

Communication

Surgical Instrument Segmentation via Segment-then-Classify Framework with Instance-Level Spatiotemporal Consistency Modeling

Tiyao Zhang, Xue Yuan * and Hongze Xu

School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China;
22115028@bjtu.edu.cn (T.Z.); hzxu@bjtu.edu.cn (H.X.)

* Correspondence: xyuan@bjtu.edu.cn

Abstract

Accurate segmentation of surgical instruments in endoscopic videos is crucial for robot-assisted surgery and intraoperative analysis. This paper presents a Segment-then-Classify framework that decouples mask generation from semantic classification to enhance spatial completeness and temporal stability. First, a Mask2Former-based segmentation backbone generates class-agnostic instance masks and region features. Then, a bounding box-guided instance-level spatiotemporal modeling module fuses geometric priors and temporal consistency through a lightweight transformer encoder. This design improves interpretability and robustness under occlusion and motion blur. Experiments on the EndoVis 2017 and 2018 datasets demonstrate that our framework achieves mIoU improvements of 3.06%, 2.99%, and 1.67% and mIoU gains of 2.36%, 2.85%, and 6.06%, respectively, over previously state-of-the-art methods, while maintaining computational efficiency.

Keywords: surgical instrument segmentation; segment-then-classify framework; instance-level spatiotemporal consistency modeling

Academic Editor: Toon Goedemé

Received: 9 September 2025

Revised: 11 October 2025

Accepted: 13 October 2025

Published: 15 October 2025

Citation: Zhang, T.; Yuan, X.; Xu, H. Surgical Instrument Segmentation via Segment-then-Classify Framework with Instance-Level Spatiotemporal Consistency Modeling. *J. Imaging* **2025**, *11*, 364. <https://doi.org/10.3390/jimaging11100364>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Minimally invasive surgery (MIS) has become a valuable alternative to conventional open procedures, such as appendectomy [1], cholecystectomy [2], and pancreas or liver resection [3], owing to its smaller incisions, reduced trauma, and faster recovery. However, MIS also poses unique challenges to surgeons, including a restricted field of view and complex hand–eye coordination, which collectively increase cognitive workload and demand high operational precision. To alleviate these challenges, accurate and temporally consistent surgical instrument segmentation is essential. Reliable segmentation facilitates real-time surgical navigation and robotic control, while also supporting postoperative analysis, skill assessment, and workflow optimization. Enhancing the robustness and interpretability of segmentation therefore directly contributes to improving the safety and automation level of computer-assisted MIS. Nevertheless, the surgical environment is highly dynamic—characterized by specular reflections, occlusions, motion blur, and cluttered anatomical structures—which makes robust and temporally stable segmentation particularly difficult to achieve. Traditional pixel-wise classification approaches often fail under these conditions, producing fragmented or temporally inconsistent masks that limit their usefulness in downstream tasks requiring temporal coherence and spatial completeness.

To address these challenges, Robot-Assisted Minimally Invasive Surgery (RAMIS) has been developed with great vigor in recent years [4,5], aiming to assist surgeons in overcoming these shortcomings more effectively, including the utilization of automatic surgical skill analysis [6], surgical stage segmentation [7], surgical scene reconstruction [8], field of view expansion [9,10], and other techniques. To accomplish these sophisticated operations, it is essential to accurately identify the location of surgical instruments within the image domain through the utilization of image segmentation methods.

In the early stages of development, techniques employed utilized handcrafted features derived from color and texture, in conjunction with machine learning models such as random forests and Gaussian mixtures [11,12]. Subsequently, convolutional neural network methods facilitated further advancement in the field of surgical instrument segmentation. ToolNet [13] employed a fully nested fully convolutional network to impose multi-scale prediction constraints. In a related vein, Laina et al. [14] put forth a multi-task convolutional neural network for parallel regression segmentation and localization. Milletari et al. [15] employed a residual convolutional neural network to integrate multi-scale features extracted from a frame through a long short-term memory (LSTM) unit.

Although deep learning-based approaches have achieved unprecedented success, most existing work treats surgical video data as independent static frames, relying solely on visual cues for segmentation. However, surgical videos contain rich temporal dynamics that can provide critical clues for improving accuracy and stability. Effectively exploiting these temporal cues and explicitly injecting them into the network has therefore become a key challenge. Recent transformer-based methods, such as MATIS [16], leverage Multiscale Vision Transformers [17] to extract global temporal consistency information from video sequences for mask classification.

In this study, we propose a segment-then-classify framework for surgical instrument segmentation in endoscopic videos. The proposed framework decouples mask generation and category classification, producing spatially coherent masks while improving classification stability. Specifically, we first generate class-agnostic instance masks using a Mask2Former-based segmentation backbone. Then, to enhance classification accuracy and temporal consistency, we introduce a bounding box-guided instance-level spatiotemporal consistency modeling module, which combines region proposal features with normalized bounding box priors. A lightweight Transformer encoder is employed to capture the temporal evolution of each instrument instance, ensuring consistent classification across consecutive frames.

The main contributions of this paper can be summarized as follows:

- (1) We propose a segment-then-classify framework that decouples segmentation and classification, improving the spatial completeness and temporal stability of surgical instrument segmentation.
- (2) We introduce a bounding box-guided temporal modeling strategy that combines spatial priors with semantic region features to enhance instance-level classification consistency.
- (3) We achieve state-of-the-art performance on the EndoVis 2017 and 2018 datasets, demonstrating the effectiveness and interpretability of the proposed framework.

The remainder of this paper is organized as follows. Section 2 reviews related work on surgical instrument segmentation and temporal modeling approaches. Section 3 presents the details of the proposed framework. Section 4 reports the experimental results and analysis. Sections 5 and 6 discuss the findings and open research questions, followed by conclusions in Section 7.

2. Related Work

2.1. Surgical Instrument Segmentation

Surgical instrument segmentation has been extensively studied as a core component of computer-assisted minimally invasive surgery. Early works primarily focused on pixel-wise classification using convolutional neural networks (CNNs), such as TerausNet [18] and U-Net-based models [19], which achieved remarkable success in static image segmentation. However, these methods often produce fragmented or incomplete masks when applied to video sequences, due to the high variability of surgical scenes and rapid instrument motion.

To improve mask quality, several methods adopted instance segmentation strategies that treat surgical instruments as independent objects rather than homogeneous pixel classes. For example, ISINet [20] introduced instance-level learning for object differentiation, while TerausNet [18] leveraged transformer-based attention mechanisms to integrate multi-scale features. MATIS [16], a recent representative work, proposed a masked-attention transformer to perform segmentation followed by classification, offering improved spatial coherence. Nevertheless, these approaches still exhibit limitations in temporal consistency—the classification of the same instrument may vary across frames due to motion blur, occlusion, or tool overlap.

In addition to the above methods, several recent works have explored transformer-based and foundation-model-inspired architectures for surgical scene understanding. For instance, AGMF-Net [21] introduced multi-scale temporal attention for video object segmentation; and Samsurg [22] leveraged the Segment Anything model for zero-shot instrument detection. These approaches highlight the trend toward integrating strong priors and global attention, which motivates our instance-level temporal consistency design.

The above methods highlight the need to go beyond frame-level segmentation and integrate temporal dynamics in an interpretable and instance-aware manner. This motivates our work, which aims to enhance instance-level temporal consistency by combining explicit geometric priors with transformer-based temporal modeling.

2.2. Temporal Modeling and Segment-then-Classify Frameworks

Temporal modeling has emerged as an effective approach to improve consistency and robustness in video-based segmentation tasks. Traditional methods typically adopt optical flow estimation or recurrent architectures (e.g., ConvLSTM) to capture temporal dependencies between frames. However, these techniques are often computationally expensive and sensitive to motion noise, making them less suitable for real-time surgical applications.

Recently, transformer-based architectures have demonstrated strong potential for learning long-range temporal relationships. For instance, MF-TAPNet [23] and TraSeTR [24] utilize multi-frame attention to propagate contextual information across frames. While these approaches successfully model temporal information at the feature level, they often lack instance-level interpretability, making it difficult to ensure consistent classification of each instrument across frames.

In contrast, the segment-then-classify paradigm offers a promising alternative by decoupling spatial mask prediction from semantic classification. Instead of performing pixel-wise classification directly, the model first generates class-agnostic instance masks and subsequently classifies them using instance-level features. This decoupling facilitates better spatial completeness and provides flexibility for integrating temporal and positional priors. Our method builds upon this principle and extends it by introducing bounding box-guided instance-level spatiotemporal consistency modeling, which explicitly captures the geometric and temporal evolution of each instrument instance using a lightweight transformer encoder. This design enhances interpretability, reduces ambiguity in instance tracking, and improves classification stability under complex motion.

3. Materials and Methods

3.1. Overview

The proposed method aims to achieve spatially complete and temporally consistent segmentation of surgical instruments by decoupling mask generation and classification.

As illustrated in Figure 1, the overall framework follows a segment-then-classify paradigm and consists of three main stages:

- (1) a class-agnostic segmentation backbone based on Mask2Former to generate instance masks and corresponding region proposal features;
- (2) a bounding box-guided instance prior construction module to combine spatial priors and semantic region features;
- (3) an instance-level spatiotemporal consistency modeling module that employs a transformer encoder to capture temporal relationships among instances across consecutive frames.

The final classification head assigns semantic categories to each predicted mask.

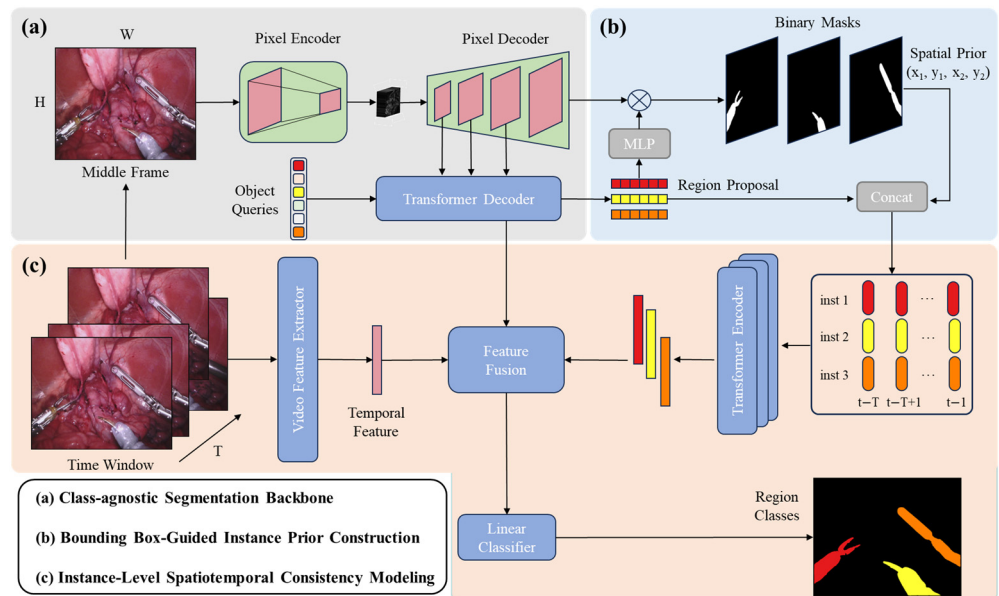


Figure 1. Overview of the proposed Segment-then-Classify framework for surgical instrument segmentation. The framework consists of three main stages: (a) a Mask2Former-based segmentation backbone to generate class-agnostic instance masks and region features; (b) a bounding box-guided instance prior construction module that fuses spatial priors with semantic features; and (c) an instance-level spatiotemporal consistency modeling module employing a transformer encoder to ensure consistent classification across consecutive frames.

3.2. Class-Agnostic Segmentation Backbone

We adopt Mask2Former as the segmentation backbone due to its ability to handle multiple instances with high spatial precision.

Given an input video frame $I_t \in R^{H \times W \times 3}$, Mask2Former extracts a feature map $F_t \in R^{C \times H' \times W'}$ through a pixel decoder and transformer-based mask decoder. A set of N instance queries is used to generate mask embeddings $\{M_t^i\}_{i=1}^N$ and corresponding region proposal features $\{f_t^i\}_{i=1}^N$.

The segmentation objective can be formulated as Equation (1):

$$\mathcal{L}_{seg} = \frac{1}{N} \sum_{i=1}^N [\mathcal{L}_{dice}(m_t^i, \hat{m}_t^i) + \mathcal{L}_{bce}(m_t^i, \hat{m}_t^i)], \quad (1)$$

where $\hat{m}_t^i$ is the predicted mask, and $\mathcal{L}_{dice}$ and $\mathcal{L}_{bce}$ denote the Dice loss and binary cross-entropy loss, respectively.

The output of this stage provides instance-level mask features f_t^i , which contain semantic and appearance information but lack explicit spatial priors or temporal awareness.

3.3. Bounding Box-Guided Instance Prior Construction

For each predicted mask $\hat{m}_t^i$, we compute its bounding box coordinates as Equation (2):

$$b_t^i = (x_{min}^i, y_{min}^i, x_{max}^i, y_{max}^i), \quad (2)$$

and normalize them by the frame dimensions, as shown in Equation (3):

$$\tilde{b}_t^i = \left(\frac{x_{min}^i}{W}, \frac{y_{min}^i}{H}, \frac{x_{max}^i}{W}, \frac{y_{max}^i}{H} \right), \quad (3)$$

Each normalized bounding box encodes geometric and positional priors of the corresponding instrument instance.

To construct an instance-level feature that integrates both semantic and spatial cues, we project $\tilde{b}_t^i$ into a high-dimensional space using a fully connected layer, as shown in Equation (4):

$$p_t^i = FC(\tilde{b}_t^i), \quad (4)$$

and concatenate it with the semantic region feature, as shown in Equation (5):

$$b_t^i = \text{Concat}(f_t^i, p_t^i). \quad (5)$$

This results in a position-aware instance representation $z_t^i \in R^d$, which provides a compact yet interpretable summary of each instrument spatial and appearance characteristic.

3.4. Instance-Level Spatiotemporal Consistency Modeling

To capture temporal evolution and maintain category consistency across frames, we design a lightweight transformer encoder that processes instance features over time.

Given a sequence of T consecutive frames, the input feature set is $Z = \{z_t^i \mid t = 1, \dots, T; i = 1, \dots, N_t\}$. We first apply temporal position encoding $PE(t)$ to preserve frame order, then feed the features into a multi-head self-attention encoder, as shown in Equations (6) and (7):

$$h_t^i = MHSA(z_t^i + PE(t)), \quad (6)$$

$$\tilde{h}_t^i = FFN(h_t^i), \quad (7)$$

where $MHSA$ denotes the multi-head self-attention operation and FFN is a feed-forward layer.

This process enables contextual aggregation across both temporal and instance dimensions, ensuring consistent representation for the same instrument in different frames.

Finally, the refined instance feature $\tilde{h}_t^i$ is passed through a classification head to predict the instrument category, as shown in Equation (8):

$$\hat{y}_t^i = \text{Softmax}(W_c \tilde{h}_t^i + b_c), \quad (8)$$

and the classification loss is defined as Equation (9):

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N y_t^i \log(\hat{y}_t^i). \quad (9)$$

The overall loss function for joint optimization is given by Equation (10):

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{cls}, \quad (10)$$

where λ balances segmentation and classification objectives.

3.5. Implementation Details

All experiments were conducted on an NVIDIA RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA) with PyTorch 2.1. The model was trained with an AdamW optimizer using an initial learning rate of 1×10^{-4} , a weight decay of 1×10^{-5} , and a batch size of 4. Each video sequence was sampled with 5 consecutive frames ($T = 5$) for temporal modeling. The bounding box projection dimension was set to 256, matching the region feature dimension. We trained for 80 epochs, with early stopping based on validation mIoU.

We evaluate our method on the EndoVis 2017 Robotic Instrument Segmentation and EndoVis 2018 Robotic Scene Segmentation datasets. Following standard practice, we report the metrics mIoU, IoU, and mcIoU for performance comparison.

The proposed model requires 65M parameters and 98 GFLOPs per frame sequence, comparable to MATIS (96 GFLOPs) and TraSeTr (102 GFLOPs), while maintaining faster inference (12.5 fps on RTX 4090).

4. Results

4.1. Datasets

We train and evaluate our method on two publicly available experimental frameworks, the Endovis 2017 and Endovis 2018 [25] datasets. Each dataset consists of 10 video sequences of porcine abdominal surgery. Each video contains 300 frames with a sampling frequency of 2 Hz and a resolution of 1280×1024 . 8×225 frames are used as the training set, and the remaining 8×75 frames and 2×300 frames are used as the test set. For a fair comparison with previous methods, we follow the evaluation criteria established in [20]. We only use the 4-fold cross-validation proposed in For Endovis 2018, we use the additional instance annotations given in and their predefined training and validation splits. For evaluation, we adopt three common segmentation metrics from [20]: mean intersection over union (mIoU), intersection over union (IoU), and mean class intersection over union (mcIoU). Finally, we present the standard deviation error between Endovis 2017 folds.

4.2. Main Results

Table 1 shows the comparison results on Endovis 2017 and Endovis 2018. Our method outperforms all previous methods in all three overall segmentation indicators, and improves the three overall indicators of mIoU, IoU, and mcIoU by 3.06%, 2.99%, and 1.67% and 2.36%, 2.85%, and 6.06%, respectively.

Table 1. Performance of the proposed method and existing approaches on the EndoVis 2017 and 2018 datasets. The best results are shown in bold. Our method achieves higher mIoU, IoU, and mcIoU scores, indicating improvements in both spatial completeness and temporal consistency.

Dataset	Method	mIoU	IoU	mcIoU	Instrument Category						
					BF	PF	LND	VS/SI	GR/CA	MCS	UP
Endovis 2017	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	0.8	17.90	21.80
	ISINet	55.62	52.20	28.96	38.70	38.50	50.09	27.43	2.1	28.72	12.56
	TraSeTr	60.40	-	32.56	45.20	56.70	55.8	38.90	11.40	31.30	18.20
	Kurmann et al.	65.70	-	-	-	-	-	-	-	-	-
	MATIS	66.73	61.79	34.99	67.17	50.36	46.53	31.49	11.08	13.57	24.79
	Ours	69.79	64.78	36.66	61.03	52.28	45.69	34.66	15.00	20.81	27.14
Endovis 2018	TernausNet	46.22	39.87	14.19	44.20	4.67	0.00	0.00	0.00	50.44	0.00
	MF-TAPNet	67.87	39.14	24.68	69.23	6.10	11.68	14.00	0.91	70.24	0.57

ISINet	73.03	70.94	40.21	73.83	48.61	30.98	37.68	0.00	88.16	2.16
TraSeTr	76.20	-	47.71	76.30	53.30	46.50	40.60	13.90	86.20	17.15
MATIS	82.31	77.01	48.57	83.55	38.65	40.48	92.56	70.38	0	14.39
Ours	84.67	79.86	54.63	83.95	41.47	66.57	92.74	74.20	0	23.50

4.3. Qualitative Results

We also visualize the segmentation results. Figure 2 shows a qualitative comparison between our method and previous pixel classification-based surgical instrument segmentation methods on the Endovis 2017 and Endovis 2018 datasets. The previous pixel-by-pixel classification strategy lacks spatial consistency, resulting in incomplete and coherent instrument masks (e.g., columns 2 and 3), or even failure to segment instruments at all (e.g., column 4). Our mask classification-based method can better utilize instance-level properties to segment complete surgical instruments.

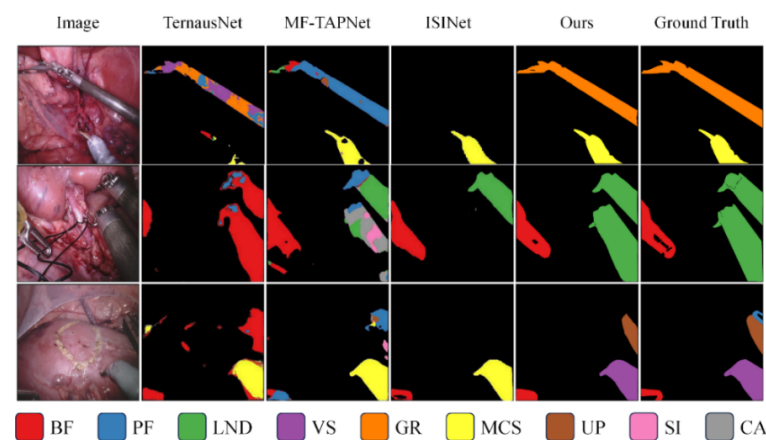


Figure 2. Qualitative comparison of segmentation results on the EndoVis 2017 and EndoVis 2018 datasets. Visualization of the proposed segment-then-classify framework versus pixel classification-based methods. Each color denotes a distinct surgical instrument category, and the color-instrument correspondence is shown in the legend below. Our method produces more complete and temporally coherent masks, particularly in regions with motion blur or overlapping tools.

Figure 3 specifically shows the qualitative comparison of the results before and after the introduction of instance-level spatiotemporal consistency information. When the binary masks are the same, this paper further enhances the spatiotemporal consistency of instances between consecutive frames by introducing the instance-level spatiotemporal consistency information of the target instrument, thereby improving the classification results.

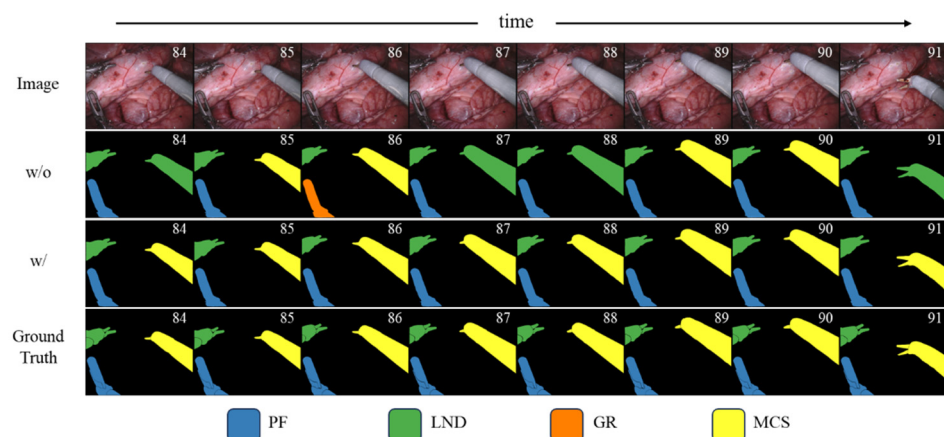


Figure 3. Effect of instance-level spatiotemporal consistency modeling on classification stability. Qualitative comparison of segmentation results before and after introducing instance-level

spatiotemporal consistency information. Without this module, the same instrument exhibits inconsistent category predictions across consecutive frames, whereas our method maintains stable and coherent classification over time.

4.4. Ablation Experiment

We show the impact of position prior information and region proposal features on model performance when used as instance-level summary features in Table 2. Both properties improve the overall performance of the model. Among them, the position prior information has a more obvious effect on improving model performance, which also shows that interpretable position prior information plays a vital role in extracting instance motion model information.

Table 2. Ablation of Location and Region Proposal Features for Instance-Level Summarization in EndoVis 2017. The best results are shown in bold. “✓” indicates the presence of the attribute and “-” indicates its absence.

Position Prior	Region Proposal	mIoU	IoU	mcIoU
-	-	66.73	61.79	34.99
✓	-	68.91 (+2.18)	63.86 (+2.07)	35.31 (+0.32)
-	✓	67.02 (+0.29)	62.10 (+0.31)	35.28 (+0.29)
✓	✓	69.79 (+3.06)	64.78 (+2.99)	36.66 (+1.67)

5. Discussion

The experimental results on the EndoVis 2017 and 2018 datasets demonstrate the effectiveness of the proposed Segment-then-Classify framework in improving both spatial completeness and temporal stability.

Compared with pixel-wise classification methods, our approach ensures more coherent mask boundaries by decoupling segmentation from classification.

Furthermore, the incorporation of bounding box-guided priors provides explicit spatial cues that enhance interpretability and reduce temporal ambiguity, especially in cases of occlusion or fast motion.

While the method shows consistent improvements, several limitations remain. First, the current spatiotemporal modeling focuses on short-term consistency within a limited frame window, which may not fully capture long-term surgical workflow dynamics. Second, the framework assumes that instance masks are of sufficient quality; inaccurate mask generation can propagate errors into the classification stage. Third, although our model improves interpretability, it still requires considerable computational resources for transformer-based temporal encoding.

The current evaluation is limited to the EndoVis datasets, which, although standard, may not cover the full variability of clinical environments. Extending validation to other datasets such as CholecSeg8k or SurgVisDom is planned for future work to further assess generalization.

Practical Managerial Significance

The proposed framework has potential clinical benefits in several real-world scenarios. Temporally stable and interpretable instrument segmentation can improve intraoperative awareness for robot-assisted surgery, enabling safer tool navigation and collision avoidance. It also facilitates postoperative analysis and skill assessment by providing consistent instrument tracking data. Furthermore, stable segmentation outputs can serve as reliable inputs for downstream tasks such as surgical phase recognition and workflow modeling, contributing to smarter, data-driven surgical assistance systems.

6. Open Research Questions and Future Directions

Despite the progress achieved, several open research questions (ORQs) remain.

- (1) Long-term temporal modeling: How can transformer-based structures efficiently capture dependencies over entire surgical procedures without excessive computation?
- (2) Generalization to unseen tools: Can the model adapt to novel instruments or surgical scenes without retraining, possibly through domain adaptation or few-shot learning?
- (3) Multi-modal integration: How can visual segmentation be combined with kinematic or force-sensing data to improve scene understanding?
- (4) Real-time deployment: What architectural or hardware-level optimizations are required to deploy the framework in real robotic surgery environments?

Addressing these questions could lead to more robust, efficient, and interpretable models for surgical video understanding.

7. Conclusions

In this paper, we proposed a segment-then-classify framework for surgical instrument instance segmentation in endoscopic videos. Unlike traditional pixel-wise classification methods, our approach decouples mask generation and category prediction, leading to more complete and temporally stable segmentation results. To further enhance classification accuracy, we introduced a bounding box-guided temporal modeling strategy that combines geometric priors with semantic region features. By leveraging a Transformer encoder to model instance-level spatiotemporal consistency across frames, our method effectively improves mask classification while maintaining high-quality mask structures. Extensive experiments on the EndoVis 2017 and 2018 datasets demonstrate that our framework outperforms state-of-the-art methods in both segmentation completeness and classification robustness, especially under challenging temporal conditions. In future work, we plan to extend the proposed framework toward long-term temporal modeling, multi-modal feature integration, and real-time clinical deployment in robot-assisted surgery.

Author Contributions: Methodology, T.Z.; Validation, T.Z.; Formal analysis, H.X.; Investigation, T.Z.; Resources, H.X.; Writing—original draft, T.Z.; Writing—review & editing, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the data used were obtained from the public databases..

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from MICCAI Endoscopic Vision Challenge and are available at <https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/> (accessed on 9 September 2025) and <https://endovissub2018-roboticscenesegmentation.grand-challenge.org/> (accessed on 9 September 2025) with the permission of MICCAI Endoscopic Vision Challenge.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Biondi, A.; Di Stefano, C.; Ferrara, F.; Bellia, A.; Vacante, M.; Piazza, L. Laparoscopic versus open appendectomy: A retrospective cohort study assessing outcomes and cost-effectiveness. *World J. Emerg. Surg.* **2016**, *11*, 44.
2. Antoniou, S.A.; Antoniou, G.A.; Koch, O.O.; Pointner, R.; Granderath, F.A. Meta-analysis of laparoscopic vs open cholecystectomy in elderly patients. *World J. Gastroenterol. WJG* **2014**, *20*, 17626.
3. Chen, Q.; Merath, K.; Bagante, F.; Akgul, O.; Dillhoff, M.; Cloyd, J.; Pawlik, T.M. A comparison of open and minimally invasive surgery for hepatic and pancreatic resections among the medicare population. *J. Gastrointest. Surg.* **2018**, *22*, 2088–2096.
4. Haidegger, T.; Speidel, S.; Stoyanov, D.; Satava, R.M. Robot-assisted minimally invasive surgery—Surgical robotics in the data age. *Proc. IEEE* **2022**, *110*, 835–846.
5. Maier-Hein, L.; Eisenmann, M.; Sarikaya, D.; März, K.; Collins, T.; Malpani, A.; Fallert, J.; Feussner, H.; Giannarou, S.; Mascagni, P.; et al. Surgical data science—from concepts toward clinical translation. *Med. Image Anal.* **2022**, *76*, 102306.
6. Zia, A.; Essa, I. Automated surgical skill assessment in RMIS training. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 731–739.
7. Twinanda, A.P.; Shehata, S.; Mutter, D.; Marescaux, J.; de Mathelin, M.; Padoy, N. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **2016**, *36*, 86–97.
8. Long, Y.; Li, Z.; Yee, C.H.; Ng, C.F.; Taylor, R.H.; Unberath, M.; Dou, Q. E-dssr: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Proceedings, Part IV 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 415–425.
9. Casella, A.; Moccia, S.; Paladini, D.; Frontoni, E.; De Momi, E.; Mattos, L.S. A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation. *Med. Image Anal.* **2021**, *70*, 102008.
10. Bano, S.; Vasconcelos, F.; Tella-Amo, M.; Dwyer, G.; Gruijthuijsen, C.; Poorten, E.V.; Vercauteren, T.; Ourselin, S.; Deprest, J.; Stoyanov, D. Deep learning-based fetoscopic mosaicking for field-of-view expansion. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1807–1816.
11. Bouget, D.; Benenson, R.; Omran, M.; Riffaud, L.; Schiele, B.; Jannin, P. Detecting surgical tools by modelling local appearance and global shape. *IEEE Trans. Med. Imaging* **2015**, *34*, 2603–2617.
12. Rieke, N.; Tan, D.J.; di San Filippo, C.A.; Tombari, F.; Alshekhali, M.; Belagiannis, V.; Eslami, A.; Navab, N. Real-time localization of articulated surgical instruments in retinal microsurgery. *Med. Image Anal.* **2016**, *34*, 82–100.
13. Garcia-Peraza-Herrera, L.C.; Li, W.; Fidon, L.; Gruijthuijsen, C.; Devreker, A.; Attilakos, G.; Deprest, J.; Poorten, E.V.; Stoyanov, D.; Vercauteren, T.; et al. Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; IEEE: New York, NY, USA, 2017; pp. 5717–5722.
14. Laina, I.; Rieke, N.; Rupperecht, C.; Vizcaíno, J.P.; Eslami, A.; Tombari, F.; Navab, N. Concurrent segmentation and localization for tracking of surgical instruments. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017; Proceedings, Part II 20; Springer International Publishing: Cham, Switzerland, 2017; pp. 664–672.
15. Milletari, F.; Rieke, N.; Baust, M.; Esposito, M.; Navab, N. CFCM: Segmentation via coarse to fine context memory. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018; Proceedings, Part IV 11; Springer International Publishing: Cham, Switzerland, 2018; pp. 667–674.
16. Ayobi, N.; Pérez-Rondón, A.; Rodríguez, S.; Arbeláez, P. Matis: Masked-attention transformers for surgical instrument segmentation. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
17. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 6824–6835.
18. Shvets, A.A.; Rakhlin, A.; Kalinin, A.A.; Iglovikov, V.I. “Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning,” In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA, 17–20 December 2018.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*; Springer International Publishing: Cham, Switzerland, 2015.
20. González, C.; Bravo-Sánchez, L.; Arbeláez, P. Isinet: An instance-based approach for surgical instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2020; pp. 595–605.

21. Song, M.; Zhai, C.; Yang, L.; Liu, Y.; Bian, G. An attention-guided multi-scale fusion network for surgical instrument segmentation. *Biomed. Signal Process. Control* **2025**, *102*, 107296.
22. Matasyoh, N.M.; Mathis-Ullrich, F.; Zeineldin, R.A. *Samsurg: Surgical Instrument Segmentation in Robotic Surgeries Using Vision Foundation Model*; IEEE Access: Cambridge, MA, USA, 2024.
23. Jin, Y.; Cheng, K.; Dou, Q.; Heng, P.A. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part V 22; Springer International Publishing: Cham, Switzerland, 2019; pp. 440–448.
24. Zhao, Z.; Jin, Y.; Heng, P.-A. Trasetr: Track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022.
25. 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* **2020**, arXiv:2001.11190.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.