2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Ada-Tracker: Soft Tissue Tracking via Inter-Frame and Adaptive-template Matching

Jiaxin Guo[1], *Student Member, IEEE*, Jiangliu Wang[1], *Member, IEEE*, Zhaoshuo Li[2], Tongyu Jia[3]
Qi Dou[4,5], *Member, IEEE*, and Yun-Hui Liu[1,5], *Fellow, IEEE*

***Abstract*— Soft tissue tracking is crucial for computer-assisted interventions. Existing approaches mainly rely on extracting discriminative features from the template and videos to recover corresponding matches. However, it is difficult to adopt these techniques in surgical scenes, where tissues are changing in shape and appearance throughout the surgery. To address this problem, we exploit optical flow to naturally capture the pixel-wise tissue deformations and adaptively correct the tracked template. Specifically, we first implement an inter-frame matching mechanism to extract a coarse region of interest based on optical flow from consecutive frames. To accommodate appearance change and alleviate drift, we then propose an adaptive-template matching method, which updates the tracked template based on the reliability of the estimates. Our approach, Ada-Tracker, enjoys both short-term dynamics modeling by capturing local deformations and long-term dynamics modeling by introducing global temporal compensation. We evaluate our approach on the public SurgT benchmark, which is generated from Hamlyn, SCARED, and Kidney boundary datasets. The experimental results show that Ada-Tracker achieves superior accuracy and performs more robustly against prior works. Code is available at `https://github.com/wrld/Ada-Tracker`.**

I. INTRODUCTION

Soft tissue tracking is an essential task in computer-assisted interventions, benefiting various downstream applications, including force estimation [1], motion compensation [2], image-guided surgery [3], tissue scanning [4], [5], autonomous tissue manipulation [6], 3D Reconstruction [7], [8], etc. Tissue tracking also enhances tissue deformation estimation but also manipulation and interaction within surgical spaces [9].

Typically, the soft tissue tracking technique involves selecting the target region of the tissue in the initial frame and then tracking the movement in subsequent frames of a surgical video sequence. Previous traditional methods employ rigid assumptions [10], [11], [7], classical descriptors [12], or model-based techniques [13], [14], [15] to recover complex deformations on tissue surfaces. Recent studies [16], [17], [18] have revealed an emerging trend that utilizes data-driven techniques, which can be roughly divided into two categories: feature-based patch tracking and flow-based point tracking. One major line of research is to consider the region of interest as an object and apply general visual object tracking (VOT) methods for tissue tracking [19], [20], [21], [22]. Such deep VOT trackers adopt a Siamese-based network [23] that independently extracts appearance features of the template and search region, then matches them via cross-correlation to determine the position with maximum similarity in the response map. However, surgical scenes possess a homogeneous appearance with texture-less soft tissue, making it difficult to correlate the template and search region to find the correspondence based on their appearance. To tackle this problem, recent works attempt to estimate optical flow and capture the movements of individual points between neighboring frames for tissue tracking [24], [25], [17]. Without relying on the appearance of objects during training, such flow-based methods offer better generalization

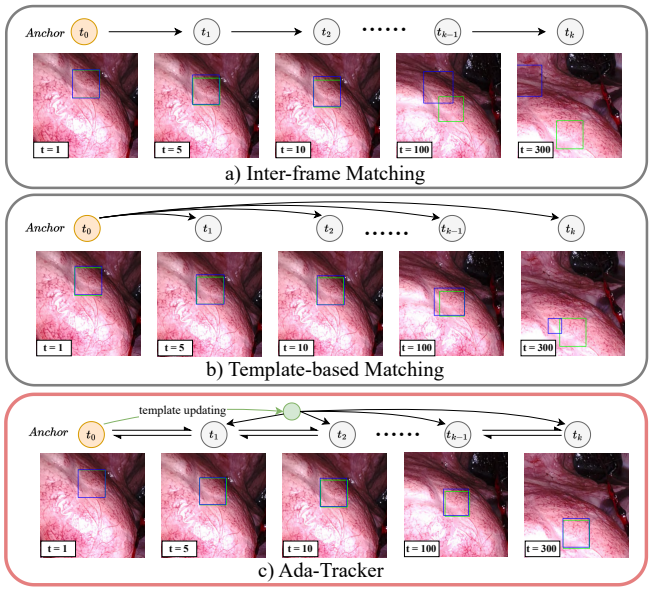

Fig. 1. **Different approaches to employ optical flow for soft tissue tracking,** with blue and green bounding box indicating the predicted result and ground truth. a) Inter-frame matching updates the target box based on optical flow from consecutive frames, prone to drift due to accumulated error. b) Template-based matching predicts the target box by matching the template image from the anchor box with every input frame, struggling with appearance variation and large motion. c) Our Ada-Tracker exploits the benefits of a) and b) for the more robust and accurate tracker.

This work is supported in part by Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part by the Research Grants Council of Hong Kong under Grant T42-409/18-R, Grant 14218322, and Grant 14207320, in part by the Hong Kong Centre for Logistics Robotics, in part by the Multi-Scale Medical Robotics Centre, InnoHK, and in part by the VC Fund 4930745 of the CUHK T Stone Robotics Institute. (*Corresponding author: Yun-Hui Liu*)

[1]CUHK T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong.
[2]Johns Hopkins University, United States.
[3]Faculty of Urology, Third Medical Center, Chinese PLA General Hospital, Beijing, China.
[4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.
[5]Hong Kong Center for Logistics Robotics, Hong Kong.

to unseen deformable patterns, avoiding the challenges of adapting to different surgery domains. However, as they focus on estimating the motion of sparse points individually, such methods lack contextual information and are susceptible to noise and large deformations.

In this work, we argue that estimating the patch-level optical flow provides a comprehensive view of tissue deformation, enabling stable motion estimation with enhanced robustness. Tissue tracking based on patch-level optical flow could be more effective in modeling complex and non-rigid deformations, making it natural to handle complex deformations and occlusions. It motivates us to answer the following question: *How to exploit patch-level optical flow for soft tissue tracking?* Two natural strategies to adapt patch-level optical flow for template-based soft tissue tracking are *inter-frame matching* and *template-based matching*. As shown in Fig. 1, the first approach updates the target box by computing optical flow between consecutive frames. While this approach excels in achieving precise and stable motion estimation for short-term video sequences, it is prone to accumulate errors during long-term tracking, which can result in issues such as drifting, divergence, and sensitivity to occlusion. The template-based matching predicts the target region by taking the template image as the reference, to match with every frame. But this method struggles with challenges posed by variations in appearance and large movement, including fluctuations in illumination, camera motion, and deformations in soft tissue.

Based on the above analysis, we then propose Ada-Tracker for patch-level, long-term soft tissue tracking that exploits the strengths of both inter-frame motion estimation and template-based refinement, correlated with a designed adaptive template updating scheme. Ada-Tracker provides a comprehensive solution to soft tissue tracking by leveraging optical flow to capture both real-time tissue dynamics and exploit template-based correspondences. Specifically, our method consists of two stages: i) the **Inter-frame matching** stage estimates optical flow between consecutive frames, capturing the immediate motion of soft tissues to extract a coarse region of interest (ROI). We evaluate the confidence of the flow estimate to down-weigh outliers and check if occlusion is present. ii) the **Adaptive-template matching** stage then adaptively updates the template to reflect appearance changes while rectifying potential inaccuracies and drift from the initial inter-frame estimates. Ada-Tracker is capable of capturing the deformations of soft tissues with robustness and accuracy, countering surgical tracking challenges such as occlusions, drift, and appearance variations. Our contributions are summarized as threefold:

1) We present a novel approach for soft tissue tracking by capitalizing on optical flow, offering a comprehensive solution that bridges optical flow with soft tissue tracking in surgical contexts.
2) Our method harnesses the strengths of both inter-frame and adaptive-template matching, which estimates the soft tissue movement effectively, catering to both short-term changes and long-term trends.
3) We perform thorough experiments to validate our method on the SurgT dataset, outperforming the previous SOTA trackers in terms of accuracy, and robustness in both 2D and 3D tissue tracking.

II. RELATED WORKS

Visual Object Tracking. Previous studies utilize CNN backbones [26], [27] to extract features and fuse them through lightweight relation modeling networks, such as Siamese [23], [28], [29] and discriminative trackers [30], [31]. However, their performance is restricted due to one-way information interaction. In response, methods like TransT [32] and STARK [33] incorporate Transformers to allow bi-directional relation modeling, enhancing accuracy but at the cost of slower inference. By considering the region of interest as an object, some recent studies adopt the general VOT to address data-driven soft tissue tracking problem [9]. According to the report of the SurgT [9], however, the low scores of TransT [32] indicate the difficulty of the generalization of the supervised learning-based method to the surgical scene. The VOT-based unsupervised tracking methods also struggle to find correspondence between the template and search region of homogeneous appearance, yielding lower performance compared to the traditional method.

Soft Tissue Tracking. Soft tissue tracking is a non-rigid tracking method, that meets the challenges of texture-less, deformable tissue. Traditional works [10], [11], [7] track the surgical scene with rigid assumptions, which fails when encountering large motions. MIS-SLAM [13] leverages deform nodes to model the surface deformations, and models like DefSLAM [34] combine meshes with classical features. To model the deformations, some work [14], [15] leverage spline and mesh to describe the deformations. Recent studies [16], [17], [18] have revealed an emerging trend that utilizes data-driven techniques using convolutional neural networks (CNN) or graph convolutional networks (GCN). However, a significant limitation of these efforts is their dependence on sparse key points or descriptors designed for specific datasets or applications. This restriction limits their usefulness in real-world dynamic surgical scenarios, which involve challenging situations such as instrument obstructions, changes in lighting, and significant distortions of soft tissue.

III. METHODOLOGY

The overview of Ada-Tracker is illustrated in Fig. 2. Ada-Tracker aims to locate the ROI of soft tissues in the form of bounding boxes at every frame, which is predefined at the start of a surgical video sequence. Our method follows a coarse-to-fine structure, consisting of two stages: 1) **Inter-frame matching** to capture the real-time dynamics of soft tissue movements, providing a coarse prediction for initialization (Sec. III-A), 2) **Adaptive-template matching** to first update the template image dynamically, and then refine the coarse prediction from the first stage. (Sec. III-B)). We lastly introduce our self-supervised training and losses in Sec. III-C.

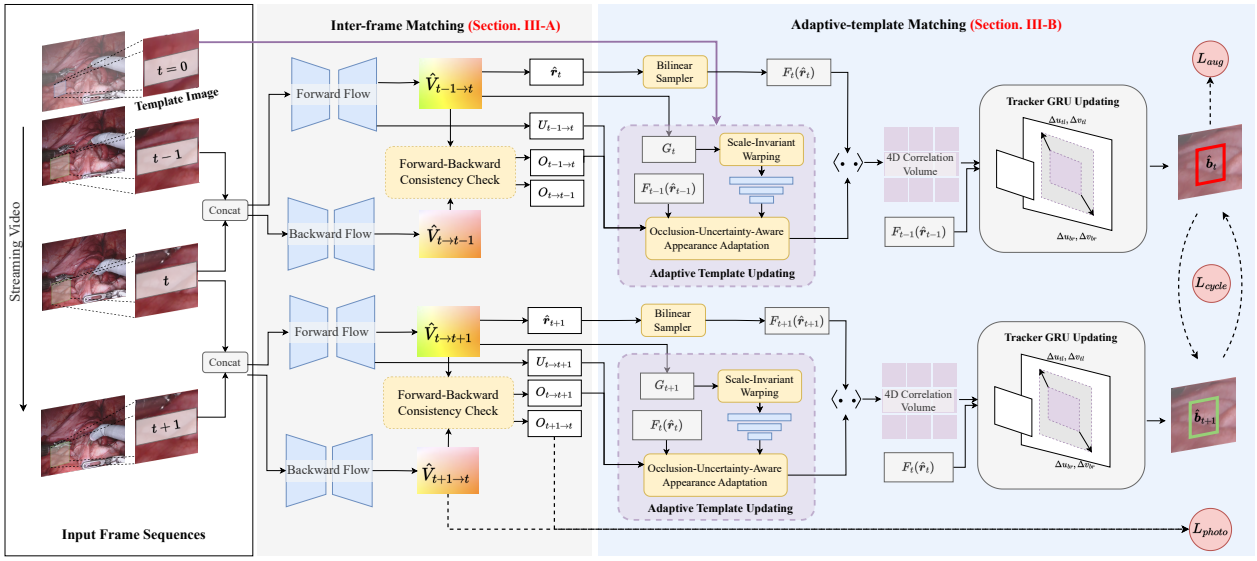

Fig. 2. **Overview of our proposed Ada-Tracker.** Given a bounding box at the start of the surgical video, we aim to locate the ROI of soft tissues in the form of bounding boxes at every frame. Our method consists of two stages: inter-frame matching and adaptive-template matching. In the first stage, we capture the immediate motion, and obtain a coarse bounding box prediction. Next, we update the template adaptively based on the flow, confidence and occlusion information from the previous stage. We finally match the updated template and coarse ROI to obtain the final prediction.

A. *Inter-frame Matching*

In the inter-frame matching stage, we aim to capture the immediate motion and obtain a coarse bounding box prediction of the tracked tissue. At the start of the surgical video, a bounding box $\mathbf{b}_0$ at I_0 is assumed given, and the template patch $P_0 \in \mathbb{R}^{H_0 \times W_0 \times 3}$ within $\mathbf{b}_0$ is extracted. At each time t, we initialize the search region in I_t to be $P_t \in \mathbb{R}^{4H_{t-1} \times 4W_{t-1} \times 3}$, which is expanded from the previous bounding box result $\hat{\mathbf{b}}_{t-1}$. The tracked patch $P_{t-1} \in \mathbb{R}^{4H_{t-1} \times 4W_{t-1} \times 3}$ from the previous image I_{t-1} is also extracted by expanding $\hat{\mathbf{b}}_{t-1}$. Both P_t and P_{t-1} are used to estimate a coarse bounding box at time t.

Uncertainty-aware Optical Flow Estimation We leverage the RAFT model [35] for our optical flow estimation to capture pixel-wise soft-tissue movement between consecutive frames P_{t-1} and P_t. Ada-Tracker begins with a feature encoder that transforms the input image pair P_{t-1} and P_t into a lower dimensional representation $F_{t-1}, F_t \in \mathbb{R}^{\frac{H_{t-1}}{8} \times \frac{W_{t-1}}{8} \times C}$ with $C = 256$ to preserve the motion-critical details. P_{t-1} is processed by the context encoder to provide spatial information during updating following [35]. Subsequently, a 4D correlation volume is constructed, capturing the similarity between every pixel pair. A GRU-based updater is employed to iteratively refine optical flow predictions by exploiting both spatial and temporal contexts.

To address the inherent uncertainty in motion estimation, we introduce an additional CNN layer to generate a confidence map $U_{t-1 \to t}$ from the output of the GRU-based updater by a sigmoid activation. The inter-frame optical flow and its associated confidence are jointly computed as:

$$\hat{V}_{t-1 \to t}, U_{t-1 \to t} = \mathcal{F}_\theta(I_{t-1}, I_t), \tag{1}$$

where $\hat{V}_{t-1 \to t}$ denotes the estimated optical flow between the origin image patch I_{t-1} and the search image patch I_t. The function $\mathcal{F}_\theta$ refers to the inter-frame optical flow model parameterized by weights θ. The confidence map $U_{t-1 \to t}$ quantifies the reliability of the predicted flows, highlighting regions with strong feature correspondences and thereby indicating areas of high stability.

Coarse ROI Prediction. Based on the inter-frame optical flow $\hat{V}_{t-1 \to t}$, we update the coarse ROI $\hat{\mathbf{r}}_t$ from the previous coarse ROI $\hat{\mathbf{r}}_{t-1}$ which is double size expanded from the previous target bounding box $\hat{\mathbf{b}}_{t-1}$. We combine the origin coordinates in $\hat{\mathbf{r}}_{t-1}$ with the predicted flow $\hat{V}_{t-1 \to t}(\hat{\mathbf{r}}_{t-1})$ as the predicted coordinates in P_t. Then $\hat{\mathbf{r}}_t$ is generated with a Min-Max strategy to extract the minimal enclosing rectangle of the predicted coordinates in P_t.

Occlusion Map. To deal with the occlusion from instruments, camera movement, and tissue deformation, we apply the forward-backward consistency check following [36]. Specifically, we calculate the forward flow $\hat{V}_{t-1 \to t}$ and backward flow $\hat{V}_{t \to t-1}$, then generate the occlusion map $O_{t-1 \to t}$ and $O_{t \to t-1}$. The occlusion map helps in identifying the occluded areas in the tracking region, serving as a mask to filter out the noisy feature with occlusion in the adaptive template updating. During online tracking, we calculate the occlusion percentage $\text{occ}_{t-1 \to t}$ in $O_{t-1 \to t}(\hat{\mathbf{b}}_{t-1})$. No prediction will be made if occlusion percentage $\text{occ}_{t-1 \to t} > \beta$, where β is an occlusion threshold.

B. *Adaptive-template Matching*

During long-term tissue tracking, matching a static template image with the search region is prone to fail when encountering drastic appearance change, deformation, and occlusion. To avoid the problem, the adaptive-template matching is introduced to dynamicly update the template image P_0 to improve the robustness under motion and appearance change. We update the template adaptively based on the flow, confidence, and occlusion information from the

previous stage. Next, we devise an anchor-based matching network to find the best match of the updated template in the coarse ROI $\hat{\mathbf{r}}_t$. This anchor-based matching network follows a RAFT-like structure but focuses on patch-level updating instead of pixel-wise motion estimation, which is robust to noises and highly efficient.

Adaptive Template Updating. To aggregate the template image with spatial and temporal information from inter-frame dynamics, we enhance the anchor patch compensating both motion and appearance variation, from multi-aspect: the accumulated flow, occlusion map, and uncertainty map.

For motion updating, we maintain the accumulated flow G_t, which contains the pixel-wise motion variation of the coarse target region from the start til now, e.g. camera shift, scale-in, scale-out, and soft tissue deformation. To compensate for the motion variation, we design a scale-invariant warping scheme with the grid of the coarse ROI $\mathbf{x}_0$, the center of the coarse ROI box $\mathbf{c}_0$, the calculated scale ratio S_t:

$$\begin{aligned} G_t &= \text{var}(G_{t-1} + V_{t-1\to t}), \\ S_t &= \text{mean}(\sqrt{\frac{\| -G_t + \mathbf{x}_0 - \mathbf{c}_0 \|^2}{\| \mathbf{x}_0 - \mathbf{c}_0 \|^2}}), \\ P_{w,t} &= P_0((-G_t + \mathbf{x}_0 - \mathbf{c}_0) \cdot S_t + \mathbf{c}_0), \end{aligned} \tag{2}$$

where $P_{w,t}$ denotes the warped template image from the template image P_0. By using this scale-invariant warping approach, we could obtain the warped template image $P_{w,t}$ with inherent preservation of intricate details, such as rotations and tissue deformations, whilst simultaneously filtering out extreme alterations or positional shifts from camera-like scale-ins, scale-outs, or shifts. Then we extract warped template feature $F_{w,t}$ from $P_{w,t}$ by the feature encoder.

To address the appearance variation during the long-term surgical tracking, we resize the confidence map $U_{t-1\to t}$ and occlusion map $O_{t-1\to t}$ to match the size of F_{t-1}. Then $U_{t-1\to t}$ and $O_{t-1\to t}$ are leveraged to determine the weight for $F_{w,t}$ to adapt the appearance by fusing the target feature of previous step $F_{t-1}(\hat{\mathbf{r}}_{t-1})$. Our designed appearance adaptation is designed as follows:

$$\begin{aligned} F_{\text{occ}} &= (1 - O_{t\to t-1}(\hat{\mathbf{r}}_{t-1})) \odot F_{t-1}(\hat{\mathbf{r}}_{t-1}), \\ F_{\text{conf}} &= U_{t-1\to t}(\hat{\mathbf{r}}_{t-1}) \odot F_{\text{occ}}, \\ F_{u,t} &= \alpha F_{w,t} + (1-\alpha) F_{\text{conf}}, \end{aligned} \tag{3}$$

where F_{occ} and F_{conf} denote the feature masked with the occlusion map, and the feature with both uncertainty-awareness and occlusion-awareness, $F_{u,t}$ refers to the final updated template feature. Therefore, we ensure that occluded or boundary regions do not overly influence the fused feature representation.

Template-based Matching. In this step, we aim to conduct an anchor-based matching between the updated template feature $F_{u,t} \in \mathbb{R}^{\frac{H_0}{8}\times\frac{W_0}{8}\times C}$ and the coarse ROI sampled search image feature $F_t(\hat{\mathbf{r}}_t) \in \mathbb{R}^{\frac{H_{t-1}}{8}\times\frac{W_{t-1}}{8}\times C}$. Specifically, to correlate the template and search region, we build a 4D correlation cost volume $\mathbf{C} \in \mathbb{R}^{\frac{H_0}{8}\times\frac{W_0}{8}\times\frac{H_{t-1}}{8}\times\frac{W_{t-1}}{8}}$. We take the center half-size bounding box of $\hat{\mathbf{r}}_t$ as the initial index for the template feature to lookup the correlation volume, which enables a faster matching process based on the coarse initialization. Finally, we take $F_{t-1}(\hat{\mathbf{r}}_{t-1})$ as the context feature to serve the network with more spatial context around the template region.

Different from the pixel-wise motion updating in RAFT, our anchor-based GRU updates the region-wise motion to find the bounding box that best matches the search region. Specifically, our method only outputs the flow of the left-top corner and right-bottom corner to update the predicted bounding box. Then the updated flow is calculated as the interpolation of the corner flow:

$$\hat{V}_z^{(i+1)} = \hat{V}_z^{(i)} + \text{Interp}(\Delta V_z^{(i)}(\mathbf{x}_{tl}), \Delta V_z^{(i)}(\mathbf{x}_{br})), \tag{4}$$

where $\hat{V}_z^{(i)}$ refers to the predicted flow at iter i, $\Delta V_z^{(i)}(\mathbf{x}_{tl})$ and $\Delta V_z^{i}(\mathbf{x}_{br})$ denote the delta flow of the left top corner and bottom right corner. With the iterative updating of the template coordinates, we aim to match the template feature with the search region feature. Then the final predicted bounding box is calculated from the left top corner $\Delta V_z(\mathbf{x}_{tl})$ and right bottom corner $\Delta V_z(\mathbf{x}_{br})$ from the anchor-based matching model $\mathcal{F}_\phi$:

$$\begin{aligned} \hat{V}_z, U_z &= \mathcal{F}_\phi(F_{u,t}, F_t(\hat{\mathbf{r}}_t), F_{t-1}(\hat{\mathbf{r}}_{t-1}), \\ \hat{\mathbf{b}}_t &= \text{bbox}(\hat{V}_z(\mathbf{x}_{tl}), \hat{V}_z(\mathbf{x}_{br})), \end{aligned} \tag{5}$$

where $\hat{V}_z$ and U_z indicate the tracking flow between the warped template image $F_{u,t}$ and the coarse search region $F_t(\hat{\mathbf{r}}_t)$. The designed anchor-based matching realizes a coarse-to-fine prediction of the tracking target, enabling improved tracking accuracy and robustness.

C. Training and Losses

Cycle Consistency Loss. As shown in Fig. 2, after predicting the target bounding box $\hat{\mathbf{b}}_{t-1\to t}$ from $t-1$ to t, we crop the I_{t+1} with four times expanded bounding box of the predicted result as the search region to track the anchor from t to $t+1$. After this forward-tracking procedure, we take $\hat{\mathbf{b}}_{t\to t+1}$ to start a backward-tracking from $t+1$ to $t-1$ to obtain the backward predicted result $\hat{\mathbf{b}}_{cycle}$. After this cycle tracking, we could take the original anchor box $\mathbf{b}$ as a pseudo label to supervise $\hat{\mathbf{b}}_{cycle}$ by a linear combination of GIoU loss [37] and l1 norm loss. Additionally, we employ a cycle template loss to penalize the reconstruction loss between the warped template image $\hat{P}_{cycle}$ after the cycle and the template image P. Then our cycle loss is represented as follows:

$$\begin{aligned} \mathcal{L}_{cycle} = \mathcal{L}_{GIoU}(\hat{\mathbf{b}}_{cycle}, \mathbf{b}) + \mathcal{L}_1(\hat{\mathbf{b}}_{cycle}, \mathbf{b}) \\ + \mathcal{L}_{recon}(\hat{P}_{cycle}, P). \end{aligned} \tag{6}$$

Total Loss. To improve the self-supervise performance, we adopt an occlusion-aware photometric loss $\mathcal{L}_{photo}$ [38] to train the inter-frame matching network, leveraging the occlusion map $O_{t-1\to t}$ and $O_{t\to t-1}$. An augmentation loss $\mathcal{L}_{aug}$ is empolyed to supervise the tracking results with the pseudo ground truth label. A smooth loss $\mathcal{L}_{smooth}$ [39]

TABLE I

COMPARISONS AGAINST PRIOR WORKS IN THE SURGT CHALLENGE ON TEST SET

Method	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$EAO\uparrow$
CSRT	0.872	0.769	5.7±2.6	0.894	3.3±3.8	0.563
Jmees	0.868	0.818	5.3±2.4	0.878	2.7±1.9	0.583
KIT	0.465	0.747	12.5±8.0	0.810	11.9±9.5	0.223
TransT	0.701	0.529	16.3±5.1	0.861	17.3±26.7	0.274
SRV	0.476	0.681	15.4±7.5	0.710	8.5±13.7	0.293
MEDCVR	0.702	0.509	7.9±5.8	0.832	9.2±39.7	0.302
ETRI	0.802	0.693	12.1±7.4	0.909	5.7±5.2	0.405
RIWOlink	0.807	0.737	8.0±4.5	0.894	5.8±9.2	0.433
ICVS-2Ai	0.872	0.816	6.7±3.5	0.901	2.6±2.3	0.573
Ours	**0.894**	**0.833**	**5.2±3.3**	**0.912**	**2.1±2.0**	**0.591**

is adopted to regularize the optical flow. Eventually, our training loss $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cycle} + \lambda_2 \mathcal{L}_{aug} + \lambda_3 \mathcal{L}_{photo} + \lambda_4 \mathcal{L}_{smooth}, \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are loss weights.

IV. EXPERIMENTS

A. Datasets

We evaluate our method in the recent proposed public SurgT dataset [9]. It provides the first standardized benchmark for assessing soft tissue tracking approaches and is generated from three datasets: Hamlyn [40], SCARED [41], and Kidney boundary datasets [42]. The SurgT dataset is composed of 157 stereo endoscopic videos with calibration parameters, including 125 videos from 12 cases for training, 12 videos from 3 cases for validation, and 20 videos from 5 cases for testing. Only the validation and testing datasets are annotated with the bounding boxes. Since no annotation is available for training, we randomly create the bounding box to train the Ada-tracker in a self-supervised manner.

B. Implementation Details

All experiments are implemented using PyTorch, and conducted on an NVIDIA RTX 3090 GPU. During training, we apply the AdamW optimizer with the learning rate as 0.000125 and the weight decay as 0.00001. For loss weights, we set $\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.1, \lambda_4 = 0.001$. For training data, we collect video sequences from the training datasets in a random range, and random crop the 256×256 size images as input. We add shift, rotation, and illumination change to the image for augmentation and obtain the augmented bounding boxes as pseudo ground truth for $\mathcal{L}_{aug}$. Following SurgT [9], we use both monocular metrics and stereo metrics to evaluate our method, including 2D/3D Accuracy, 2D/3D Error, 2D/2D Robustness, and Expected Average Overlap (EAO).

C. Comparison with State-of-the-art

We compare Ada-Tracker with 7 methods submitted to the SurgT challenge, as well as CSRT [43], TransT [32] as two additional baselines. These methods can be grouped into three categories: i) traditional discriminative correlation filters-based tracker (CSRT [43], Jmees), ii) feature-based patch tracking method (TransT [32], ETRI, MEDCVR, SRV, KIT, RIWOlink), and iii) optical flow based point tracking method (ICVS-2Ai).

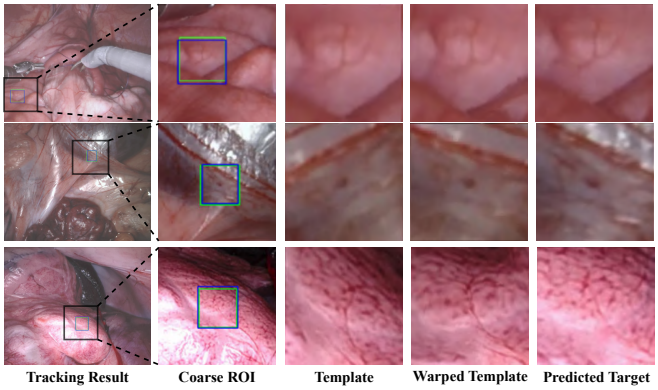

Fig. 3. **Qualitative results**. We visualize the tracking in different cases, including tissue deformation, camera movement, and illumination variations.

As shown in Tab. I, Ada-Tracker performs favorably against prior work in all metrics on case 1-5. The methods in the first category have achieved the most competitive performance, with Jmees achieving the best performance during the challenge. However, Jmees requires stereo frames to calculate the disparity information of consecutive frames to determine the size of the updated bounding boxes. They also require the pre-trained segmentation network to leverage prior tool information to determine the tracking occlusion. VOT-based methods leverage the generalization of models trained on the general computer vision datasets, which differ drastically from surgical scenes due to domain gaps. Regarding the optical flow-based ICVS-2Ai, while optical flow for point tracking improves the performance against appearance variation, their sensitivity to noises makes ICVS-Ai suffer from sudden scale change and occlusions during long-term tracking, as evidenced by the low robustness in Case 5 shown in Tab. II and Tab. III. Without patch-level optic flow to help determine the patch boundary, ICVS-Ai relies on left and right disparity to determine the predicted bounding box scale, which can affect the robustness in Tab. III.

On the contrary, by combining the benefits of inter-frame and adaptive-template matching, our Ada-Tracker addresses the limitations of both methods and performs more robustly than previous work. Our patch-level soft-tissue tracking tracks the target bounding box by estimating a holistic deformation of the template patch, overcoming surgical tracking challenges such as large tissue deformation (Case 3, 4), and random camera movements (Case 1, 2, 5), illumination and reflective surfaces (Case 1, 5). Without the use of stereo disparity information, our method is still capable of achieving competitive performance by utilizing scale-invariant warping in adaptive template updating. This maintains a constant scale reference point to assist in determining the target box boundary. As shown in Fig. 3, while the current frame encounters camera rotation and illumination variation, our Ada-Tracker still achieves stable and accurate tracking, with the warped template remaining in the same motion state as the predicted target. More tracking visualization is provided in our supplementary video.

TABLE II

COMPARISONS AGAINST PRIOR WORK IN THE SURGT CHALLENGE ON TEST 2D CASES

Method	Case 1			Case 2			Case 3			Case 4			Case 5		
	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$
CSRT	0.92	0.72	**7±3**	0.87	0.77	8±3	0.87	0.81	5±2	**1.0**	0.85	**3±1**	0.70	0.68	6±3
Jmees	0.92	0.76	**7±3**	**0.89**	0.77	**7±3**	0.88	**0.88**	5±2	0.94	**0.90**	**3±1**	0.70	0.75	5±3
KIT	0.26	0.61	19±13	0.17	0.61	23±17	0.81	0.84	9±7	0.84	0.71	11±6	0.16	0.58	20±10
TransT	0.72	0.46	18±4	0.64	0.42	24±10	0.82	0.61	16±4	0.79	0.61	11±7	0.47	0.48	11±4
SRV	0.39	0.49	27±13	0.19	0.55	16±11	0.71	0.80	11±6	0.98	0.69	13±5	0.07	0.56	22±12
MEDCVR	0.74	0.54	7±5	0.65	0.48	16±12	0.82	0.68	4±4	0.77	0.61	10±7	0.46	0.57	6±4
ETRI	0.88	0.56	21±13	0.82	0.69	14±10	0.85	0.82	7±3	0.94	0.73	7±4	0.49	0.60	10±6
RIWOlink	0.89	0.66	9±5	0.71	0.78	9±6	0.85	0.86	**3±1**	0.99	0.69	10±5	0.55	0.59	9±6
ICVS-2Ai	0.94	0.76	9±4	0.84	0.79	9±6	**0.91**	**0.88**	5±3	**1.0**	**0.90**	**3±1**	0.65	0.72	8±5
Ours	**0.95**	**0.79**	7±5	0.88	**0.82**	7±6	**0.91**	**0.88**	4±3	**1.0**	0.88	**3±1**	**0.72**	**0.78**	**4±3**

TABLE III

COMPARISONS AGAINST PRIOR WORK IN THE SURGT CHALLENGE ON TEST 3D CASES

Method	Case 1		Case 2		Case 3		Case 4		Case 5	
	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$
CSRT	0.91	5±4	0.83	2±1	0.93	**1±1**	**1.0**	**1±1**	**0.78**	9±16
Jmees	**1.0**	**2±1**	0.85	2±1	0.93	**1±1**	0.95	**1±1**	0.72	5±4
KIT	0.68	20±16	0.70	23±14	**0.97**	2±2	**1.0**	6±5	0.68	20±19
TransT	0.91	23±8	0.82	24±53	0.92	6±14	0.92	10±56	0.70	31±26
SRV	0.66	16±13	0.37	16±12	0.96	3±3	**1.0**	2±1	0.46	18±66
MEDCVR	0.90	8±22	0.75	20±125	0.88	2±9	0.92	10±62	0.67	14±24
ETRI	0.97	6±4	**0.90**	7±6	0.94	2±3	**1.0**	3±3	0.71	13±13
RIWOlink	0.96	7±4	0.81	6±5	0.93	**1±1**	0.99	5±34	0.72	11±7
ICVS-2Ai	0.98	4±2	0.84	**1±1**	0.94	1±2	**1.0**	**1±1**	0.71	6±6
Ours	0.99	2±2	0.84	2±1	0.96	**1±1**	**1.0**	**1±1**	0.76	**4±3**

TABLE IV

ABLATION OF ADA-TRACKER ON VALIDATION SET. "I":INTER-FRAME MATCHING ONLY, "T": TEMPLATE-BASED MATCHING ONLY, "A": ADAPTIVE TEMPLATE UPDATING.

I	T	A	$Rob_{2D}\uparrow$	$Acc_{2D}\uparrow$	$Err_{2D}\downarrow$	$Rob_{3D}\uparrow$	$Err_{3D}\downarrow$	$EAO\uparrow$
✓			0.323	0.512	14.2±5.4	0.587	16.9±13.5	0.088
	✓		0.541	0.627	8.9±7.3	0.620	9.8±19.7	0.164
	✓	✓	0.643	0.639	7.2±6.5	0.693	7.9±17.4	0.279
✓	✓	✓	**0.740**	**0.845**	**4.2±5.6**	**0.783**	**2.8±2.5**	**0.318**

D. Ablation Study

We evaluate the effectiveness of different components in our proposed method as shown in Tab. IV.

a) Inter-frame Matching: It updates the target box in the next frame based on the previous prediction and the estimated flow, resulting lowest robustness and accuracy in both 2D and 3D tracking. As shown in Fig. 3, when encountering large movements, the trajectory will get noise accumulation and eventually get a divergent trajectory.

b) Template-based Matching: Template-based matching estimates the target box based on finding the correspondences between the template image and the current frame, which improves the accuracy and robustness by setting a reference. It is shown in Fig. 3 that in the first case, the trajectory of the template-based tracker could closely follow the ground truth trajectory most of the time. However, it eventually fails when the template appearance has a large difference with the search region.

c) Adaptive Template Updating: After adding the template updating scheme, we could improve the template image to make it consistent with the search region. As shown in Fig. 3, the warped template realizes consistent deformation and motion with the predicted target, making it easier to find correspondence for matching. The trajectory in Fig. 4 shows that our trajectory closely matches the ground truth, verifying the effectiveness of our method.

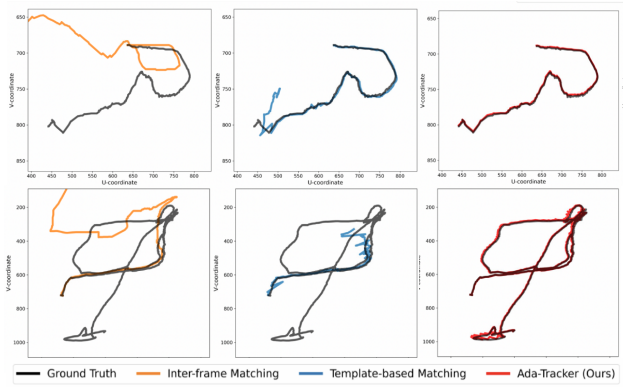

Fig. 4. **Qualitative comparison** of different approaches in both short-term (first row) and long-term (second row) tracking case.

V. CONCLUSIONS

In this paper, we propose Ada-Tracker to enable patch-level optical flow in soft tissue tracking, combining the strengths of both inter-frame and dynamic-template matching. Our method could identify movement from frame to frame and refine its accuracy through anchor-based modifications. Our Ada-Tracker achieves high robustness and accuracy encountering surgical obstacles like unexpected tissue movements or appearance alterations. Our work has the great potential to be applied to computer-assisted interventions, helping make surgeries safer and more streamlined.

REFERENCES

[1] S. Giannarou, M. Ye, G. Gras, K. Leibrandt, H. J. Marcus, and G.-Z. Yang, "Vision-based deformation recovery for intraoperative force estimation of tool–tissue interaction for neurosurgery," *International journal of computer assisted radiology and surgery*, vol. 11, pp. 929–936, 2016.

[2] R. Richa, A. P. Bó, and P. Poignet, "Towards robust 3d visual tracking for motion compensation in beating heart surgery," *Medical Image Analysis*, vol. 15, no. 3, pp. 302–315, 2011.

[3] M. C. Yip, D. G. Lowe, S. E. Salcudean, R. N. Rohling, and C. Y. Nguan, "Tissue tracking and registration for image-guided surgery," *IEEE transactions on medical imaging*, vol. 31, no. 11, pp. 2169–2182, 2012.

[4] C. Wang, J. Cartucho, D. Elson, A. Darzi, and S. Giannarou, "Towards autonomous control of surgical instruments using adaptive-fusion tracking and robot self-calibration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2395–2401.

[5] J. Zhan, J. Cartucho, and S. Giannarou, "Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 11 147–11 154.

[6] Z. Wang, X. Li, D. Navarro-Alarcon, and Y.-h. Liu, "A unified controller for region-reaching and deforming of soft objects," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 472–478.

[7] A. Marmol, A. Banach, and T. Peynot, "Dense-arthroslam: Dense intra-articular 3-d reconstruction with robust localization prior for arthroscopy," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 918–925, 2019.

[8] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE transactions on medical imaging*, vol. 38, no. 1, pp. 79–89, 2018.

[9] J. Cartucho, A. Weld, S. Tukra, H. Xu, H. Matsuzaki, T. Ishikawa, M. Kwon, Y. Jang, K.-J. Kim, G. Lee, *et al.*, "Surgt: Soft-tissue tracking for robotic surgery, benchmark and challenge," *arXiv preprint arXiv:2302.03022*, 2023.

[10] O. G. Grasa, J. Civera, and J. Montiel, "Ekf monocular slam with relocalization for laparoscopic sequences," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4816–4821.

[11] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE transactions on medical imaging*, vol. 33, no. 1, pp. 135–146, 2013.

[12] Y. Li, F. Richter, J. Lu, E. K. Funk, R. K. Orosco, J. Zhu, and M. C. Yip, "Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2294–2301, 2020.

[13] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4068–4075, 2018.

[14] W.-K. Wong, B. Yang, C. Liu, and P. Poignet, "A quasi-spherical triangle-based approach for efficient 3-d soft-tissue motion tracking," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 5, pp. 1472–1484, 2012.

[15] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. E. Bowden, "3d reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical optics express*, vol. 8, no. 4, pp. 2106–2123, 2017.

[16] A. Schmidt, O. Mohareri, S. DiMaio, and S. E. Salcudean, "Fast graph refinement and implicit neural representation for tissue tracking," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1281–1288.

[17] Schmidt, Mohareri, DiMaio, and Salcudean, "Recurrent implicit neural graph for deformable tracking in endoscopic videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 478–488.

[18] S. Lin, A. J. Miao, J. Lu, S. Yu, Z.-Y. Chiu, F. Richter, and M. C. Yip, "Semantic-super: A semantic-aware surgical perception framework for endoscopic tissue identification, reconstruction, and tracking," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4739–4746.

[19] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2137–2155, 2016.

[20] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 300–317.

[21] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. ˇCehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey, *et al.*, "The sixth visual object tracking vot2018 challenge results," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[22] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[23] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 850–865.

[24] M. Neoral, J. Šerỳch, and J. Matas, "Mft: Long-term tracking of every pixel," *arXiv preprint arXiv:2305.12998*, 2023.

[25] C. Doersch, A. Gupta, L. Markeeva, A. Recasens, L. Smaira, Y. Aytar, J. Carreira, A. Zisserman, and Y. Yang, "Tap-vid: A benchmark for tracking any point in a video," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 610–13 626, 2022.

[26] 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.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[28] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[29] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 771–787.

[30] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6182–6191.

[31] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4660–4669.

[32] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8126–8135.

[33] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 448–10 457.

[34] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on robotics*, vol. 37, no. 1, pp. 291–303, 2020.

[35] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[36] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[37] 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.

[38] K. Luo, C. Wang, S. Liu, H. Fan, J. Wang, and J. Sun, "Upflow: Upsampling pyramid for unsupervised optical flow learning," in *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1045–1054.

[39] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4884–4893.

[40] D. Recasens, J. Lamarca, J. M. Fácil, J. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[41] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia, *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[42] G. Hattab, M. Arnold, L. Strenger, M. Allan, D. Arsentjeva, O. Gold, T. Simpfendörfer, L. Maier-Hein, and S. Speidel, "Kidney edge detection in laparoscopic image data for computer-assisted surgery: Kidney edge detection," *International journal of computer assisted radiology and surgery*, vol. 15, pp. 379–387, 2020.

[43] A. Lukezic, T. Vojir, L. ˇCehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6309–6318.