arXiv:2403.09551v3 [cs.CV] 17 Mar 2026

Label-supervised surgical instrument segmentation using temporal equivariance and semantic continuity

Qiyuan Wang[1,2], Yanzhe Liu[5], Shang Zhao[2], Rong liu[5] and S. Kevin Zhou[1,2,3,4]

***Abstract*— In robotic surgery, instrument presence labels are typically recorded alongside video streams, offering a cost-effective alternative to manual annotations for segmentation tasks. Label-supervised surgical instrument segmentation (SIS), a weakly supervised segmentation setting where only instrument presence labels are available, remains underexplored due to its inherently ill-posed nature. Temporal information plays a vital role in capturing sequential dependencies, thereby enhancing representation learning even under incomplete supervision. This paper extends a two-stage label-supervised segmentation framework by leveraging the temporal characteristics of surgical videos from three perspectives. First, a temporal equivariance constraint is introduced to enforce pixel-level consistency across adjacent frames. Second, a class-aware semantic continuity constraint is applied to preserve coherence between global and local regions over time. Third, temporally-enhanced pseudo masks are generated from consecutive frames to suppress irrelevant regions and improve segmentation accuracy. We evaluate our method on two surgical video datasets: the Cholec80 cholecystectomy benchmark and a real-world robotic left lateral segmentectomy (RLLS) dataset. Instance-level instrument annotations, sampled at regular intervals and validated by an experienced clinician, provide a reliable basis for evaluation. Experimental results demonstrate that our method consistently achieves favorable performances over state-of-the-art methods. These findings highlight the effectiveness of incorporating temporal constraints into label-supervised frameworks, offering a promising strategy to reduce annotation costs and advance surgical video analysis.**

I. INTRODUCTION

Automated visual understanding of laparoscopic and robotic surgery videos is essential to enable autonomous procedures and provide advanced intraoperative assistance. At its core, this capability is dependent on accurate surgical instrument segmentation (SIS). In recent years, SIS has evolved rapidly: early conventional methods [1] have been replaced by deep learning-based approaches [2], [3], [4]. In addition, segmentation architectures originally developed for natural images such as DETR [5] and Mask2Former [6] have been adapted to the surgical domain, resulting in further improvements [7], [8], [9], [10], [11]. However, most current methods rely on fully supervised learning, which requires extensive manual annotations to achieve the robustness and precision demanded by clinical applications. To reduce this annotation burden, recent studies have proposed various incompletely supervised learning strategies. In the unsupervised paradigm, Liu et al. [12] relied on handcrafted cues such as color, objectness, and location to learn implicit segmentation representations. Other methods [13], [14] used annotated simulated data and applied domain adaptation to real-world surgical scenarios. These approaches can only identify category-agnostic segments and lack semantic information. In the semi-supervised paradigm, Jin et al. [15] generated pseudo-masks for unlabeled frames by transferring predictions based on temporal priors. Wei et al. [16] extended a widely used semi-supervised classification framework to SIS. Their method applies unsupervised loss to high-confidence pixels in weakly augmented images by contrasting them with adversarially augmented versions. However, these approaches still require a substantial quantity of pixel-level annotations. In the weakly supervised paradigm, Yang et al. [17] reduced the need for dense segmentation annotations by using scribble-level labels. Matalla et al. [18] proposed a joint detection and segmentation framework based on bounding box annotations and partial presence labels. More recent work [19] has advanced label-supervised localization and detection using only instrument presence information. However, label-supervised surgical instrument segmentation

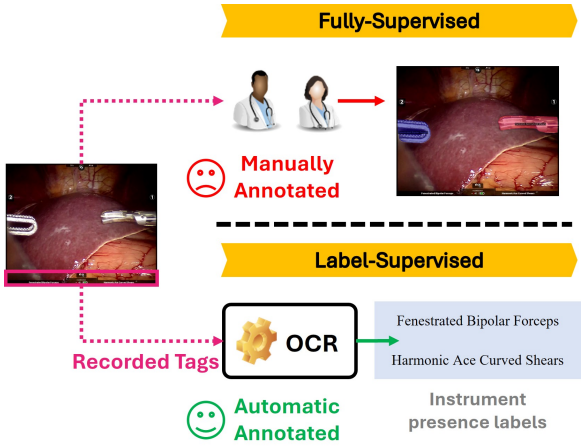


Fig. 1. Motivation of label-supervised surgical instrument segmentation which only relies on instrument presence labels. The recorded presence information can be obtained automatically through Optical Character Recognition (OCR) methods at a low cost. This paradigm can alleviate manual pixel-wise annotations for segmentation methods.

[1]School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei Anhui, 230026, P.R. China. [2]Center for Medical Imaging, Robotics, Analytic Computing & Learning(MIRACLE), Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R. China. [3]Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei Anhui, 230026, P.R. China. [4]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Institute of Computing Technology, CAS, Beijing, 100190, P.R. China. [5]Faculty of Hepato-Biliary-Pancreatic Surgery, The First Medical Center, Chinese PLA General Hospital, Beijing, China.

Corresponding author: S. Kevin Zhou (skevinzhou@ustc.edu.cn); You can apply for annotations through https://forms.gle/AZKQBw4HC5uXgVmg8.

remains an underexplored task. It requires more robust modeling of pixel-level and region-level relationships.

Robotic surgical videos often record both video streams and instrument presence annotations, offering substantial potential to reduce manual annotation costs, as illustrated in Fig. 1. However, label-supervised surgical instrument segmentation remains an under-constrained and challenging task. Temporal properties can help alleviate these issues by modeling sequential dependencies and patterns over time under incomplete supervision. To address these challenges, we propose a novel label-supervised surgical instrument segmentation (LS-SIS) framework that relies solely on instrument presence annotations. Our approach incorporates temporality-based constraints to guide the learning process. First, we introduce a temporal equivalence constraint to enhance pixel-level consistency across consecutive features. Second, a class-aware semantic continuity constraint is applied to improve region-level semantic consistency over time for each instrument category. By focusing on uncertain local regions, the model is encouraged to activate non-discriminative parts of the target. Finally, instead of generating frame-wise pseudo-masks, we construct temporal-enhanced masks from a clip of consecutive frames. This strategy helps suppress irrelevant areas while emphasizing foreground regions corresponding to the annotated categories.

Our main contributions can be summarized as follows: 1) We propose a label-supervised surgical instrument segmentation method, LS-Surg, that relies solely on instrument presence annotations. This approach is well-suited to the surgical domain, where presence annotations are recorded alongside video streams, significantly reducing the cost of manual annotations; 2) We incorporate temporal properties by introducing two temporality-based constraints: temporal equivariance and class-aware temporal semantic continuity constraints, combined with a temporal-enhanced pseudo-mask generation mechanism; 3) Our method demonstrates performance improvements on an open benchmark (Cholec80) and a real robotic left lateral segment liver surgery dataset (RLLS), outperforming prior state-of-the-art methods. Additionally, we contribute instance-wise annotations that were manually labeled and double-checked by an experienced clinician.

II. METHODS

The overall architecture of LS-Surg is depicted in Fig. 2. It is built upon a two-stage label-supervised segmentation framework (LSSS), Multi-Class Token Transformer (MCT) [20], which was designed for static images. To mitigate highly under-constrained challenges, our method takes temporal attributes of surgical video into account with two temporality-based constraints: temporal equivariance constraint (**TER**) and class-aware semantic continuity constraint (**CSCR**). Furthermore, temporal-enhanced pseudo masks generation (**TMG**) replaces frame-wise one to suppress irrelevant information.

A. Multi-Class Token Transformer

MCT uses multiple class tokens instead of a single one, which is typical in traditional transformers. This design enables class-specific localization by allowing transformer attention to occur between each class token and the patch tokens. To incorporate temporal constraints during training, we extend this architecture to receive temporal pair inputs, as illustrated in the purple section of Fig. 2.

Given a temporal pair (I^t, I^T), multi-class tokens p aggregate semantic information from corresponding patch tokens through multiple transformer layers. This process yields output class tokens $(z_c^t, z_c^T) \in \mathbb{R}^D$ and output patch tokens $(z_p^t, z_p^T) \in \mathbb{R}^{D\times h\times w}$. Here, D denotes the embedding size, while h and w represent height and width, respectively; t and T signify the reference and target time points. These output tokens then follow distinct transformation pathways. The class tokens are average-pooled to generate classification predictions $(y_c^t, y_c^T) \in \mathbb{R}^C$, where C is the number of classes. Concurrently, the patch tokens are processed through CNN layers to produce additional class predictions $(y_p^t, y_p^T) \in \mathbb{R}^C$. Finally, the multi-label soft margin loss is computed between these class predictions and the ground-truth annotations $(\hat{y}^t, \hat{y}^T)$, as formally expressed in Eq.(1), where σ represents the sigmoid function.

$$
\begin{aligned}
L_{CLS} &= \psi(y_c^t, \hat{y}^t) + \psi(y_p^t, \hat{y}^t) + \psi(y_c^T, \hat{y}^T) + \psi(y_p^T, \hat{y}^T) \\
\psi(y, \hat{y}) &= -\frac{1}{C}\sum_{i=1}^{C}\{\hat{y}^i \log\sigma(y^i) + (1-\hat{y}^i)\log(1-\sigma(y^i))\}.
\end{aligned} \tag{1}
$$

We also employ contrastive class-token enhancement, a technique adopted from [21]. After the training phase, final class-specific localization maps are generated by combining two key components: class-specific attention maps from the transformer encoder and class activation maps derived from patch tokens. This integration process is detailed in the same aforementioned work. For consistency and clarity, we use the term "CAM" to refer to these composite localization maps throughout this paper.

B. Temporal Equivariance Constraint

The **TER** aims to enhance pixel-wise temporal consistency among patch tokens across video frames, as illustrated in the orange section of Fig. 2. Conceptually, TER works by transforming features from the reference time $t' = t$ to the current target time $t' = T$. This ensures that the features from the reference frame align with those of the target frame. This dense transition operation can be effectively implemented using a Temporal Propagator.

Class-Aware Projection: To explicitly establish connections with semantic properties, this operation aims to cluster patch tokens based on multi-class tokens p, which can be conceptualized as semantic prototypes. To account for background regions, we **incorporate a background token** alongside the original multi-class tokens. Utilizing an optimal-transport clustering algorithm [22], adjacent patch tokens are clustered online based on these semantic prototypes across temporal views. Given a pair of reference output patch

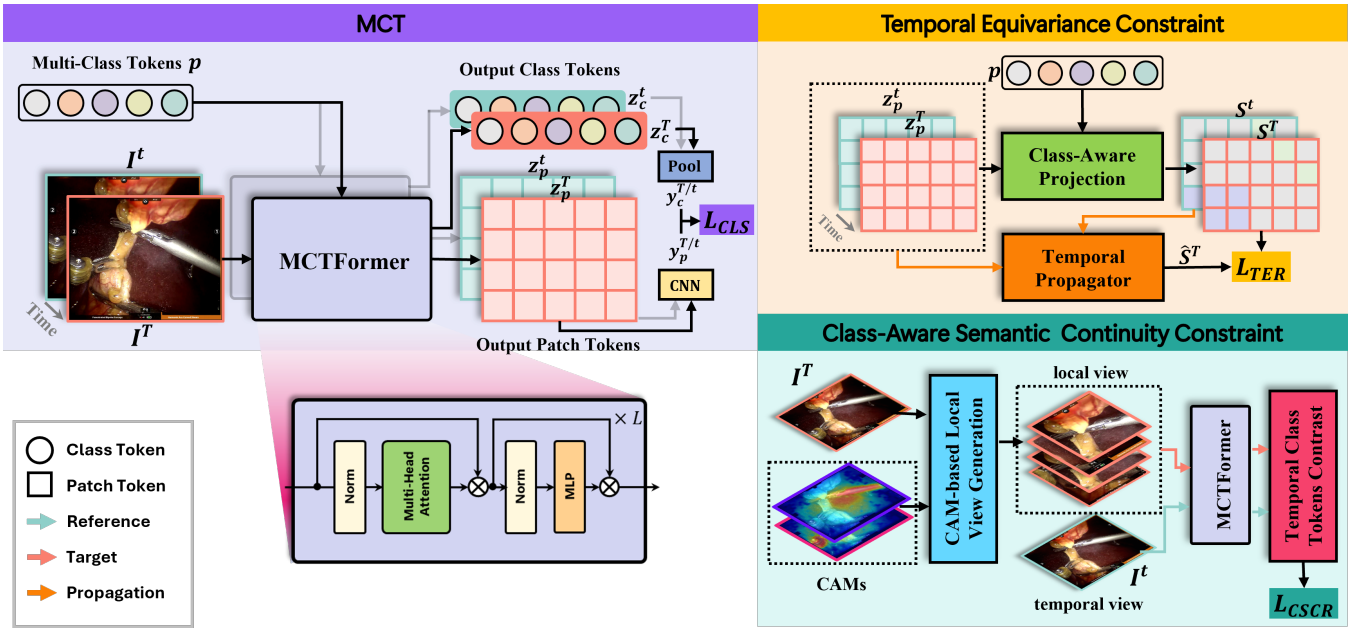


Fig. 2. Schematic illustration of an overview pipeline of our method. Two temporality-based constraints are inserted into a two-stage label-supervised architecture.

tokens z_p^t and target output patch tokens z_p^T, projection results (S^t, S^T) are derived following this online clustering process.

Temporal Propagator: Considering the challenges of appearance variation and fast motion across surgical frames, we adopt a forward warp scheme. This scheme is based on local spatial similarities and avoids the high computational cost in training process, similar to the approach in [23]. This propagator can also be alternatively implemented using optical-flow techniques depending on the specific scenario. Given a pair of output patch tokens (z_p^t, z_p^T), spatial similarities $\mathbf{Q}$ across time are calculated pixel-wise as formulated in Eq.(2), where Θ represents the cosine similarity operator. Here, i and j denote token index items.

$$\mathbf{Q} = [\Theta(z_{p,i}^t, z_{p,j}^T)]_{hw \times hw} \tag{2}$$

The Temporal Propagator transfers S^t to $\hat{S}^T$ based on the calculated spatial similarities $\mathbf{Q}$, as shown in Eq.(3), where Φ represents the propagation function. A local-window thresholding function $\mathcal{N}$ is applied to constrain short-term spatial smoothness, indicating items within a k-sized window around item j. Subsequently, TER can be formulated as Eq.(4) to regulate temporal equivariance. g denotes the argmax function.

$$\begin{aligned} \Phi((z_p^t, z_p^T), S^t) &= \hat{S}^T \\ \hat{S}^T(j) &= \sum_i \frac{\exp(\mathcal{N}(\mathbf{Q}_{ij}))}{\sum_i \exp(\mathcal{N}(\mathbf{Q}_{ij}))} S^t(i), \quad i \in \{1, .., k\} \end{aligned} \tag{3}$$

$$L_{\text{TER}} = -\sum_{i,j}^{hw} g(\hat{S}^T) \log(S^T). \tag{4}$$

C. Class-aware Semantic Continuity Constraint

For surgical frames, there are commonly some **non-discriminative** object regions that are challenging to differentiate, which can lead to recognition inconsistency across the temporal dimension. Enhancing region-level semantic consistency across time can help the model focus on these more non-salient regions. To this end, we introduce the Class-Aware Semantic Continuity Constraint (CSCR). This constraint facilitates the semantic similarities of region-centric representations between local, uncertain target regions and reference global views for each common category, as depicted in the green portion of Fig. 2.

$$L_{CSCR} = \sum_{i=1}^{B_l} \sum_{j=1}^{B_g} \mathrm{I}_{ij} \log \frac{\exp(x_i^{l,T} \cdot x_j^{g,t}/\tau)}{\exp(x_i^{l,T} \cdot x_j^{g,t}/\tau) + \sum_{j'=1}^{B_g} \bar{\mathrm{I}}_{ij'} \exp(x_i^{l,T} \cdot x_{j'}^{g,t}/\tau))} \tag{5}$$

$$B_l = B \times C \times L, B_g = B \times C$$

CAM-based Local View Generation: Specified by the derived CAM, L local crops from the target frame I^T are sampled, with a focus on more uncertain regions. The pseudo multi-label annotation for each crop is then estimated based on the activation area ratio of its corresponding category CAM. If this area ratio exceeds a predefined threshold, the respective category is considered positive within that specific crop.

Temporal Class Tokens Contrast: As mentioned in subsec. II-A, output class tokens contain region-centric semantic information for all categories. From this, we establish CSCR as the multi-label contrastive formulation. This is achieved by minimizing the difference between global reference and local target class tokens. Specifically, the global and local class tokens, derived from batches of global reference views and L local target crops, are first processed into low-dimensional features $x^{g,t} \in \mathbb{R}^{B \times C \times d}$ and $x^{l,T} \in \mathbb{R}^{B \times L \times C \times d}$ through pro-

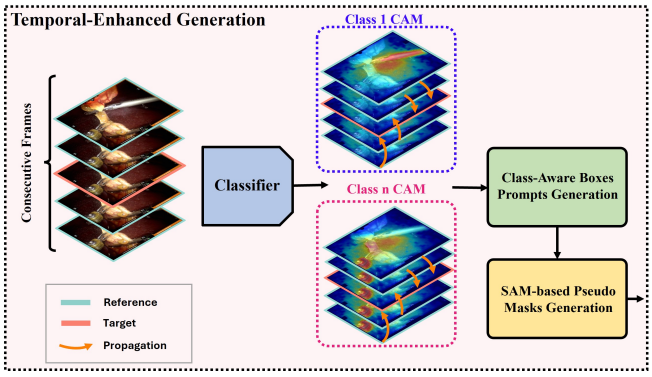


Fig. 3. The flowchart of our temporal-enhanced pseudo masks generation.

jection heads (B denotes batch size, and d represents the projection dimension).

The CSCR can be formulated as Eq.(5), where $\tau \in \mathbb{R}^+$ denotes a temperature parameter. $\mathrm{I} \in \{0,1\}^{B_l \times B_g}$ and $\bar{\mathrm{I}} \in \{1,0\}^{B_l \times B_g}$ represent intra-class matrix and inter-class indicator matrix, respectively. If two items belong to the same category, the corresponding entry in I is assigned a value of 1, whereas the analogous entry in $\bar{\mathrm{I}}$ is assigned a value of 0.

The overall training loss can be expressed as Eq.6.

$$L_{\text{overall}} = L_{CLS} + L_{TER} + L_{CSCR}. \tag{6}$$

D. Temporal-enhanced Pseudo Masks Generation

After training, we generate **temporal-enhanced pseudo masks** rather than frame-wise ones, as depicted in Fig. 3. Given a target frame I^T and its consecutive frames $\{I^{t'}\}_{t'=T-\delta t}^{T+\delta t}$, our architecture first derives the output patch tokens $\{z_p^{t'}\}_{t'=T-\delta t}^{T+\delta t}$ and subsequently generates their corresponding CAMs $\{M^{t'}\}_{t'=T-\delta t}^{T+\delta t}$. CAMs from frames other than I^T are then propagated to time T similar to Eq.(3). The temporal-enhanced CAMs can be expressed as Eq.(7). This scheme can effectively attenuate background information.

$$M^T = \frac{1}{2\delta t+1}\left(M^T + \sum_{t'=(T-\delta t)/T}^{T+\delta t} \Phi((z_p^{t'}, z_p^T), M^{t'})\right) \tag{7}$$

Through a thresholding operation, coarse pseudo masks are derived from CAMs, which subsequently generate class-aware box prompts based on connected components. Finally, SAM [24] without fine-tuning operation processes class-aware bounding box prompts and produces more precise pseudo masks.

E. Main Results

III. EXPERIMENTS AND RESULTS

A. Datasets and Evaluation Metrics

Datasets: In this paper, we conduct experiments based on two surgical video datasets, which were originally designed for phase recognition tasks, including one cholecystectomy surgery benchmark (**Cholec80**) [27] and one real robotic left lateral segment liver surgery dataset (**RLLS**) [28]. Instrument presence labels for RLLS were extracted via Optical Character Recognition (OCR) and noise filtering, while Cholec80 provided those labels. To evaluate LS-SIS methods, we manually annotated instance-wise instrument masks at fixed time steps for both datasets. Specifically, we sampled one frame per minute, excluding *in-vitro* frames, resulting in 1,295 annotated frames for Cholec80 and 2,350 for RLLS. A clinician with three years of experience double-checked all annotations to ensure precision.

TABLE I

COMPARISON OF OURS AGAINST DIFFERENT METHODS ON CHOLEC80, CHOLECINSTANCE, AND RLLS ON SEMANTIC SEGMENTATION(SS) AND INSTANCE SEGMENTATION(IS) METRICS.

Method	SS			IS		
	IoU_C	IoU_S	IoU_{mc}	AP_{50}	AP_{75}	AP
Dataset Cholec80						
WLSTM[19]	48.17	47.55	41.72	44.24	34.15	31.82
MCT[20]	76.21	73.67	59.73	69.22	56.91	54.71
MCT+[21]	77.17	74.87	57.77	69.23	62.20	58.11
MCST[25]	71.94	69.00	66.71	79.01	67.00	61.31
ToCo[26]	66.79	63.99	40.10	58.31	46.21	43.87
Ours	**82.08**	**80.10**	**73.46**	**87.73**	**78.37**	**73.49**
Dataset CholecInstance						
WLSTM[19]	50.20	49.37	37.48	33.69	23.39	22.23
MCT[20]	69.10	66.75	52.03	58.84	47.92	43.29
MCT+[21]	65.94	62.77	47.09	57.67	47.87	42.90
MCST[25]	69.11	65.13	56.12	67.85	52.03	**56.12**
ToCo[26]	65.60	62.02	35.95	51.78	40.13	36.80
Ours	**76.58**	**72.82**	**59.35**	**75.30**	**62.64**	55.98
Dataset RLLS						
WLSTM[19]	65.53	62.68	37.95	53.27	38.66	36.78
MCT [20]	73.42	69.76	42.36	57.48	41.58	40.49
MCT+[21]	74.89	72.12	49.04	66.13	45.95	44.67
MCST[25]	67.71	63.27	42.55	54.11	23.68	27.19
ToCo[26]	61.59	56.99	40.35	56.86	30.00	31.08
Ours	**77.10**	**73.03**	**49.89**	**68.84**	**53.04**	**50.23**

Metrics: We evaluate the proposed method from not only semantic segmentation (SS) but also instance segmentation (IS) perspectives. Our research selects the challenge IoU_C, IoU_S and IoU_{mc} metrics defined in [7] for SS. The formulas of metrics are represented in Eq.(8), wherein P/G mean segmentation predictions/ground truths, $C/\tilde{C}_i$ mean all classes/ground truth classes of the current frame and K represents number of frames. AP metrics in [6] are applied for IS.

$$\begin{aligned}
\text{IoU}_C &= \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{\tilde{C}_i}\sum_{c=1}^{\tilde{C}_i}\frac{P_{i,c}\cap G_{i,c}}{P_{ic}\cup G_{i,c}}\right)\\
\text{IoU}_S &= \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{C}\sum_{c=1}^{C}\frac{P_{i,c}\cap G_{i,c}}{P_{ic}\cup G_{i,c}}\right)\\
\text{IoU}_{mc} &= \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{K}\sum_{i=1}^{K}\frac{P_{i,c}\cap G_{i,c}}{P_{i,c}\cup G_{i,c}}\right)
\end{aligned} \tag{8}$$

Implement Details: Following the two-stage LSSS paradigm, we first train the classification network with instrument presence annotations and generate temporal-enhanced pseudo masks. For Cholec80, the last 40 surgical videos are used for training, while the first 40 videos are reserved for validation. For RLLS, the dataset split aligns with that in [28]. Subsequently, a supervised segmentation model is trained with these pseudo masks (excluding validation frames) and then evaluated with ground-truth annotations.

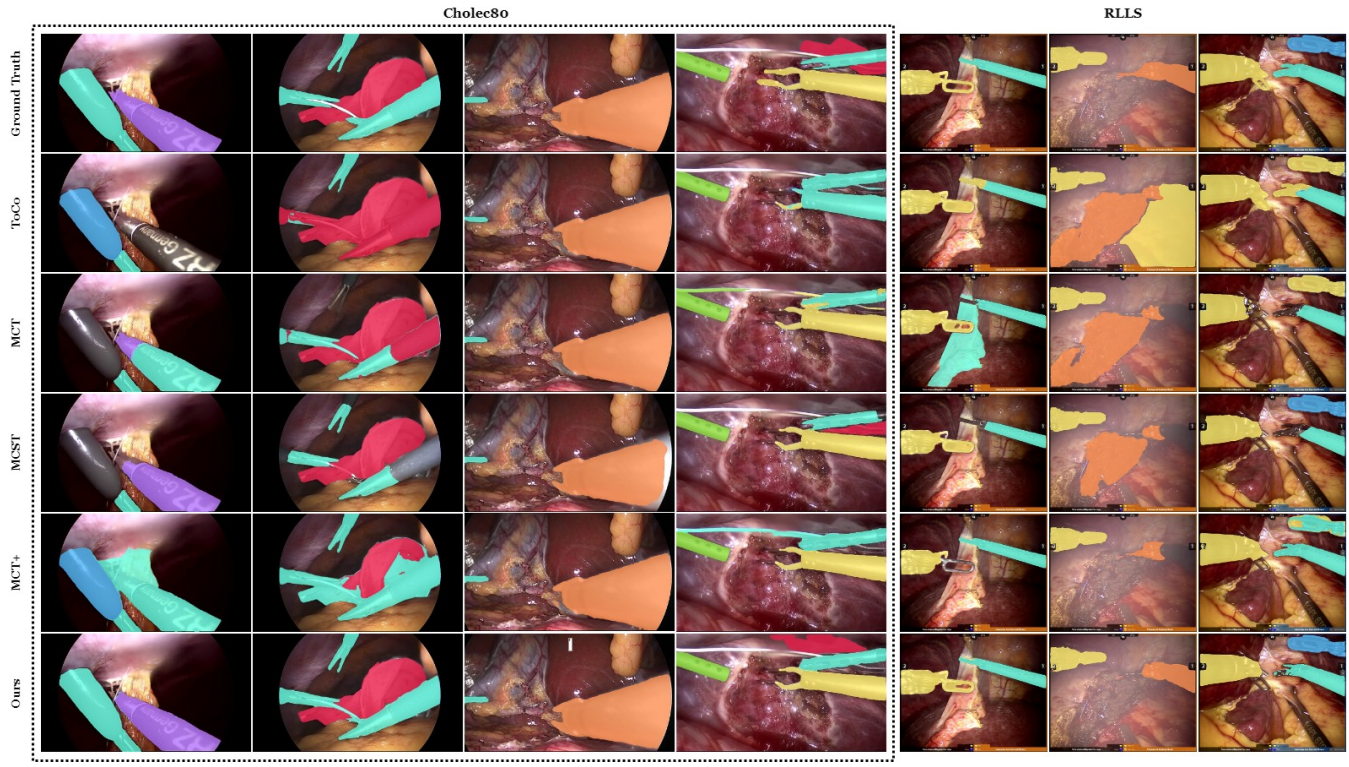


Fig. 4. Qualitative comparisons of segmentation networks for both Cholec80 and RLLS. In comparison, our proposed network achieves the best prediction in most situations.

The inference time is contingent upon the chosen segmentation model's architecture. In this study, we select Mask2Former [6] as the segmentor for joint semantic and instance segmentation, achieving a inference speed of approximately 0.01 seconds per frame (1 A800 GPU).

Cholec: Results of different methods are presented in Table I. **Bold** indicates the optimal results, while underlined represents the suboptimal results. In this paper, we select not only some typical LSSS methods like MCT [20], MCT+ [21], ToCo [26], MCST [25], but also other similar works in surgical domain like WLSTM [19] for comparison. For SS, LS-Surg outperforms all other methods by a considerable margin. In particular, our method improves over MCT+ by 6% IoU_C, 7% IoU_S while it improves over MCST by 10% IoU_{mc}. For IS, LS-Surg is superior to other methods similarly. Ours improves over MCST by 11% AP_{50}, 16% AP_{75}, and 19% AP.

We also evaluate these trained methods with the validation set of CholecInstance [29], a novel instance segmentation dataset for cholecystectomy surgery, to present the generality of effectiveness. Similar trends can be observed. Following the official dataset split, we also trained an identical segmentation model under supervised learning framework, achieving segmentation metrics of *85.17* IoU_C, *81.28* IoU_S, *64.58* IoU_{mc} for SS and *85.78* AP_{50}, *78.76* AP_{75}, *70.28* AP for IS. The performance gap between ours and supervised one remains within acceptable margins.

RLLS: As presented in Table I, for SS, Ours surpasses the sub-optimal method with 2.9% improvement in IoU_C, 1.3% increase in IoU_S and substantial 1.6% enhancement in IoU_{mc}. Similarly, in the context of IS, our method improves 4.1% AP_{50}, 15.4% AP_{75} and 12.4% AP. These metrics validate the importance of temporal properties in surgical video for label-supervised SIS.

B. Ablation Study

The ablation study results of key components are presented in Table. II. The first row shows the performance of MCT+. Purely incorporating the temporal equivariance relation via TER, our method significantly improves effectiveness by 5.2% IoU_C, 4.4% IoU_S, 10.9% IoU_{mc} for SS and by 20.4% AP_{50}, 19.8% AP_{75}, 21.1% AP for IS. The CSCR further enhances the baseline across all metrics, with a particularly notable increase in IoU_{mc}. Replacing frame-wise pseudo masks with our TMG continues to boost results, peaking at *82.08* IoU_C, *80.10* IoU_S on SS, and *87.73* AP_{50}, *78.37* AP_{75}, *73.49* AP on IS. The IoU_{mc} metric, however, shows a slight decrease.

Beyond the final segmentation results, we also evaluate SS on both pseudo masks derived from CAM seeds (PM(C)) and refined pseudo masks from SAM (PM(S)). Given only instrument presence annotations, IoU_C metrics are similar to IoU_S. As shown in Table. II, consistent improvement trends are observed across all metrics. Specifically, results show an increase of *11.6* IoU_S and *15.4* IoU_{mc} for PM(C), and *7.8* IoU_S and *15.5* IoU_{mc} for PM(S). Compared to PM(S), the method demonstrates a more significant improvement for PM(C).

TABLE II

PERFORMANCE OF KEY COMPONENTS ON CHOLEC80. TER MEANS TEMPORAL EQUIVARIANCE CONSTRAINT, CSCR MEANS CLASS-AWARE SEMANTIC CONTINUITY CONSTRAINT AND TMG MEANS TEMPORAL-ENHANCED PSEUDO MASKS GENERATION. PM(C) REPRESENTS PSEUDO MASKS FROM CAMS SEEDS WHILE PM(S) REPRESENTS REFINED PSEUDO MASKS AFTER SAM.

Key Componets			SS			IS			PM(C)		PM(S)	
TER	CSCR	TMG	IoU_C	IoU_S	IoU_{mc}	AP_{50}	AP_{75}	AP	IoU_S	IoU_{mc}	IoU_S	IoU_{mc}
			77.17	74.87	57.77	69.23	62.20	58.11	49.02	43.95	61.89	52.95
✓			81.15	78.15	64.11	83.38	74.55	70.37	56.29	55.94	68.13	62.39
✓	✓		81.60	79.51	**73.85**	84.44	76.93	71.62	60.53	57.27	69.48	65.02
✓	✓	✓	**82.08**	**80.10**	73.46	**87.73**	**78.37**	**73.49**	**60.71**	**59.44**	**69.70**	**68.53**

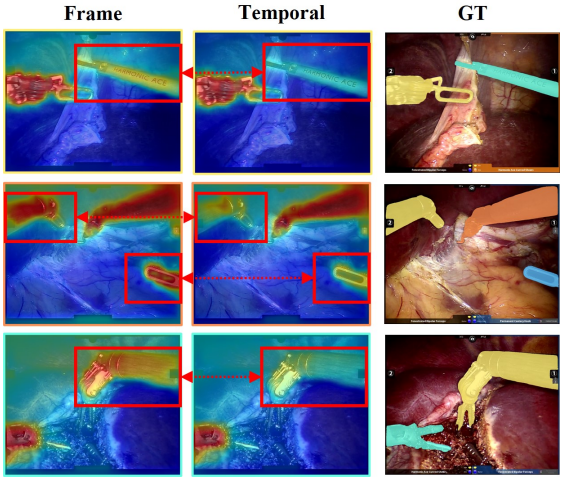


Fig. 5. Qualitative comparison for frame-wise and temporal-enhanced CAMs. The red box represents the irrelevant regions that need to be suppressed for the current category.

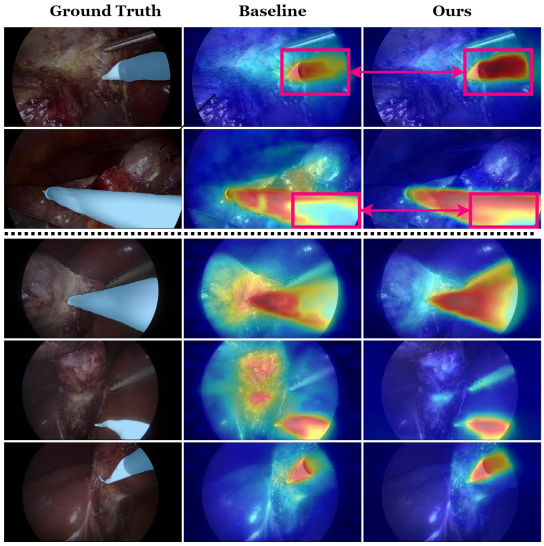


Fig. 6. Qualitative comparison of CAMs on low confidence and appearance variation situations. The first two rows represent low confidence regions while the last three rows mark the consecutive three frames with appearance variation.

C. Visualization

Fig. 4 illustrates the final segmentation results of various methods compared on the Cholec80 and RLLS datasets. Our method consistently demonstrates superior segmentation performance in most cases, exhibiting accurate recognition and comprehensive shape preservation across multiple categories. To further substantiate the effectiveness of the TMG mechanism, we present several examples of temporal-enhanced CAMs alongside their corresponding frame-wise CAMs. As highlighted by the red boxes in Fig. 5, temporal-enhanced CAMs effectively suppress irrelevant regions for the current category. This mechanism enhances focus on category-specific foreground information, thereby improving the accuracy of PM(C) generation.

In the context of label-supervised SIS, low-confidence regions and appearance variations commonly undermine methodological efficacy. From this, we also visualize the CAMs of both the baseline and our proposed method under these challenging situations to validate robustness. As shown in the first two rows of Fig. 6, our method can activate more low-confidence object regions than the baseline. For the last three rows, representing three adjacent frames, the CAMs generated by our method demonstrate greater robustness against appearance variations over time.

IV. CONCLUSION

In this paper, we introduce LS-Surg, a novel approach for label-supervised surgical instrument segmentation (LS-SIS) that relies exclusively on instrument presence annotations. LS-Surg extends a two-stage LSSS framework by incorporating temporal equivariance and class-aware semantic continuity constraints, alongside a temporal-enhanced pseudo-mask generation mechanism. Our extensive experiments on Cholec80 and RLLS datasets validate the effectiveness of LS-Surg. However, the proposed method has several limitations. (1) The lack of tissue presence annotations restricts its applicability to label-supervised tissue segmentation. (2) The two-stage design increases training time. Despite these limitations, our temporal learning strategies can be seamlessly integrated into other frameworks. Future work will focus on adapting them to a single-stage architecture to reduce computational cost and on extending the method to open-vocabulary tissue and organ segmentation using vision-language models. We also plan to explore the use of segmentation outputs in downstream high-level understanding tasks.

REFERENCES

[1] N. Rieke, D. J. Tan, C. A. di San Filippo, F. Tombari, M. Alsheakhali, V. Belagiannis, A. Eslami, and N. Navab, "Real-time localization of articulated surgical instruments in retinal microsurgery," *Medical image analysis*, vol. 34, pp. 82–100, 2016.

[2] S. K. Hasan and C. A. Linte, "U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2019, pp. 7205–7211.

[3] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 412–420.

[4] F. Isensee and K. H. Maier-Hein, "Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images," *arXiv preprint arXiv:2004.12668*, 2020.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[6] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[7] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: an instance-based approach for surgical instrument segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 595–605.

[8] N. Ayobi, A. Pérez-Rondón, S. Rodríguez, and P. Arbeláez, "Matis: Masked-attention transformers for surgical instrument segmentation," *arXiv preprint arXiv:2303.09514*, 2023.

[9] B. Baby, D. Thapar, M. Chasmai, T. Banerjee, K. Dargan, A. Suri, S. Banerjee, and C. Arora, "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*, 2023, pp. 6191–6201.

[10] R. R. Dhanakshirur, K. A. Shastry, K. Borgavi, A. Suri, P. K. Kalra, and C. Arora, "Learnable query initialization for surgical instrument instance segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 728–738.

[11] Z. Zhao, Y. Jin, and P.-A. Heng, "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)*. IEEE, 2022, pp. 11 186–11 193.

[12] D. Liu, Y. Wei, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 657–667.

[13] M. Sahu, A. Mukhopadhyay, and S. Zachow, "Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation," *International journal of computer assisted radiology and surgery*, vol. 16, no. 5, pp. 849–859, 2021.

[14] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno, and N. Padoy, "Fun-sis: A fully unsupervised approach for surgical instrument segmentation," *Medical Image Analysis*, vol. 85, p. 102751, 2023.

[15] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*. Springer, 2019, pp. 440–448.

[16] M. Wei, C. Budd, L. C. Garcia-Peraza-Herrera, R. Dorent, M. Shi, and T. Vercauteren, "Segmatch: A semi-supervised learning method for surgical instrument segmentation," *arXiv preprint arXiv:2308.05232*, 2023.

[17] Z. Yang, R. Simon, and C. Linte, "A weakly supervised learning approach for surgical instrument segmentation from laparoscopic video sequences," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 12034. SPIE, 2022, pp. 412–417.

[18] R. Sanchez-Matilla, M. Robu, I. Luengo, and D. Stoyanov, "Scalable joint detection and segmentation of surgical instruments with weak supervision," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 501–511.

[19] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1059–1067, 2019.

[20] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4310–4319.

[21] L. Xu, M. Bennamoun, F. Boussaid, H. Laga, W. Ouyang, and D. Xu, "Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation," *arXiv preprint arXiv:2308.03005*, 2023.

[22] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[23] M. Salehi, E. Gavves, C. G. Snoek, and Y. M. Asano, "Time does tell: Self-supervised time-tuning of dense image representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 536–16 547.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[25] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Learning multi-modal class-specific tokens for weakly supervised dense object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 596–19 605.

[26] L. Ru, H. Zheng, Y. Zhan, and B. Du, "Token contrast for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3093–3102.

[27] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[28] S. Zhao, Y. Liu, Q. Wang, D. Sun, R. Liu, and S. K. Zhou, "Murphy: relations matter in surgical workflow analysis," *arXiv preprint arXiv:2212.12719*, 2022.

[29] O. Alabi, K. K. Z. Toe, Z. Zhou, C. Budd, N. Raison, M. Shi, and T. Vercauteren, "Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery," *arXiv preprint arXiv:2406.16039*, 2024.