pixel-wise predictions remain prone to topological fragmentation from glare and motion blur. We address this via a fully connected CRF post-processing workflow (Fig. 2) modeled as joint probabilistic inference. Rather than proposing a novel mathematical CRF variant, our architectural contribution lies in coupling high-level foundation model semantics with low-level structural regularization. This hybrid design leverages dense spatial and appearance consistency to rectify fragmented masks and discontinuities. Specifically, the optimal label assignment $\mathbf{y}$ is obtained by minimizing the Gibbs energy $E_{\mathbf{y}}$ [19]:

$$E_{\mathbf{y}} = \sum_i \psi_u(y_i) + \sum_{(i,j)} \psi_p(y_i, y_j). \tag{3}$$

Here, the unary potential $\psi_u(y_i) = -\log P(y_i|I)$ is derived from the HQ-Mask decoder's probability map. The pairwise potential $\psi_p(y_i, y_j)$ serves as a structural regularizer, penalizing different labels for pixels with similar color and spatial proximity in image I. Minimizing Eq. 3 effectively couples high-level semantics with low-level visual cues, rectifying discontinuities in thin instruments and suppressing isolated noise.

3 Experiments and Results

3.1 Experimental Setup

Datasets and Metrics. We strictly validate our method on three public benchmarks. EndoVis2017 [4] follows the standard 4-fold cross-validation protocol; En-

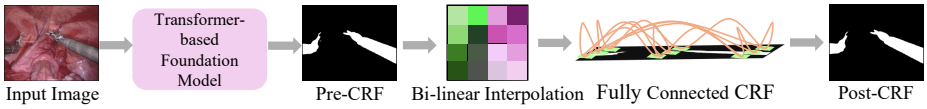

Fig. 2: The CRF post-processing workflow. It refines the coarse mask by modeling pixel-wise relationships to enforce spatial and appearance consistency.

doVis2018 [3] adopts the official training/validation split; and CATARACTS [2] includes 16 patients, with 12 for training and 4 for validation to ensure patient independence. Performance is evaluated using mean class IoU (mcIoU) [5] for semantic accuracy and Challenge IoU [8] for instance-level localization. Additionally, parameter count is reported to assess model efficiency.

Implementation Details. Input images were resized to 1024×1024, with high-level features extracted via the frozen SAM ViT-H encoder [10]. Consistent with Surgical-SAM [24], we set the prototype count $n = 2$ for EndoVis2017 and CATARACTS, and $n = 4$ for EndoVis2018. The contrastive loss was used with a temperature $\tau = 0.07$. During training, only the prompt encoder and mask decoder were fine-tuned using the Adam optimizer with a batch size of 32. The learning rates were set to 0.001 for EndoVis2017 and 0.0001 for EndoVis2018. For the CRF post-processing, the inference is performed for 2 iterations, with the standard deviations θ_γ, θ_α, and θ_β set to 3, 20, and 5, respectively. All experiments were conducted on a single NVIDIA GeForce RTX 4090D GPU.

3.2 Overall Model Performance

Comparison and Quantitative Analysis. We compare CSAM-HQ with two categories of methods: (1) specialized CNNs, including ISINet [11], TernausNet [15], and MF-TAPNet [16]; and (2) foundation models, represented by SurgicalSAM [24]. Qualitative results are presented in Fig. 3. As shown in Tables 1, CSAM-HQ achieves superior segmentation accuracy. On EndoVis2017, it yields a Challenge IoU of 71.19%, surpassing SurgicalSAM by 1.25%.

Although CSAM-HQ inherits the class-prototype prompt mechanism from SurgicalSAM [24], the baseline maintains a slight advantage on large, textureless objects (e.g., *UP*) and a few specific categories across both datasets (such as *PF* in EndoVis2017 or *BF* in EndoVis2018). This trade-off reflects an inherent design choice: while SurgicalSAM relies purely on uninterrupted prototype smoothing for global region consistency, our subsequent HQ refinement and pixel-level CRF explicitly shift the optimization focus toward recovering high-frequency geometric details and sharp boundary contours over global smoothness in homogeneous areas. Additionally, performance fluctuations on fine-structured tools like *PF* and *MCS* are inherently exacerbated by intraoperative motion blur and severe fluid specular reflections that distort prototype feature matching.

Nevertheless, on EndoVis2018, the IoU for *Suction Instrument* increases by 8.9% (from 84.43% to 93.33%) and *CA* by 14.4% (from 31.64% to 46.10%). This confirms that our coarse-to-fine strategy effectively mitigates boundary ambiguity in regions where the baseline method struggles.

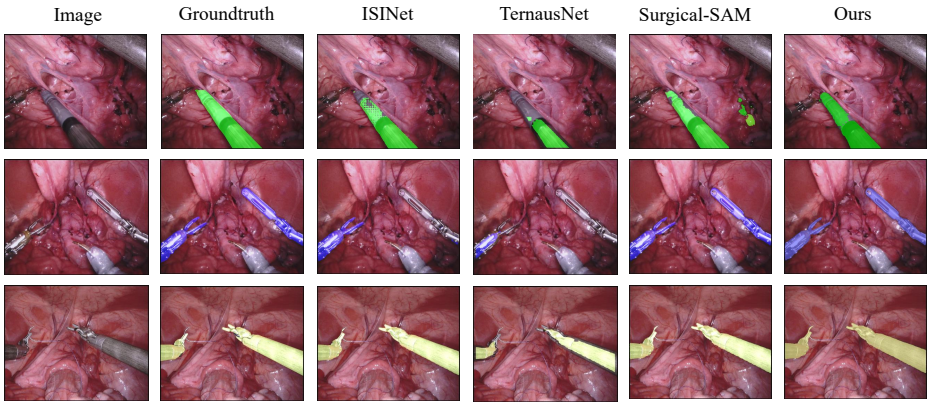

Fig. 3: Qualitative segmentation results on EndoVis2017, with each row showing a distinct challenging surgical scene.

In Table 1, 'Params' denotes strictly trainable parameters; with SAM's ViT encoder frozen, our framework requires only 4.99M tunable parameters. To evaluate computational bottlenecks, we profile the fully connected CRF. It averages 589.46 ms per frame on a CPU, introducing a predictable overhead to the end-to-end pipeline. This latency-accuracy trade-off currently hinders real-time deployment, a bottleneck we plan to alleviate via future GPU-parallel CRF implementations.

Table 1: Performance comparison of various models on EndoVis2017 and EndoVis2018 datasets. All metrics are reported in percent (%), except for Params. The best results are highlighted in **bold**. Arrows indicate the preferred direction.

Models	**Challenge IoU ↑**	**IoU ↑**	**McIoU ↑**	**cIoU per class ↑**									**Params ↓**
				BF	PF	LND	VS	GR	SI	MCS	UP	CA	
EndoVis2017 Dataset [4]													
ISINet [11]	55.62	52.20	28.96	38.70	38.50	50.09	27.43	2.10	-	28.72	12.56	-	162.52M
TernausNet [15]	35.27	12.67	10.17	13.45	12.39	20.51	5.97	1.08	-	1.00	16.76	-	32.20M
MF-TAPNet [16]	37.25	13.49	10.77	16.39	14.11	19.01	8.11	0.31	-	4.09	13.40	-	37.73M
Surgical-SAM [24]	69.94	69.94	**67.03**	68.30	**51.77**	75.52	68.24	57.63	-	**86.95**	**60.80**	-	**4.65M**
CSAM-HQ(Ours)	**71.19**	**71.19**	62.51	**72.01**	39.60	**76.14**	**69.16**	**63.65**	-	67.78	49.21	-	4.99M
EndoVis2018 Dataset [3]													
ISINet [11]	73.03	70.94	40.21	73.83	48.61	30.98	-	-	37.68	-	-	0.00	162.52M
TernausNet [15]	46.22	39.87	14.19	44.20	4.67	0.00	-	-	0.00	-	-	0.00	32.20M
MF-TAPNet [16]	67.87	39.14	24.68	69.23	6.10	11.68	-	-	14.00	-	-	0.91	37.73M
CSAM-HQ (Ours)	**74.54**	**74.54**	**71.92**	75.00	**56.53**	88.63	-	-	**93.33**	-	-	**46.10**	4.99M

BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, VS: Vessel Sealer (only EndoVis2017), GR: Grasping Retractor (only EndoVis2017), SI: Suction Instrument (only EndoVis2018), MCS: Monopolar Curved Scissors, UP: Ultrasound Probe, CA: Clip Applier (only EndoVis2018). *Note:* Challenge IoU and standard IoU are identical for SAM-based methods because the class-prototype prompt mechanism effectively eliminates false positive masks for absent categories across all frames, making the calculation scopes completely overlap.

Table 2: Generalization performance comparison on the CATARACTS dataset. The best results are highlighted in **bold**. Arrows indicate the preferred direction.

Models	Challenge IoU ↑	IoU ↑	McIoU ↑	cIoU per class ↑							Params
				Spatula	PT	LI	CF	IK	SK	KF	
Surgical-SAM	55.59	55.59	57.30	57.04	61.91	57.29	50.53	64.88	**56.59**	1.01	4.65M
CSAM-HQ (Ours)	**57.57**	**57.57**	**61.67**	**57.32**	**66.19**	**58.91**	**54.60**	**72.17**	53.43	**17.35**	4.99M

PT: Phacoemulsification Tip, LI: Lens Injector, CF: Capsulorhexis Forceps, IK: Incision Knife, SK: Slit Knife, KF: Katena Forceps.

Table 3: Component-wise ablation study on the EndoVis2017 dataset. The best performance is highlighted in **bold**.

Model	HQ Module	CRF	Resolution	IoU (%)
Surgical-SAM [24]	✗	✗	64×64	69.94
Surgical-SAM [24]	✗	✓	64×64	70.37
CSAM-HQ (Ours)	✓	✗	256×256	66.56
CSAM-HQ (Ours)	✓	✗	64×64	70.89
CSAM-HQ (Ours)	✓	✓	64×64	**71.19**

3.3 Generalization on Cross-Domain Dataset

To further validate the robustness and transferability of our proposed modules, we extended our evaluation to the CATARACTS [2] dataset. Unlike the endoscopic data in EndoVis, this dataset represents a significant domain shift involving cataract surgery instruments. In this experiment, our primary objective is to verify whether the proposed HQ-Token and CRF modules can consistently enhance the performance of the foundation model (SurgicalSAM) under cross-domain scenarios. Therefore, we conduct a direct comparison between our CSAM-HQ and the strong baseline SurgicalSAM. As shown in Table 2, CSAM-HQ demonstrates superior generalization capabilities, visualizing the agreement between model predictions and ground truth, are provided in Fig. 4. Despite the distinct visual gap between datasets, our method improves the global IoU and McIoU by 1.98% and 4.37%, respectively, compared to the baseline. Notably, for the *KF* class, which is challenging to segment due to its fine structure, our model achieves a remarkable improvement from 1.01% to 17.35%. This substantial gain confirms that our HQ and CRF modules effectively capture high-frequency boundary details that are often lost by the original foundation model during domain transfer.

3.4 Ablation Experiments

We conducted ablation studies on EndoVis2017 to validate the contributions of the CRF module and input resolution as shown in Table 3.

Effectiveness of CRF Refinement. Integrating CRF boosts Challenge IoU by 0.3%. By enforcing pixel-level appearance consistency, it suppresses isolated false positives and repairs specular reflection-induced topological discontinuities, ensuring global structural integrity where the raw decoder fails.

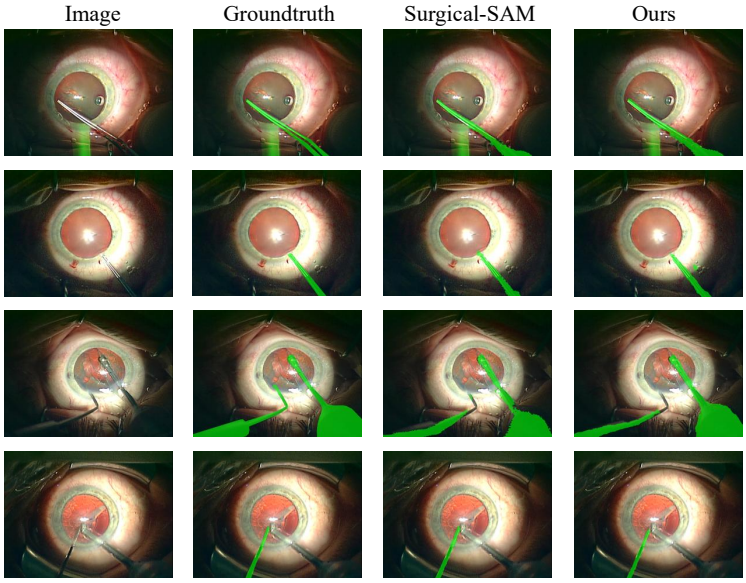

Fig. 4: Visual generalization results on the unseen CATARACTS dataset. Compared to the Surgical-SAM baseline, CSAM-HQ demonstrates superior robustness in cross-domain scenarios, accurately capturing thin instruments and fine tips that are often missed or fragmented by the baseline.

Impact of Feature Resolution. Surprisingly, 64×64 resolution outperforms 256×256 by 4.33% in IoU. We attribute this to *semantic-geometric decoupling*: high-resolution inputs introduce high-frequency noise (e.g., fluid textures) that distracts prototype matching. The 64×64 setting acts as a semantic bottleneck, robustly filtering noise for localization, while missing spatial details are fully recovered by the HQ module via high-res features $\Phi_{upsampled}$. Thus, low-resolution features ensure robust anchoring without compromising final mask quality.

4 Conclusion

In this paper, we proposed CSAM-HQ, a high-quality refinement framework built upon SurgicalSAM for precise surgical instrument segmentation. By integrating a learnable HQ-Token for boundary-aware mask refinement and a fully connected CRF for structural optimization, the proposed method effectively alleviates boundary ambiguity while maintaining high parameter efficiency. Extensive experiments on EndoVis2017, EndoVis2018, and CATARACTS demonstrate consistent improvements over strong baselines, particularly for thin and fine-structured instruments under cross-domain settings. Although our framework exhibits minor performance trade-offs on large homogeneous targets and introduces computational latency during CRF inference, it establishes a robust coarse-to-fine baseline. Future work will explore GPU-accelerated parallel processing and temporal modeling to enhance real-time efficiency and robustness in video-based surgical scene understanding.

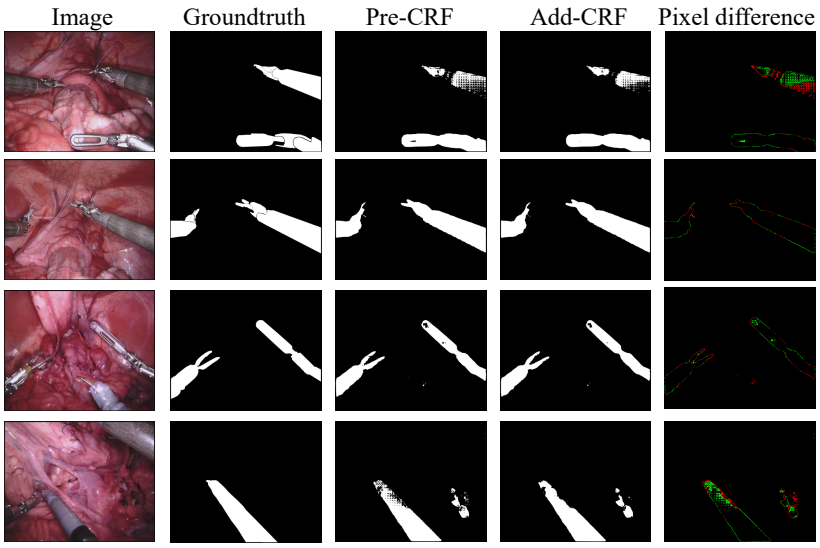

Fig. 5: Qualitative segmentation results on EndoVis2017. The last column highlights the CRF refinement: **red** pixels highlight noise or artifacts suppressed by the CRF, while **green** pixels represent instrument details and holes recovered by the CRF.

Acknowledgments. This research was supported by Guangxi Natural Science Foundation under Grant No. 2026GXNSFAA00640963, the Guangdong University Featured Innovation Program Project (No. 2025KTSCX255), and Shenzhen Institute of Information Technology Projects (Nos. SZIIT2025KJ012, SZIIT2025KJ016).

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

References

1. Ahmed, F.A., Yousef, M., Ahmed, M.A., Ali, H.O., Mahboob, A., Ali, H., Shah, Z., Aboumarzouk, O., Al Ansari, A., Balakrishnan, S.: Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review. Artificial Intelligence Review **58**(1), 1 (2024)
2. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al.: Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical Image Analysis **52**, 24–41 (2019)
3. 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)
4. 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)
5. 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)
6. Cao, Y., Zhang, J., Li, H., Ren, B., Zhao, Y., Gao, C., Zhou, Y.: A review of critical deep learning-based image guidance technologies for surgical robots. IEEE/ASME Transactions on Mechatronics **31**(1), 60–81 (2026)
7. Comptour, A., Chauvet, P., Grémeau, A.S., Figuier, C., Pereira, B., Rouland, M., Samarakoon, P., Bartoli, A., De Antonio, M., Bourdel, N.: Retrospective case control study on the evaluation of the impact of augmented reality in gynecological laparoscopy on patients operated for myomectomy or adenomyomectomy. Computer Assisted Surgery **30**(1), 2509686 (2025)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
9. Covaciu, F., Gherman, B., Rus, G., Vaida, C., Zima, I., Pisla, D.: Development of a virtual reality simulator for a robotic-assisted laparoscopic surgery. In: 2024 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR). pp. 1–6 (2024)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
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 (MICCAI). pp. 595–605 (2020)
12. Han, J., Eoh, K.J., An, M.C., Lee, S.H., Yang, S.: Enhancing endoscopic surgical imaging: Ai-based smoke removal and image enhancement. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 7055–7057 (2024)
13. Hasan, M.K., Calvet, L., Rabbani, N., Bartoli, A.: Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Medical Image Analysis **70**, 101994 (2021)
14. Huang, B., Nguyen, A., Wang, S., Wang, Z., Mayer, E., Tuch, D., Vyas, K., Giannarou, S., Elson, D.S.: Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. IEEE Transactions on Medical Robotics and Bionics **4**(2), 335–338 (2022)
15. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
16. 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 (MICCAI). pp. 440–448 (2019)
17. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. In: Advances in Neural Information Processing Systems. vol. 36, pp. 29914–29934 (2023)
18. 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 (ICCV). pp. 4015–4026 (2023)

19. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 24, pp. 109–117 (2011)
20. Ryu, Y., Hong, J., Kee, H., Park, J., Park, S.: Deep learning–driven tumor boundary estimation using robotic palpation in minimally invasive surgery. IEEE Robotics and Automation Letters **11**(3), 3350–3357 (2026)
21. Song, R., Li, Y., Meng, M.Q.H., et al.: Segmentation learning framework for multi-class surgical instrument. In: Efficient Medical Artificial Intelligence: First International Workshop, EMA4MICCAI 2025, Held in Conjunction with MICCAI 2025. p. 320 (2026)
22. Wang, Q., Zhao, S., Xu, Z., Zhou, S.K.: Lacoste: Exploiting stereo and temporal contexts for surgical instrument segmentation. Medical Image Analysis **99**, 103387 (2025)
23. Yin, Y., Luo, S., Zhou, J., Kang, L., Chen, C.Y.C.: Ldcnet: Lightweight dynamic convolution network for laparoscopic procedures image segmentation. Neural Networks **170**, 441–452 (2024)
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, P., Zhang, Z., Tao, Y.: Surgical instrument segmentation network by sam and gradient feature. In: 2024 IEEE International conference on Medical Artificial Intelligence (MedAI). pp. 21–25 (2024)
26. Zhu, Z., Wan, L., Xu, R., Zhang, Y., Chen, H., Dou, Z., Lin, C., Liu, Y., Wei, M.: Partsam: A scalable promptable part segmentation model trained on native 3d data. In: International Conference on Learning Representations (ICLR) (2026)