

TEDi: Temporal Memory-Enhanced and Denoising Transformer for Surgical Instrument Segmentation

Jiahong Yuan^{1*}, Weiming Mi^{2*}, Tao Zhang^{1**}, and Haoyin Zhou^{2**}

¹ Department of Automation, Tsinghua University, Beijing, China
yuanjh25@mails.tsinghua.edu.cn, taozhang@tsinghua.edu.cn

² Surgical Planning Laboratory, Brigham and Women's Hospital, Harvard Medical School, United States
{wemi,zhouhaoyin}@bwh.harvard.edu

Abstract. Query-based segmentation methods have shown promising potential for surgical instrument segmentation and recognition, which is essential for scene understanding and downstream tasks in computer-assisted surgery. However, most existing approaches predominantly rely on per-frame predictions and overlook cross-frame temporal priors as well as temporal-consistency constraints. This limitation often leads to unstable query representations and suboptimal category recognition. In this paper, we propose **TEDi**, a **T**emporal memory-**E**nhanced and **D**enoising transformer for surgical instrument segmentation that addresses these issues through Memory Search Enhancement and Temporal Consistency Denoising. The former introduces a query-level memory bank and a memory search enhancement encoder to retrieve discriminative representations from historical frames, enriching current-frame features. The latter constructs a temporally consistent reference as a cross-frame semantic anchor to suppress temporally unstable predictions and promote semantic coherence across frames. Extensive experiments on two benchmark datasets, EndoVis 2017 and EndoVis 2018, demonstrate that TEDi consistently outperforms state-of-the-art methods, highlighting its potential to further advance computer-assisted surgery. Our code is available at <https://github.com/argon-xixi/TEDi>.

Keywords: Surgical Instrument Segmentation · Query-based Segmentation · Transformers · Deep Learning.

1 Introduction

Robot-assisted minimally invasive surgery (RAMIS) leverages dexterous articulated instruments and high-fidelity endoscopic imaging to enable precise manipulation while reducing surgical trauma and accelerating postoperative recovery [8,19]. Accurate surgical instrument segmentation is fundamental to surgical

* Co-first authors

** Corresponding authors

scene understanding [12,4], supporting downstream tasks such as instrument tracking [23], pose estimation [20], and trajectory prediction [25], and serving as a critical component for next-generation surgical robotic systems [13,18,17].

Early methods formulated instrument segmentation as pixel-wise classification, which often suffers from spatial class inconsistency [24,27]. Recently, query-based segmentation (QBS) frameworks, such as Mask2Former [6], have gained increasing attention due to their unified set-prediction paradigm that jointly models instance masks and categories. Several studies have adapted QBS models to surgical scenarios. ISINet [10] builds upon Mask R-CNN [11] for instrument segmentation, while other approaches improve QBS-based models by enhancing the discriminability of query representations [5] and improving query initialization [9]. Beyond single-frame prediction, temporal cues have been introduced to further improve performance and cross-frame consistency, including leveraging historical queries for tracking [28] and extending Mask2Former with video Transformers [3] or spatio-temporal context modeling to strengthen temporal consistency and robustness [26].

However, empirical evidence suggests that, in surgical videos, QBS models typically achieve accurate localization while remaining prone to severe category prediction errors [3,26]. These errors manifest as (i) misclassification among visually confusable instruments and (ii) inconsistent class predictions for the same instance across adjacent frames, resulting in temporally unstable semantics. This indicates that query representations learned primarily under single-frame supervision are insufficient to exploit temporal cues as stable priors for fine-grained category discrimination.

To address these limitations, we propose a temporal-enhanced framework, TEDi, for surgical instrument segmentation, built upon two key principles: **Memory Search Enhancement (MSE)** and **Temporal Consistency Denoising (TCD)**. By explicitly mining temporal cues, TEDi systematically reduces category prediction errors and improves temporal semantic stability.

Our main contributions are summarized as follows:

- We propose **query-level memory bank** and **memory search enhancement encoder** that retrieve target representations from historical frames to provide cross-frame semantic priors for current-frame segmentation.
- We develop a **window-based temporal-consistency denoising decoder** that constructs a temporally stable reference and enforces identity-aligned temporal correction to suppress unreliable query representations.

Extensive experiments on two EndoVis benchmark datasets demonstrate that our method achieves competitive or superior performance compared with state-of-the-art approaches, with particularly notable improvements on confusable categories and temporal stability.

2 Method

Our framework follows a query-centric pipeline. For each frame, we first extract instance-aware query embeddings and refine them via Memory Search En-

hancement (MSE), which retrieves cross-frame semantic priors from a query-level memory bank to strengthen the current-frame representations. We then apply Temporal Consistency Denoising (TCD), a local temporal-window mechanism that denoises unreliable queries, suppresses unstable predictions, and promotes semantic coherence across frames. Finally, the denoised queries are fed into the mask prediction head to produce the segmentation masks.

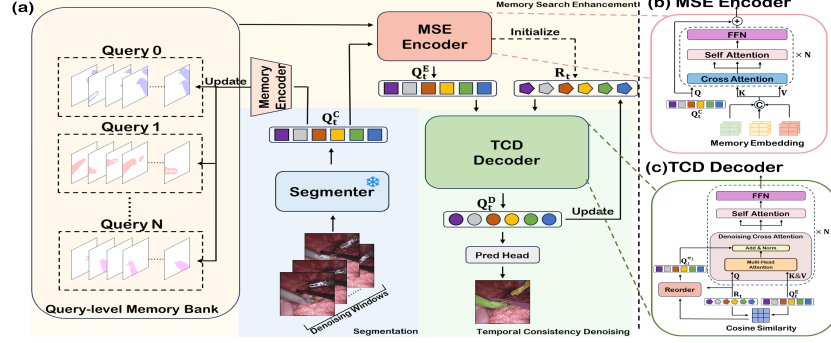


Fig. 1. The overview of TEDi. Panel (a) illustrates the overall pipeline of TEDi, while (b) and (c) depict the architectures of the MSE Encoder and the TCD Decoder.

2.1 Preliminaries

Baseline Segmentation Model. We adopt Mask2Former [6] as the baseline, a representative query-based segmentation framework that formulates segmentation as predicting categories and masks for a fixed set of learnable object queries. Given an input image $I \in \mathbb{R}^{H \times W}$, the backbone and pixel decoder produce a feature map F , which interacts with N object queries $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^N$, $\mathbf{q}_i \in \mathbb{R}^d$ through a transformer decoder. Each query outputs a mask prediction $\mathbf{m}_i \in [0, 1]^{H \times W}$ and a classification score $\mathbf{c}_i \in \mathbb{R}^C$, forming a set of segmentation candidates.

2.2 Proposed Methodology

Memory Search Enhancement Encoder. Memory mechanisms are widely adopted in video segmentation to enhance current predictions using representations from previous frames [21,15]. However, storing dense pixel-level features often introduces redundancy and weakens retrieval effectiveness [7,29]. We instead propose a **query-level memory bank** with a **memory search enhancement encoder**, storing compact instance-aware queries. By encoding predicted masks and injecting mask-aware representations into the queries, the memory remains both efficient and spatially informative.

Firstly, we fine-tune the segmenter and then freeze its parameters to produce a set of query-level predictions pairs $\mathcal{H}_t = \{\mathbf{q}_{t,i}, \mathbf{c}_{t,i}, \mathbf{m}_{t,i}\}_{i=1}^N$ for frame t .

Secondly, we encode the mask $\mathbf{m}_{t,i}$ into the query embedding space:

$$\mathbf{z}_{t,i} = \phi_m(\mathbf{m}_{t,i}) \in \mathbb{R}^d, \quad (1)$$

and fuse it with the original query $\mathbf{q}_{t,i}$ using a learnable gating function to obtain the query-level memory:

$$\mathbf{e}_{t,i} = \mathbf{g}_{t,i} \odot \mathbf{q}_{t,i} + (1 - \mathbf{g}_{t,i}) \odot \mathbf{z}_{t,i}, \quad \mathbf{g}_{t,i} = \sigma(\phi_g(\mathbf{q}_{t,i}, \mathbf{z}_{t,i})) \in \mathbb{R}^d, \quad (2)$$

where $\phi_m(\cdot)$ and $\phi_g(\cdot)$ denote learnable projection functions, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. Hence, the memory of frame t is defined as $\mathbf{E}_t = \{\mathbf{e}_{t,i}\}_{i=1}^N \in \mathbb{R}^{N \times d}$.

Thirdly, a fixed-capacity memory bank $\mathcal{M}_t$ containing query embeddings from the recent K frames is maintained by updating online in a FIFO manner.

$$\mathcal{M}_t = \{\mathbf{E}_\tau\}_{\tau \in \mathcal{T}_t}, \quad \mathcal{T}_t = \{\tau \mid t - K \leq \tau \leq t - 1\}. \quad (3)$$

Finally, given the current-frame queries $\mathbf{Q}_t^C = \{\mathbf{q}_{t,i}\}_{i=1}^N \in \mathbb{R}^{N \times d}$ and the concatenated stored memories along the query dimension $\mathbf{M}_t = \text{Stack}(\mathcal{M}_t) \in \mathbb{R}^{KN \times d}$, we obtain enhanced queries by retrieving cross-frame semantic priors via stacked memory attention blocks:

$$\mathbf{Q}_t^E = \text{MemAttn}(\mathbf{Q}_t^C, \mathbf{M}_t). \quad (4)$$

The enhanced queries $\mathbf{Q}_t^E = \{\mathbf{q}_{t,j}^E\}_{j=1}^N \in \mathbb{R}^{N \times d}$ are then used for subsequent temporal denoising and prediction.

Temporal Consistency Denoising Decoder. The baseline segmenter is trained in a frame-wise manner and therefore lacks explicit temporal constraints. When applied to videos, its query predictions may become temporally inconsistent or overly confident yet incorrect. Such instability can propagate across frames and impair segmentation reliability [26]. To address this limitation, we introduce a **temporal consistency denoising decoder** that refines query embeddings through identity-aligned temporal correction, enforcing coherence across frames and suppressing per-frame noisy query representations.

For frame t , we maintain a Temporal Consistency Reference $\mathbf{R}_t = \{\mathbf{r}_{t,i}\}_{i=1}^N \in \mathbb{R}^{N \times d}$, which is defined as the denoised queries from the previous frame and serves as a temporally stable semantic anchor:

$$\mathbf{R}_t = \begin{cases} \mathbf{Q}_t^E, & t = 0, \\ \mathbf{Q}_{t-1}^D, & t > 0. \end{cases} \quad (5)$$

Then, we evaluate the similarity between the Reference and the enhanced queries, and obtain a one-to-one assignment π_t via the Hungarian algorithm:

$$\pi_t = \arg \max_{\pi} \sum_{i=1}^N \mathbf{S}_{i,\pi(i)}, \quad \mathbf{S}_{i,j} = \frac{\langle \mathbf{r}_{t,i}, \mathbf{q}_{t,j}^E \rangle}{\|\mathbf{r}_{t,i}\|_2 \|\mathbf{q}_{t,j}^E\|_2}. \quad (6)$$

Hence, the enhanced queries can be subsequently reordered to align with $\mathbf{R}_t$:

$$\mathbf{Q}_t^{\pi_t} = \text{Reorder}(\mathbf{Q}_t^E, \pi_t). \quad (7)$$

Finally, we refine the queries with L cascaded denoising blocks using temporal consistency cues, where the core denoising cross-attention is defined as:

$$\tilde{\mathbf{Q}}_t^\ell = \tilde{\mathbf{Q}}_t^{\ell-1} + \text{CA}(\mathbf{R}_t, \mathbf{Q}_t^E, \mathbf{Q}_t^E), \quad \tilde{\mathbf{Q}}_t^0 = \mathbf{Q}_t^{\pi_t}, \quad (8)$$

here, the residual term is applied on $\tilde{\mathbf{Q}}_t^{\ell-1}$, ensuring that the decoder preserves the current-frame representation and only injects temporally consistent information. Within each block, we further apply self-attention and a FFN Network.

After L iterations, the denoised queries become the output of the decoder $\mathbf{Q}_t^D = \tilde{\mathbf{Q}}_t^L$, as well as the reference $\mathbf{R}_{t+1}$ for the next frame. To prevent long-term error accumulation, the temporal reference is reset every T frames.

Optimization. We train TEDi using a classification loss $\mathcal{L}_{Cls}$ and a mask prediction loss $\mathcal{L}_{Mask}$. Specifically, we adopt focal loss [14] for $\mathcal{L}_{Cls}$, and combine cross-entropy (CE) loss and Tversky loss [22] for mask supervision. The overall loss is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{Cls} + \lambda_2 \mathcal{L}_{ce} + \lambda_3 \mathcal{L}_{Tversky}, \quad (9)$$

where λ_1 , λ_2 , and λ_3 denote the loss weights. In practice, we set $\lambda_1 = 2$, $\lambda_2 = 5$, and $\lambda_3 = 5$.

3 Experiments

All experiments are conducted using publicly available datasets. We evaluate TEDi on two public benchmarks for surgical instrument segmentation, EndoVis 2017(EV17) [2] and EndoVis 2018(EV18) [1]. To ensure fair comparisons with prior work, we strictly follow the commonly adopted dataset splits and label settings used in previous studies. For EV17, we follow [24] and report results under the 4-fold cross-validation protocol. For EV18, We adopt the validation splits and the instrument category annotations provided in [10].

We report performance using three commonly used metrics in this field: challenge IoU (Ch_IoU), ISINet IoU (ISI_IoU), and mean class IoU (mc_IoU).

3.1 Implementation Details

We use Swin-Small [16] as the backbone and set the number of object queries to $N = 10$ to reduce redundancy. The memory bank capacity and denoising window length are set to $K = 3$ and $T = 3$, respectively. Memory bank initialized from 3 preceding frames. Training is performed with Adam at a learning rate of 6×10^{-5} . Inference is restricted to the last frame of each window to maintain strict causality. All experiments are conducted on a single NVIDIA A100 GPU.

3.2 Comparison with State-of-the-Art Methods

We compare TEDi with representative approaches, including single-frame models (QPD [9], S3Net [5]) and temporal models (MATIS [3], LACOSTE [26]). As shown in Table 1, TEDi consistently outperforms prior methods on EV18 across all three metrics. Compared with QPD, TEDi improves Ch_IoU, ISI_IoU, and mc_IoU by 8.92, 6.39, and 13.57 points, respectively. Notably, for challenging classes such as Clip Applier and Large Needle Driver, TEDi achieves substantially higher class IoU, demonstrating the effectiveness of cross-frame memory retrieval and temporal-consistency denoising in mitigating category confusion. Compared with temporal approaches such as LACOSTE and MATIS, TEDi further achieves clear gains in Ch_IoU and ISI_IoU, highlighting the benefit of memory interaction and consistency-guided refinement in the query embedding space. On EV17, TEDi attains 80.71% Ch_IoU and 78.30% ISI_IoU, surpassing existing state-of-the-art methods on the main metrics. Improvements over LACOSTE and MATIS remain consistent across evaluation protocols, confirming the robustness of the proposed framework.

Fig. 2 presents qualitative comparisons. In challenging scenarios with blurred boundaries, partial visibility, or temporal inconsistencies, existing methods often exhibit class flickering or semantic drift, whereas TEDi produces more stable

Table 1. Performance comparison on EV17/18 datasets. M2F refers to the baseline Mask2Former model, while N denotes the number of learnable object queries.

Model	Ch	ISI	mc	Instrument Categories IoU						
	IoU	IoU	IoU	BF	PF	LND	VS/SI	GR/CAM	MCS	UP
Dataset EV17										
TernausNet[24]	35.27	39.87	14.19	44.20	4.67	0.00	0.00	0.00	50.44	0.00
Dual-MF[27]	45.80	-	26.40	34.40	21.50	64.30	24.10	0.80	17.90	21.80
ISINet[10]	55.62	52.20	28.96	38.70	38.50	50.09	27.43	2.01	28.72	12.56
S3Net[5]	72.54	71.99	46.55	75.08	54.32	61.84	35.50	27.47	43.23	28.38
MATIS(Full)[3]	71.36	66.28	41.09	68.37	53.26	53.55	31.89	27.34	21.34	26.53
LACOSTE(S)[26]	76.32	72.37	48.22	73.24	52.04	60.41	38.73	0.00	54.53	67.88
QPD[9]	77.80	79.58	49.92	70.61	45.84	80.01	63.41	33.64	66.57	35.28
M2F($N = 100$)	75.12	71.68	44.48	60.42	62.97	60.88	36.29	3.14	30.22	44.48
M2F($N = 10$)	77.79	73.05	49.63	76.33	60.06	65.06	35.69	3.81	42.75	63.68
TEDi	80.71	78.30	52.11	75.55	58.55	65.78	39.32	15.76	40.64	69.13
Dataset EV18										
TernausNet[24]	35.27	12.67	10.17	13.45	12.39	20.51	5.97	1.08	1.00	16.76
Dual-MF[27]	70.40	-	35.09	74.10	6.80	46.00	30.10	7.60	80.90	0.10
ISINet[10]	73.03	70.97	38.73	73.83	40.35	30.98	37.68	0.00	88.16	0.16
S3Net[5]	75.81	74.02	42.58	77.22	50.87	19.83	50.59	0.00	92.12	7.44
MATIS(Full)[3]	84.26	79.12	54.04	83.52	41.90	66.18	70.57	0.00	92.96	23.12
LACOSTE(S)[26]	85.20	82.41	55.92	85.21	70.75	68.02	62.64	12.81	91.98	0.00
QPD[9]	77.77	78.43	43.84	82.80	60.94	19.96	49.70	0.00	93.93	0.00
M2F($N = 100$)	82.93	80.54	49.77	85.23	69.42	45.37	56.35	0.00	91.99	0.00
M2F($N = 10$)	84.18	81.26	48.91	86.90	43.47	60.46	54.00	4.27	93.29	0.00
TEDi	86.69	84.82	57.41	87.45	56.22	69.91	69.96	14.82	93.94	9.59

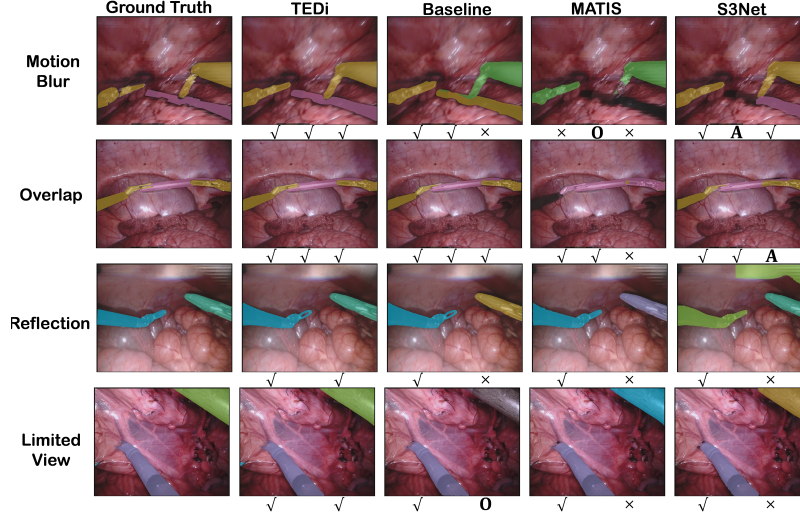


Fig. 2. Qualitative comparison of TEDi with other SOTA methods. ✓ indicates that the instrument is classified and segmented correctly. O indicates missed instance, × shows the misclassified instance and A indicates ambiguous instance.

predictions and improved category discrimination for confusable instruments. To evaluate computational efficiency, we conduct experiments using inputs at a resolution of 512×512 . TEDi contains 60.58M parameters and requires 78.92 GFLOPs, compared with 56.91 GFLOPs for the Mask2Former baseline. TEDi achieves an inference latency of 57.96 ms per clip, with the memory bank introducing only approximately 5.5 ms of additional overhead and a peak GPU memory footprint of 666.1 MB.

3.3 Query Visualization Analysis

We further analyze the baseline’s query predictions during inference and identify two major sources of category errors: (i) high-confidence but incorrectly classified queries that introduce noisy semantics, and (ii) correctly classified yet low-confidence queries whose scores are suppressed due to motion blur or occlusion. The latter are prone to being overlooked during decoding, reducing category separability.

To validate the effectiveness of the proposed modules, we visualize the accuracy of high-confidence queries on EV18 seq_2, which are selected as those with class confidence score $s_{Cls} > 0.8$ and mean class confidence score $s_{Mask} > 0.9$. As shown in Fig. 3, blue denotes correctly classified queries and yellow denotes mismatches. After incorporating MSE and TCD, the number of high-confidence correct queries increases substantially, while high-confidence incorrect queries are markedly reduced. These results suggest that MSE compensates for am-

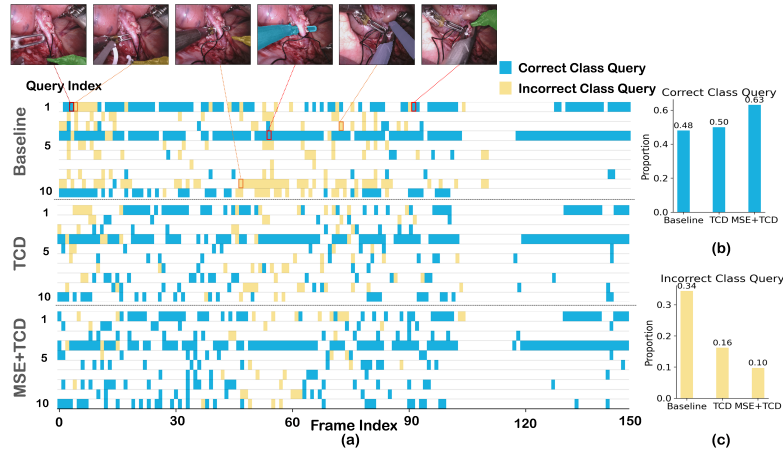


Fig. 3. The visualization analysis of query predictions. Panel (a) visualizes whether high-confidence query category predictions are correctly classified on EV18 seq_2 (frames 000–149). Panel (b) reports the statistics of queries with correct category predictions. Panel (c) reports the statistics of queries with incorrect category predictions.

biguous frames via cross-frame semantic retrieval, and TCD suppresses noisy representations through temporal-consistency regularization.

3.4 Ablation Studies and Hyper-parameter Analysis

We conduct ablation studies on EV18 to evaluate the contribution of each component and analyze the impact of the memory length K and denoising window size T . The results are reported in Table 2.

Table 2. Ablation analysis on EV18 dataset. For column 2-3, K denotes the length of the memory bank, while T denotes the denoising window size.

Model	Params.		Ch_			Instrument Categories IoU						
	K	T	IoU	ISI	mc	BF	PF	LND	SI	CA	MCS	UP
M2F(N=10)	-	-	84.18	81.26	48.91	86.90	43.47	60.46	54.00	4.27	93.29	0.00
MSE	3	-	85.65	83.26	51.67	88.40	54.77	65.87	52.83	0.00	94.11	5.72
TCD	-	3	85.59	83.07	51.70	88.16	49.51	62.86	57.94	0.00	93.90	9.51
MSE+TCD	1	5	85.65	83.69	52.56	88.00	61.19	69.79	47.90	0.00	94.07	6.97
MSE+TCD	2	4	85.79	83.79	52.91	88.43	56.57	66.57	55.18	0.00	94.32	9.30
MSE+TCD	4	2	86.47	84.28	55.23	88.80	58.47	70.94	54.18	2.32	94.16	17.72
MSE+TCD	5	1	85.44	83.50	51.60	88.47	56.41	67.38	50.39	0.00	94.25	4.30
MSE+TCD	3	3	86.69	84.82	57.41	87.45	56.22	69.91	69.96	14.82	93.94	9.59

Overall, introducing the query-level memory bank with MSE consistently improves category discrimination, while the TCR-guided denoising decoder further

enhances temporal semantic consistency and overall segmentation performance. In terms of hyper-parameters, the best performance is achieved with $K = 3$ and $T = 3$. Overly long memory bank or denoising window tends to introduce more historical noise and causes error accumulation over time, whereas an overly short denoising window is insufficient to fully exploit temporal-consistency cues.

4 Conclusions

We propose TEDi, a temporally enhanced framework for query-based surgical instrument segmentation that integrates MSE and TCD to leverage cross-frame priors and enforce temporal coherence. Experiments on two benchmarks show consistent state-of-the-art performance, improving category prediction stability while remaining efficient and validating the benefits of query-level memory interaction and temporal correction for surgical scene understanding.

Acknowledgments. Jiahong Yuan and Tao Zhang received support from the 7th People’s Hospital of Zhengzhou. Weiming Mi and Haoyin Zhou received support from NIH grants R00EB027177 and R01EB036996.

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. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: Matis: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
4. Ayobi, N., Rodríguez, S., Pérez, A., Hernández, I., Aparicio, N., Dessevres, E., Peña, S., Santander, J., Caicedo, J.I., Fernández, N., et al.: Pixel-wise recognition for holistic surgical scene understanding. Medical Image Analysis p. 103726 (2025)
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. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)

7. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European conference on computer vision. pp. 640–658. Springer (2022)
8. Dagnino, G., Kundrat, D.: Robot-assistive minimally invasive surgery: trends and future directions. *International Journal of Intelligent Robotics and Applications* **8**(4), 812–826 (2024)
9. Dhanakshirur, R.R., Shastry, K.A., Borgavi, K., Suri, A., Kalra, P.K., Arora, C.: Learnable query initialization for surgical instrument instance segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2023)
10. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 595–605. Springer (2020)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
12. Huang, Y., Bai, L., Cui, B., Yuan, K., Wang, G., Hoque, M.I., Padoy, N., Navab, N., Ren, H.: Surgtpgs: Semantic 3d surgical scene understanding with text promptable gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 584–594. Springer (2025)
13. de Jong, R., Carolus, H., Franciscus, H., van Jaarsveld, R.C., van Hillegersberg, R., Josien, P., de With, P.H., al Khalil, Y., van Der Sommen, F., et al.: Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis* p. 103873 (2025)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
15. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931* (2024)
16. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
17. Lu, B., Li, B., Chen, W., Jin, Y., Zhao, Z., Dou, Q., Heng, P.A., Liu, Y.: Toward image-guided automated suture grasping under complex environments: A learning-enabled and optimization-based holistic framework. *IEEE Transactions on Automation Science and Engineering* **19**(4), 3794–3808 (2021)
18. Nagy, T.D., Haidegger, T.: A dvrk-based framework for surgical subtask automation. *Acta Polytechnica Hungarica* pp. 61–78 (2019)
19. Westebring-van der Putten, E.P., Goossens, R.H., Jakimowicz, J.J., Dankelman, J.: Haptics in minimally invasive surgery—a review. *Minimally Invasive Therapy & Allied Technologies* **17**(1), 3–16 (2008)
20. Rai, U., Xu, H., Giannarou, S.: Surgpose: Generalisable surgical instrument pose estimation using zero-shot learning and stereo vision. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6875–6881. IEEE (2025)
21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
22. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International workshop on machine learning in medical imaging. pp. 379–387. Springer (2017)

23. Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N.: A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. *IEEE Robotics and Automation Letters* **6**(2), 2938–2945 (2021)
24. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). pp. 624–628. IEEE (2018)
25. Toussaint, M., Ha, J.S., Oguz, O.S.: Co-optimizing robot, environment, and tool design via joint manipulation planning. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 6600–6606. IEEE (2021)
26. 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)
27. Zhao, Z., Jin, Y., Gao, X., Dou, Q., Heng, P.A.: Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2020)
28. Zhao, Z., Jin, Y., Heng, P.A.: Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: 2022 International conference on robotics and automation (ICRA). pp. 11186–11193. IEEE (2022)
29. Zhou, J., Pang, Z., Wang, Y.X.: Rmem: Restricted memory banks improve video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18602–18611 (2024)