TraSeTR: Track-to-Segment Transformer with Contrastive Query for Instance-level Instrument Segmentation in Robotic Surgery

Zixu Zhao, Yueming Jin, and Pheng-Ann Heng

Abstract-Surgical instrument segmentation - in general a pixel classification task - is fundamentally crucial for promoting cognitive intelligence in robot-assisted surgery (RAS). However, previous methods are struggling with discriminating instrument types and instances. To address above issues, we explore a mask classification paradigm that produces per-segment predictions. We propose TraSeTR, a novel Track-to-Segment Transformer that wisely exploits tracking cues to assist surgical instrument segmentation. TraSeTR jointly reasons about the instrument type, location, and identity with instance-level predictions *i.e.*, a set of class-bbox-mask pairs, by decoding query embeddings. Specifically, we introduce the prior query that encoded with previous temporal knowledge, to transfer tracking signals to current instances via identity matching. A contrastive query learning strategy is further applied to reshape the query feature space, which greatly alleviates the tracking difficulty caused by large temporal variations. The effectiveness of our method is demonstrated with state-of-the-art instrument type segmentation results on three public datasets, including two **RAS** benchmarks from EndoVis Challenges and one cataract surgery dataset CaDIS.

Index Terms—AI-based methods, Transformer, surgical instrument segmentation, medical robotics

I. INTRODUCTION

Robot-assisted surgery (RAS) has revolutionized the minimally invasive surgery by remarkably extending the dexterity and overall capability of surgeons. The robotic system controls the movement of surgical instruments, enabling efficient manipulation and vivid observation for many surgical tasks [1]-[3]. Intelligent parsing of such instruments, e.g., identifying their types or positions, is highly desired for promoting cognitive assistance to surgeon perception [4], operating workflow optimization [5], and skill assessment [6], [7]. To this end, the instance-level semantic segmentation of instruments, which can separate instruments to different types, is required as a fundamental task to support many downstream applications, such as tool pose estimation [8], tracking [9], trajectory prediction [10], [11], and even facilitate the surgical task automation [12], [13] for next generation of operating intelligence.

The similar instrument types, with small inter-class discrepancy, are challenging to recognize, especially in complex



Fig. 1. Two paradigms for surgical instrument type segmentation: (i) **pixel** classification predicts a single class for each pixel (C classes in total). (ii) mask classification predicts a set of binary masks and assigns a single class to each mask. To account for the large temporal variations of instruments, we exploit tracking cues with prior queries for video-level mask classification.

surgical scenes. Most of previous instrument segmentation methods follow a *pixel classification* paradigm in which the deep learning model predicts the probability distribution over all classes for each pixel in a frame. The key idea of them [14]–[18] is to modify the neural network (*e.g.*, U-Net [19]) or differentiate instrument types by exploring spatial or temporal cues, including depth maps [20], pose estimation [21], optical flows [15], and motion flows [16]. Nevertheless, as shown in Fig. 1, these solutions are struggling with the spatial class inconsistency problem, where one instrument may be assigned multiple instrument types.

An alternative paradigm – mask classification that predicts a set of binary masks, and each associated with a single class – has been increasingly adopted for instance-level segmentation. In robotic surgery, ISINet [22] takes the first step to predict a single class for each instrument segment based on Mask-RCNN [23]. One main challenge is to maintain the class consistency over time. The relabelling strategy in [22] takes into account the predictions of previous frames, but tends to misassign labels to similar instances due to large temporal variations (see Fig. 1). How to correctly perform video-level mask classification for surgical instruments, to tolerate the intra-class variations across time, is crucial yet still remains unexplored.

Recently, Transformers have shed light on mask classification by jointly reasoning about a number of query embeddings for instance predictions via the encoder-decoder attention mechanism [24]. The development of DETR [25] and its variants [26] have widely demonstrated promising performance on object tracking [27] and instance segmentation [28], [29]. Specifically in RAS scenarios, there emerge

This work was supported by the Key-Area Research and Development Program of Guangdong Province, China under Grant 2020B010165004, Hong Kong RGC TRS project T42-409/18-R, and a grant from the National Natural Science Foundation of China with Project No. U1813204.

Z. Zhao and P. A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, P. A. Heng is also with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Y. Jin is with the Department of Computer Science, University College London, UK.

Corresponding author: Zixu Zhao (zxzhao@cse.cuhk.edu.hk)



Fig. 2. **TraSeTR overview.** We use a backbone to extract frame feature \mathcal{F} and a pixel decoder to produce pixel-level embeddings $\mathcal{E}_{\text{pixel}}$. A transformer module computes from \mathcal{F} and N queries to yield instance embeddings $\mathcal{E}_{\text{inst}}$, which are combined with $\mathcal{E}_{\text{pixel}}$ to output N binary masks with bounding boxes and classes. TraSeTR performs instance tracking for surgical instrument segmentation. It is realized by a two-stage identity matching mechanism that allows the *prior queries* to be precisely transformed to the current predictions, and the *current queries* to infer newly-entered instances either initializing prior queries for future frame. Based on the matching set σ , the contrastive query learning is applied to instance embeddings between time t and $t - \tau$.

some works actively exploring transformer-based solutions for surgical phase recognition [30], tool detection [31], and surgical scene reconstruction [32]. All these successes motivate us to explore the potential of Transformer for discriminating instrument types in a way of leveraging the temporal knowledge of dynamic instances.

In this paper, we propose TraSeTR, a novel track-tosegment transformer that dynamically integrates tracking cues to assist instance-level surgical instrument segmentation. Rooted in the mask classification paradigm, TraSeTR employs a Transformer module [24] to infer a set of instance predictions from current frame features and query embeddings, each consisting of a class prediction, a bounding box prediction, and a binary mask prediction. The query embeddings are initialized as (i) prior queries to encode prior knowledge of previous frame instances and (ii) current queries to detect newly-appeared instances in a frame. Furthermore, they are learned to be temporally contrastive to tolerate the dynamic changes of instruments. TraSeTR performs instance tracking via a two-stage identity matching between prediction set and ground-truth set such that the current instances can be precisely inferred from their corresponding prior queries. Our main contributions include:

- We propose a novel track-to-segment transformer to wisely discriminate both *instances* and *types* for accurate surgical instrument segmentation.
- The keys to the success of TraSeTR are the new mechanisms of *identity matching* and *contrastive query learning*, which are carefully designed to track surgical instruments with large temporal variations.
- We design a link-by-link inference strategy to infer instrument segments with dynamically changed prior queries in an auto-regressive fashion.
- We extensively evaluated the proposed method on two RAS benchmarks EndoVis17 [33] and EndoVis18 [34],

and a public eye surgery dataset CaDIS [35]. TraSeTR set new state-of-the-art results on instrument type segmentation tasks at a fast speed of 23 FPS.

II. METHODS

TraSeTR is a mask classification model for instance-level surgical instrument segmentation. In this section, we first describe how to formulate the track-to-segment problem. Then, we introduce the architecture and matching mechanism of the model. Finally, we illustrate the training and inference strategies which are specifically designed for TraSeTR.

A. Track-to-Segment Formulation

TraSeTR segments the instruments by (i) partitioning the frame into N instances where N is significantly larger than the real instrument number \tilde{N} , and (ii) merging the selected instances into one segmentation map over C classes. The predictions of N instances are a set of N probability-bbox-mask pairs $z = \{(p_i, b_i, m_i)\}_{i=1}^N$, where the probability distribution p_i contains an auxiliary "no object" label (\emptyset) to denote instances that do not correspond to any classes; $b_i \in [0, 1]^4$ is the bounding box of the instance; and m_i is the binary mask. To train the mask classification model, a matching set σ between the prediction set z and the ground-truth set $\tilde{z} = \{(\tilde{c}_i, \tilde{b}_i, \tilde{m}_i)\}_{i=1}^{\tilde{N}}$ is required, where $\tilde{c}_i \in \{1, ..., C\}$. The predictions that are not matched by σ are assigned with the ground-truth label \emptyset .

For accurate video-level mask classification, TraSeTR incorporates tracking cues to assist segmentation by tracking instrument instances $\{z_1, ..., z_k\}$ with identity $K = \{1, ...k\}$. Specifically, TraSeTR initialize N query embeddings for instance prediction. Among them, N^{prior} prior queries are initialized with the output embeddings of previous frame, and N^{cur} current queries are randomly initialized to learn to detect newly-entered instruments. Once the matching set σ

assigns the predictions from prior queries to corresponding ground truths, the instances are successfully tracked.

B. TraSeTR Architecture

The overall architecture of TraSeTR is simple and depicted in Fig. 2. We now describe three basic components that enable instance-level segmentation.

Frame extractor. Taking a frame at time t as input, the CNN backbone generates a feature map $\mathcal{F} \in \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times C_{\mathcal{F}}}$, where $C_{\mathcal{F}}$ is the number of channels, S is the sampling stride, and $H \times W$ is the frame size. The feature map is then upsampled via a decoder to produce pixel-level embeddings $\mathcal{E}_{\text{pixel}} \in \mathbb{R}^{HW \times C_{\mathcal{F}}}$, where $C_{\mathcal{E}}$ is the number of channels.

Transformer module. We employ a standard encoderdecoder Transformer [24] to compute from frame features \mathcal{F} and N query embeddings (including N^{prior} prior queries from time $t - \tau$ and N^{cur} current queries). We add \mathcal{F} with positional embeddings as Transformer is permutation invariance. Thanks to the self- and cross-attention mechanisms in Transformer, its output, *i.e.*, N instance embeddings $\mathcal{E}_{\text{inst}} \in \mathbb{R}^{N \times C_Q}$ of dimension C_Q , encode global information about all instruments appeared in the surgical scene. Note that the channel number of query embeddings is C_Q as well.

Instance fusion module. It first maps instance embeddings \mathcal{E}_{inst} to three types of predictions. For class prediction, we apply a Multi-Layer Perceptron (MLP) followed by a softmax function to yield class probability predictions $\{p_i\}_{i=1}^N$. For bounding box prediction, we use a MLP to produce N bounding boxes $\{b_i\}_{i=1}^N$. For mask prediction, we convert \mathcal{E}_{inst} to mask embeddings $\mathcal{E}_{mask} \in \mathbb{R}^{N \times C_Q}$ with a MLP, which are used to generate N binary mask predictions $\{m_i\}_{i=1}^N$ via a dot production with pixel-level embeddings \mathcal{E}_{pixel} , followed by a sigmoid activation. Finally, instances are fused to instrument type segmentation. We assign each pixel [h, w] to one of the probability-bbox-mask pair from the matching set σ via $\operatorname{argmax}_{i \in \sigma} p_i(c_i) \cdot m_i[h, w]$. Here, c_i is the predicted class $c_i = \operatorname{argmax}_{c \in \{1, \dots, C, \emptyset\}} p_i(c)$.

C. Tracking with Identity Matching

TraSeTR infers a dynamic-size set of N predictions in a single pass of the model, where N is changing with the number of prior queries. To score the N predictions with respect to the \tilde{N} ground truths $(N > \tilde{N})$, we need to find a matching set σ such that the j^{th} ground truth matches the prediction with index $\sigma(j)$. Essentially, the correct link between the prediction from prior query and the current ground truth can be regarded as a tracking task, which is crucial to enhance the temporal class consistency.

However, tracking instruments is difficult because of the large temporal variations (*e.g.*, tool tips) caused by the zoom in and zoom out of endoscopic camera. The *bipartite matching* used in DETR [25] and TrackFormer [27], *i.e.*, minimizing an assignment cost between the prediction $z_{\sigma(j)}$ and ground truth \tilde{z}_j , $\mathcal{L}_{assign}(z_{\sigma(j)}, \tilde{z}_j) = p_{\sigma(j)}(\tilde{c}_j) + \mathcal{L}_{box}(b_{\sigma(j)}, \tilde{b}_j)$, tends to cause invalid tracking by matching the prediction from current query to the ground truth.

To correctly track instances, we propose a two-stage identity matching strategy. At the first stage, we search for a matching subset σ_1 between predictions from prior queries and the current ground truths. A naive matching is possible if all instances belong to different classes. In this case, the j^{th} ground truth matches to $\sigma_1(j)^{\text{th}}$ prediction if their classes are the same. A more general method can be achieved by employing the prior ground truths $\{\tilde{z}_i^{\text{prior}}\}_{i=1}^{N^{\text{prior}}}$ of these predictions for cost calculation, such that instrument identities can be explicitly associated with current ground truths $\{\tilde{z}_i\}_{i=1}^{\tilde{N}}$ via a matching cost defined as:

$$\mathcal{L}_{\mathrm{match}}(z_{\sigma_1(j)}, \tilde{z}_j) = \mathbb{1}_{\tilde{c}_{\sigma_1(j)}^{\mathrm{prior}} = \tilde{c}_j} |\tilde{x}_{\sigma_1(j)}^{\mathrm{prior}} - \tilde{x}_j|, \qquad (1)$$

where \tilde{x}_j and $\tilde{x}_{\sigma_1(j)}^{\text{prior}}$ are the horizontal coordinates of the j^{th} ground truth's center and $\sigma_1(j)^{\text{th}}$ prior ground truth's center, derived from the bounding boxes \tilde{b}_j and $\tilde{b}_{\sigma_1(j)}^{\text{prior}}$, respectively. We then define the *relative horizontal distance* (RHD) as $|\tilde{x}_{\sigma_1(j)}^{\text{prior}} - \tilde{x}_j|$ to distinguish instances, especially those of the same class, since the horizontal displacement of the same instance is smaller than that of the other. It can be inherently supported by the "dual-arm distribution", *i.e.*, one instrument moves within the right-half scene, while the other (of the same class) within the left-half scene. Unlike the IoU-based bounding box loss [36], RHD could still track the fast moving instruments in robotic surgery. The optimal subset then can be found by minimizing the matching cost:

$$\hat{\sigma}_1 = \operatorname*{arg\,min}_{\sigma_1} \sum_j \mathcal{L}_{\mathrm{match}}(z_{\sigma_1(j)}, \tilde{z}_j), \ \sigma_1(j) \in [1, N^{\mathrm{prior}}].$$
(2)

At the second stage, we search for a matching subset σ_2 between ground truths that are not matched by $\hat{\sigma}_1$ and predictions of current queries, to detect the newly-entered instance. Here, we optimize a bipartite matching-based assignment:

$$\hat{\sigma}_2 = \operatorname*{arg\,min}_{\sigma_2} \sum_j \mathcal{L}_{\mathrm{assign}}(z_{\sigma_2(j)}, \tilde{z}_j), \ \sigma_2(j) \in (N^{\mathrm{prior}}, N].$$
(3)

We assign a new identity (*i.e.*, ID) for the newly-entered instance if its class appears in the video for the first time or there have already exist instances of the same class. Otherwise, it will be assigned a previous identity according to the class. Overall, the optimal matching set is a union of the subsets, *i.e.*, $\sigma = \hat{\sigma}_1 + \hat{\sigma}_2$, where $\hat{\sigma}_2$ can be \emptyset .

D. Contrastive Learning-based Training

Given the matching set σ , we encourage the model to transform the prior query to its current position, even with large temporal variations. Basically, we could compute the *Hungarian loss* for all matched pairs, which are commonly used in prior works [25], [27], [29]. This loss contains a negative log-likelihood for class prediction, a bounding box loss and a mask loss for all instances:

$$\mathcal{L}_{\text{Hung}}(z,\tilde{z}) = \sum_{j=1}^{\tilde{N}} [-\log p_{\sigma(j)}(\tilde{c}_j) + \mathcal{L}_{\text{box}}(b_{\sigma(j)},\tilde{b}_j) + \mathcal{L}_{\text{mask}}(m_{\sigma(j)},\tilde{m}_j)].$$
(4)

Here, we use the same \mathcal{L}_{mask} and \mathcal{L}_{box} as DETR [25]. Specifically, \mathcal{L}_{mask} is a linear combination of dice loss [37] and focal loss [38]. \mathcal{L}_{box} is a linear combination of ℓ_1 loss and a generalized IoU loss \mathcal{L}_{iou} [36].

Contrastive query learning. To further improve the tracking capability of TraSeTR based on the matching index σ , we push the model towards maximizing the query agreements on different temporal views of the instance. Let us denote N instance embeddings \mathcal{E}_{inst} as $\{e_i\}_{i=1}^N$ and some of them are decoded from prior query embeddings. Unlike previous video-level contrastive learning [39], [40], our key idea (see Fig. 2) is to pull the instance embedding e_i ($i \in \sigma$) with its prior query e_i^{prior} , while pushing e_i with the remaining instance embeddings e_j where $c_i \neq c_j$. Formally, we construct two types of contrastive pairs, including the positive pairs (e_i, e_i^{prior}) and the negative pairs (e_i, e_j), which allow us to learn contrastive query embeddings via a contrastive loss:

$$\mathcal{L}_{\rm ctr} = -\frac{1}{|\sigma|} \log \sum_{i \in \sigma} \frac{\phi(e_i, e_i^{prior})}{\phi(e_i, e_i^{prior}) + \sum_j \phi(e_i, e_j)}, \quad (5)$$

where $\phi(\cdot, \cdot)$ is a similarity function and can be achieved by a dot production between two embeddings. In Eq. (5), we omit the calculation of instance embeddings that are matched by σ but not have prior embeddings, *i.e.*, those newly-entered instances. We now define the overall training loss as $\mathcal{L}_{Hung} + \lambda_{ctr}\mathcal{L}_{ctr}$, where λ_{ctr} is a balancing weight.

E. Link-by-link Inference

We describe a link-by-link inference procedure specifically designed for instance-level instrument segmentation. At the beginning, TraSeTR predicts all the instruments that appear in the first frame with the identity $K_1 = \{1, ..., k\}$ being a subset of all K in the video sequence. It only decodes N^{cur} instance embeddings and then select k instances whose classification scores are above τ_d . The k output embeddings are used to initialize the prior queries for the next frame. At time t, TraSeTR outputs $N^{cur} + N^{prior}$ instance embeddings. Apart from detecting newly-appeared instruments with classification threshold τ_d , TraSeTR tracks prior instances whose classification scores are above τ_t . Note that N^{prior} changes between frames as prior queries are removed or new instances are detected. We remove prior queries if their classification scores drop below τ_t for more than 50 time steps. The time tolerance allows TraSeTR to infer an instance with multiple prior queries collected from different time steps, which provides the long-range temporal information of one certain instrument.

III. EXPERIMENTS AND RESULTS

We evaluated TraSeTR's performance on instrument type segmentation using three public datasets [33]–[35].

A. Datasets and Evaluation Metrics

EndoVis17. The EndoVis17 dataset [33], a benchmark of instrument type segmentation, contains 8 robot-assisted surgery videos recorded from da Vinci Xi Surgical System. We used the instance annotations generated by [22].

EndoVis18. The EndoVis18 dataset [34] provides 15 videos of different porcine procedures acquired by da Vinci

Xi Surgical System, and corresponding semantic annotations of the whole scene. The instruments are annotated with their parts (*shaft, wrist* and *jaws*). To distinguish among instrument types, we followed prior work [22] to generate the additional instance annotations for 7 instrument types.

CaDIS. The CaDIS dataset [35] includes 25 surgical videos recording cataract surgery by an OPMI Lumera T microscope. We used the semantic annotations of instruments and converted them to instance annotations by extracting each instrument from the scene and assigning it one of the 10 instrument types in Table II. Our annotations can be transformed to the MS-COCO standard dataset format.

As the ground-truth instance ID is not provided, we only assess the segmentation quality following prior works [14]–[16], [22], [35]. Specifically, we adopted two commonly used metrics including mean intersection-over-union (mIoU) and Dice coefficient (Dice) that only consider the classes presented in a frame. For fair comparisons, EndoVis17 was evaluated by 4-fold cross-validation using the standard folds described in [14]. EndoVis18 and CaDIS were evaluated using the same data splitting as [22] and [35], respectively.

B. Implementation Details

Model settings. TraSeTR is compatible with any backbone architecture. In this work, we use the ResNet-50 [41] backbone. The Transformer consists of 6 encoder and 6 decoder layers with 8 attention heads. As TraSeTR predicts $1 \sim 4$ instruments for each frame, the query number N^{cur} is 20 (see query number ablation in supplementary video).

Training. The time interval τ between prior frame and current frame is in the range of [1, 10]. Our model is implemented in Pytorch and trained with a NVIDIA Titan Xp GPU. We initialize the backbone with pretrained weights on COCO [42]. The initial learning rates of Transformer and backbone are 1e-4 and 1e-5, which will be multiplied by 0.1 after 50 epochs. We use the same balancing weights as DETR [25] for $\mathcal{L}_{\text{Hung}}$. The hyper-parameter λ_{ctr} is set as 0.2. To increase model robustness, we augment query embeddings by adding false negatives with a probability of 0.4 and false positives with a probability of 0.1, following [27].

Inference. During inference, we set the detection threshold τ_d as 0.9. To tolerate large temporal variations of instruments, we set the track threshold τ_t as 0.6. We also apply non-maximum suppression (NMS) with a high IoU threshold of 0.9 to filter out overlapped instances. The inference speed could be 23 FPS without extra acceleration.

C. Main Results

1) EndoVis 17&18: In Table I, we compare our TraSeTR with (i) pixel classification approaches [14]–[16] and (ii) mask classification approaches [22], [26], [27] for instrument type segmentation. For DETR [26] and TrackFormer [27], we use the source code provided by the authors to train the models. Other results are reported from the original papers. For EndoVis17, TraSeTR outperforms Dual-MF [16] by 14.6% mIoU and 9.1% Dice, indicating that mask classification formulation has great potential for instrument

TABLE I

INSTRUMENT TYPE SEGMENTATION RESULTS OF DIFFERENT METHODS ON ENDOVIS17 AND ENDOVIS18 DATASETS (7 CLASSES).

Dataset	Method	Bbox	mIoU	Dice	Instrument classes (mIoU)						
					BF	PF	LND	VS / SI	GR / CA	MCS	UP
EndoVis17	TernausNet [14]		35.3	44.9	13.3	12.4	20.5	6.0	1.1	1.0	16.8
	MF-TAPNet [15]		37.4	48.0	16.4	14.1	19.0	8.1	0.3	4.1	13.4
	Dual-MF [16]		45.8	56.1	34.4	21.5	64.3	24.1	0.8	17.9	21.8
			53.1	58.0	36.5	37.2	54.5			23.3	- <u>11.3</u>
	TrackFormer [27]	\checkmark	54.9	59.7	37.6	38.0	53.1	25.5	2.8	24.6	15.7
	ISINet [22]	\checkmark	55.6	62.8	38.7	38.5	50.1	27.4	2.0	28.7	12.6
	TraSeTR (ours)		60.4 (+4.8)	65.2 (+2.4)	45.2	56.7	55.8	38.9	<u>1</u> 1.4	31.3	18.2
EndoVis18	TernausNet [14]		46.2	53.2	44.2	4.7	0.0	0.0	0.0	50.4	0.0
	MF-TAPNet [15]		67.9	72.5	69.2	6.1	11.7	14.0	0.9	70.2	0.6
	Dual-MF [16]		70.4	76.9	74.1	6.8	46.0	30.1	7.6	80.9	0.1
				72.5	70.3	15.9	31.6	- 16.7		80.2	- 0.0 -
	TrackFormer [27]	\checkmark	71.1	77.3	75.8	20.1	38.5	30.6	4.8	82.5	1.5
	ISINet [22]	\checkmark	73.0	78.3	73.8	48.9	31.0	37.7	0.0	88.2	2.2
	TraSeTR (ours)		76.2 (+3.2)	81.0 (+2.7)	76.3	-53.3	46.5	40.6	<u> </u>	86.3	⁻ 17.5

Instrument classes include: Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Vessel Sealer (VS), Suction Instrument (SI), Grasping Retractor (GR), Clip Applier (CA), Monopolar Curved Scissors (MCS), and Ultrasound Probe (UP).



Fig. 3. **Qualitative comparisons** of TernausNet [14], MF-TAPNet [15], Dual-MF [16], DETR [26], TrackFormer [27], ISINet [22], and our TraSeTR on EndoVis17 (top) and EndoVis18 (bottom) videos. Each color represents one instrument type. More results can be found in supplementary video.

type segmentation. Compared with transformer-based approaches [26], [27], TraSeTR shows great improvements by tracking instances such that prior queries can be leveraged to infer current instances. TraSeTR is also superior to ISINet [22], achieving a new state-of-the-art of 60.4% mIoU and 65.2% Dice. For EndoVis18, TraSeTR still outperforms the prior state-of-the-art [22] by 3.2% mIoU and 2.7% Dice. In particular, the improvements of some types, *e.g.*, Clip Applier and Ultrasound Probe, are more than 10% mIoU. Fig. 3 shows the qualitative comparisons. As expected, TraSeTR maintains both spatial and temporal class consistency of instruments, while other methods fail to do so.

2) CaDIS: Table II compares the type segmentation results of TraSeTR and three strong baselines reported in [35], including DeepLabV3+ [43], UPerNet [44], and HRNetV2 [45]. Observe that TraSeTR achieves the best overall results of 69.9% mIoU. For some certain types, *e.g.*, Ph. Handpiece and I/A Handpiece, the similar tool tips could be better distinguished by the high-resolution representations of HRNetV2 [45]. But promisingly, TraSeTR peaks the segmentation performance of 7 instrument types. This result suggests that our method is robust to various surgical instruments and surgical scenes.

3) Instance-level Tracking: Fig. 4 visualizes the tracking process of our method. Each color bar represents an instrument with a specific identity (ID) and class. The length of the

TABLE II Instrument Type Segmentation Results (mIoU) of Different Methods on CaDIS Dataset (10 classes).

Instrument	DeepLabV3+	UPerNet	HRNetV2	TraSeTR
classes	[43]	[44]	[45]	(ours)
Cannula	48.9	50.0	49.5	53.1
Cap. Cystotome	55.7	54.5	61.7	64.4
Tissue Forceps	70.0	74.0	78.0	79.2
Primary Knife	86.1	89.5	89.3	92.3
Ph. Handpiece	75.0	77.6	77.9	76.2
Lens Injector	78.5	81.0	82.8	87.1
I/A Handpiece	74.0	73.6	75.3	71.3
Secondary Knife	69.0	68.2	79.5	82.7
Micromanipulator	59.3	63.6	64.4	66.3
Cap. Forceps	28.9	23.0	27.2	26.1
Total	64.6	65.5	68.6	69.9

bar indicates the time span of the instrument showing in the video. Compared with ground truth, TraSeTR correctly tracks instruments that newly-appear or re-appear in the scenes. More importantly, it assigns the instance ID to each of them, thereby achieving satisfying instance-level tracking.

D. Ablation Studies

1) Bipartite matching vs. *identity matching:* In Table III (a), we verify that the main gains of TraSeTR come from tracking instances with identity matching. We start by comparing (i) TraSeTR uses bipartite matching, and (ii) TraSeTR



Fig. 4. **Visualization of instance-level tracking** on EndoVis17 and CaDIS videos. The color bar represents the time steps that one instrument appearing in the scene. Different color indicates different instrument types. Besides the tracking process, we also show the instrument type segmentation results of intermediate frames. TraSeTR achieves 88.8% and 92.3% tracking accuracy (*i.e.*, how many instances being correctly tracked) on two example videos.

uses identity matching. We report the *mean tracking rate* of two methods at training time, *i.e.*, how many predictions from prior queries can be matched to the current ground truths. Only 27.3% instances can be successfully tracked via bipartite matching, which means that the model cannot always use prior queries for prediction. On the contrary, our identity matching tracks all instances by associating their identities at training time. As a result, it leads to 3.5% mIoU and 2.1% Dice improvements on EndoVis17 dataset, suggesting the necessity of fully exploiting temporal information to discriminate instrument types.

 TABLE III

 Ablation studies of TraSeTR on EndoVis17 Dataset.

(a)	Mean tra	mIoU	Dice		
Bipartite matching		56.9	63.1		
Identity matching		60.4	65.2		
(b)	Query Embeddings Current Prior Contrastive			mIoU	Dice
TraSeTR-NT	 ✓ 			54.6	62.0
TraSeTR-NC	 ✓ 	\checkmark		59.6	64.7
TraSeTR	✓	\checkmark	\checkmark	60.4	65.2

2) Types of Query embeddings: Table III (b) analyzes the different types of query embeddings in TraSeTR. We implement three configurations: (i) TraSeTR-No Tracking (NT): TraSeTR with current queries only, and trained with Hungarian loss \mathcal{L}_{Hung} ; (ii) *TraSeTR-Non Contrastive* (NC): TraSeTR with current and prior queries, and trained with \mathcal{L}_{Hung} ; (iii) TraSeTR: TraSeTR with contrastive queries, and trained with a combination of \mathcal{L}_{Hung} and contrastive loss \mathcal{L}_{ctr} . TraSeTR-NT achieves instance-level instrument segmentation but tends to predict wrong classes for some segments. Adding prior queries alleviates this issue as the temporal information can be explicitly leveraged. The contrastive query embeddings further improve the model's discrimination capability of largely changed instruments, peaking the segmentation results on two metrics. As shown in Fig. 5, contrastive query learning inherently strengthens the encoder-decoder attention mechanism in TraSeTR, such

that the instance attention regions can be precisely found.



Fig. 5. Attention maps of TraSeTR-NC and TraSeTR. We visualize instance embeddings whose indexes are in the matching set σ using the projection algorithm [27]. The green box indicates the largely changed instrument.

IV. CONCLUSION AND FUTURE WORK

This paper presents a novel transformer-based mask classification approach to dynamically track instances in robotic surgical video for accurate semantic segmentation of instruments. Our method addresses the difficulties in this task, *i.e.*, most notably the small inter-class discrepancy and large intra-class variations of instruments, by fully leveraging the set prediction mechanism in the designed transformer to produce per-instance predictions, and a identity matching strategy to incorporate tracking cues. TraSeTR was evaluated on three public datasets, including two RAS datasets and one cataract surgery dataset that contains different instrument types and surgical techniques performed in diverse platforms. TraSeTR outperforms the state-of-the-art performance by up to 5% mIoU and promisingly tracks the positions of instruments entering or leaving the scene. The improvements can facilitate the real-world RAS task automation, such as suturing and dissection, which greatly benefit from instance-level perception. Ablation studies demonstrated the effectiveness of our transformer design and the necessity of contrastive query inductions to tolerate temporal variations. We plan to further investigate the alternative guidance for instancelevel instrument segmentation. One potential direction is to integrate the multi-modal data, e.g., the kinematics data (the position, velocity of the tool tips) from the robotic systems, into the flexible transformer architecture. We will also extend TraSeTR into a unified framework to generate instrument trajectory maps online, which can be applied to downstream scenarios such as robot motion planning in RAS.

REFERENCES

- A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science translational medicine*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.
- [2] A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, and P. Valdastri, "Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study," *IEEE Robotics* and Automation Letters, vol. 5, no. 4, pp. 6528–6535, 2020.
- [3] Y. Qin, M. Allan, Y. Yue, J. W. Burdick, and M. Azizian, "Learning invariant representation of tasks for robust surgical state estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3208–3215, 2021.
- [4] X. Gao, Y. Jin, Z. Zhao, Q. Dou, and P.-A. Heng, "Future frame prediction for robot-assisted surgery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 533–544.
- [5] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Transactions on Medical Imaging*, 2021.
- [6] C. E. Reiley and G. D. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in *International* conference on medical image computing and computer-assisted intervention. Springer, 2009, pp. 435–442.
- [7] A. Zia, A. Hung, I. Essa, and A. Jarc, "Surgical activity recognition in robot-assisted radical prostatectomy using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 273–280.
- [8] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno, and N. Padoy, "A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2938–2945, 2021.
- [9] T. Cheng, W. Li, W. Y. Ng, Y. Huang, J. Li, C. S. H. Ng, P. W. Y. Chiu, and Z. Li, "Deep learning assisted robotic magnetic anchored and guided endoscope for real-time instrument tracking," *IEEE Robotics* and Automation Letters, vol. 6, no. 2, pp. 3979–3986, 2021.
- [10] T. Osa, N. Sugita, and M. Mitsuishi, "Online trajectory planning and force control for automation of surgical tasks," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 675–691, 2017.
- [11] M. Toussaint, J.-S. Ha, and O. S. Oguz, "Co-optimizing robot, environment, and tool design via joint manipulation planning," *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.
- [12] T. D. Nagy and T. Haidegger, "A dvrk-based framework for surgical subtask automation," *Acta Polytechnica Hungarica*, pp. 61–78, 2019.
- [13] B. Lu, B. Li, W. Chen, Y. Jin, Z. Zhao, Q. Dou, P.-A. Heng, and Y. Liu, "Toward image-guided automated suture grasping under complex environments: A learning-enabled and optimization-based holistic framework," *IEEE Transactions on Automation Science and Engineering*, 2021.
- [14] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018, pp. 624–628.
- [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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 440–448.
- [16] Z. Zhao, Y. Jin, X. Gao, Q. Dou, and P.-A. Heng, "Learning motion flows for semi-supervised instrument segmentation from robotic surgical video," in *International Conference on Medical Image Computing* and Computer-Assisted Intervention. Springer, 2020, pp. 679–689.
- [17] Z. Zhao, Y. Jin, B. Lu, C.-F. Ng, Q. Dou, Y.-H. Liu, and P.-A. Heng, "One to many: Adaptive instrument segmentation via meta learning and dynamic online adaptation in robotic surgical video," *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.
- [18] Z. Zhao, Y. Jin, J. Chen, B. Lu, C.-F. Ng, Y.-H. Liu, Q. Dou, and P.-A. Heng, "Anchor-guided online meta adaptation for fast one-shot instrument segmentation from robotic surgical videos," *Medical Image Analysis*, vol. 74, p. 102240, 2021.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Confer-*

ence on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

- [20] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and Ø. Hovde, "Streoscennet: surgical stereo robotic scene segmentation," in *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10951. International Society for Optics and Photonics, 2019, p. 109510P.
- [21] T. Kurmann, P. M. Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *International Conference* on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 505–513.
- [22] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: An instancebased approach for surgical instrument segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2020, pp. 595–605.
- [23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [24] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [25] 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.
- [26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2020.
- [27] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," *arXiv preprint* arXiv:2101.02702, 2021.
- [28] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.
- [29] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," arXiv preprint arXiv:2107.06278, 2021.
- [30] X. Gao, Y. Jin, Y. Long, Q. Dou, and P.-A. Heng, "Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
- [31] S. Kondo, "Lapformer: surgical tool detection in laparoscopic surgical video using transformer architecture," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–6, 2020.
- [32] Y. Long, Z. Li, C. H. Yee, C. F. Ng, R. H. Taylor, M. Unberath, and Q. Dou, "E-dssr: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
- [33] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.
- [34] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen et al., "2018 robotic scene segmentation challenge," arXiv preprint arXiv:2001.11190, 2020.
- [35] M. Grammatikopoulou, E. Flouty, A. Kadkhodamohammadi, G. Quellec, A. Chow, J. Nehme, I. Luengo, and D. Stoyanov, "Cadis: Cataract dataset for surgical rgb-image segmentation," *Medical Image Analysis*, vol. 71, p. 102053, 2021.
- [36] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.
- [37] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

- [38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [39] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018, pp. 1134–1141.
- [40] Z. Zhao, Y. Jin, and P.-A. Heng, "Modelling neighbor relation in joint space-time graph for video correspondence learning," in *Proceedings* of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9960–9969.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer* vision and pattern recognition, 2016, pp. 770–778.
- [42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in ECCV, 2018, pp. 801–818.
- [44] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.
- [45] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.