Self-Supervised Cyclic Diffeomorphic Mapping for Soft Tissue Deformation Recovery in Robotic Surgery Scenes

Shizhan Gong, Yonghao Long, Kai Chen, Jiaqi Liu, Yuliang Xiao, Alexis Cheng, Zerui Wang, Qi Dou

***Abstract*— The ability to recover tissue deformation from visual features is fundamental for many robotic surgery applications. This has been a long-standing research topic in computer vision, however, is still unsolved due to complex dynamics of soft tissues when being manipulated by surgical instruments. The ambiguous pixel correspondence caused by homogeneous texture makes achieving dense and accurate tissue tracking even more challenging. In this paper, we propose a novel self-supervised framework to recover tissue deformations from stereo surgical videos. Our approach integrates semantics, cross-frame motion flow, and long-range temporal dependencies to enable the recovered deformations to represent actual tissue dynamics. Moreover, we incorporate diffeomorphic mapping to regularize the warping field to be physically realistic. To comprehensively evaluate our method, we collected stereo surgical video clips containing three types of tissue manipulation (i.e., pushing, dissection and retraction) from two different types of surgeries (i.e., hemicolectomy and mesorectal excision). Our method has achieved impressive results in capturing deformation in 3D mesh, and generalized well across manipulations and surgeries. It also outperforms current state-of-the-art methods on non-rigid registration and optical flow estimation. To the best of our knowledge, this is the first work on self-supervised learning for dense tissue deformation modeling from stereo surgical videos. Our code will be released.**

***Index Terms*— soft-tissue deformation recovery, robotic surgery, diffeomorphic mapping, cycle consistency**

I. Introduction

Soft tissue deformation is everywhere in minimally invasive interventions performed by surgical robots [1]. Recovery of such deformation in real-time is fundamental for various downstream applications including quantification of instrument-tissue interaction [2], evaluation of surgical skills [3], estimation of anatomy biomechanical property [4], automation of soft-tissue manipulation [5], [6], etc. Aiming to represent the 3D geometry of tissue and further track its continuous changes over time, soft tissue deformation recovery has been a long-standing research topic in computer vision and computer-assisted surgery [1]. Despite being studied over a decade, it is still largely unsolved given challenges from complex dynamics in tissue manipulation, scene occlusions by surgical instruments, the unstructured environment in surgery, changes of illuminations, and artifacts of reflection.

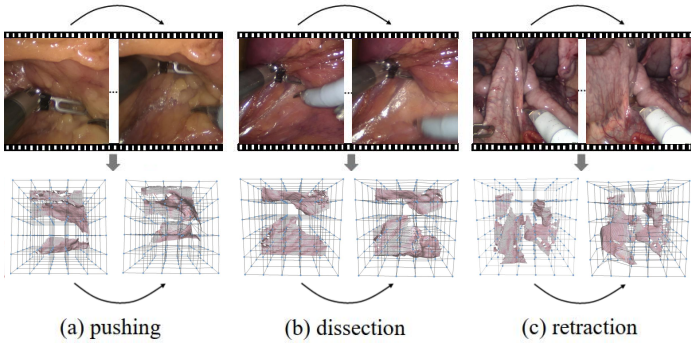

Fig. 1. Illustration of the soft tissue deformation recovery task with examples of different manipulations. Our method represents the learned deformation field in 3D.

Early methods adopted multi-sensory inputs to compensate visual information with tactile or force sensors [7], [8], however, these sensors are still not equippable to robotic systems for real surgeries. The current research trend is to explore methods of purely vision-based deformation modeling. For instance, Lagneau et al. [9] utilized a model-free visual servoing method to manipulate soft objects towards a desired shape. Li et al. [5] proposed a surgical perception framework that maps a deformation field with geometric information. These methods, however, represent the deformation in a parameterized way designed exclusively for visual servoing, which limits their potential for wider applications.

For a deformation recovery method to be considered qualified, not only should the recovered deformation reflect the correct tissue motion patterns, but also it adheres to primary physical rules. In other words, it should achieve both temporal-consistency and physical-plausibility. However, existing methods to recover 3D tissue deformation often fail to meet one or both criteria. Giannarou et al. [10] recovered the deformation field based on 3D thin-plate splines interpolation with Harris-Laplace matching. Zhou et al. [11] employed ORB feature matching to assist non-rigid iterative closest points. However, as sparse feature-based methods, homogeneous textures in

Shizhan Gong, Yonghao Long, Kai Chen, Jiaqi Liu, Yuliang Xiao, and Qi Dou are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: {szgong22, yhlong, kaichen}@cse.cuhk.edu.hk, {jiaqiliu, ylxiao, qidou}@cuhk.edu.hk).

Alexis Cheng and Zerui Wang are with Cornerstone Robotics Limited, Hong Kong, China (e-mail: {alexis.cheng,jerry.wang}@csrbtx.com).

surgical scenes make it hard to select informative feature points that reflect semantics. The partial observability due to occlusions and motion blur makes it more difficult to continuously track features over time, thus struggling to maintain temporal consistency. To address these limitations, several studies [12]–[14] have attempted to seek denser correspondence to achieve better temporal-consistency. However, the correspondence is established purely based on adjacent frames, failing to make use of long-term dependencies. In addition, dense mapping solely based on pixel values from homogeneous textures lacks semantic awareness and topology sensitivity, which makes the matching unrealistic and results in intersections or extreme displacements. This issue can be particularly problematic in situations that prioritize physical-plausibility, such as biomechanical quantification.

Recently, learning-based methods for deformation estimation and modeling have made promising progress in related tasks such as video reconstruction and image registration. This inspired us to learn the deformation field directly from data instead of relying on parameterized deformation models, thereby supporting more flexible and complex deformation and allowing for explicitly modeling 3D deformations. However, previous performance gain usually relies on strong supervision signals, such as requiring a large amount of synthetic data with ground truth labels [15], [16] or manual annotation of the landmark feature points [17], [18]. To reduce labeling cost which is often expensive in surgical scenarios, we consider recovering tissue deformations by purely learning from raw data without using any manual annotations. The technical challenges rooted in the intricate tool-tissue interaction that causes high complexity in deformation, but self-supervised models for semantics and motion extraction are promising.

In this paper, we propose a novel self-supervised framework for soft tissue deformation recovery of robotic surgery videos. To achieve this goal, we design a cyclic mechanism that prompts the learning of deformation over time from vision inputs explicitly. Furthermore, we combine semantics, inter-frame pixel-wise motion flow, and long-range temporal context to accurately estimate the deformation field, ensuring that the recovered deformation reflects actual tissue movements. More importantly, to encourage a physical-plausible deformation, we leverage diffeomorphic mapping to represent the tissue deformations, which can intrinsically enforce primary physical properties of the deformation field, such as topology-preserving and invertibility. To evaluate our method, we collect a dataset of robotic surgery videos and cut them into shorter clips, with each clip representing a complete manipulation action (see Fig. 1). The experimental results, both quantitatively and qualitatively, show that our approach can achieve impressive deformation recovery performance, surpassing all comparison methods. Our contributions are summarized as follows:

- We present a novel self-supervised learning framework to recover soft tissue deformation from surgical stereo videos, for the first time, to simultaneously emphasize both temporal-consistency and physical-plausibility.
- We design an effective scheme to integrate semantics, inter-frame flow, and long-range temporal information to model a sequence of deformation fields in real-time.
- We evaluate our approach on a robotic surgery dataset, demonstrating that such data-driven deformation modeling is generalizable across different types of tissue manipulations and surgical procedures.

II. Related Work

Soft tissue deformation recovery. Accurate modeling of the 3D motion of soft tissue in dynamic scenes is a fundamental yet challenging problem. Tissue deformation can be caused by the cardiac or respiratory cycle, tissue tool interaction, or muscular contraction [1]. Deformation caused by physiological cycles can be modeled as quasi-periodic or periodic signals [19], such as Fourier series [20], vector auto-regressive models [20], and Taken's theorem [21]. For non-periodic deformation, optical video-based methods through minimization of non-rigid matching and smoothing costs are often used [22]. Non-rigid ICP methods [23] are widely used to match the input with a template. However, it cannot track fast tissue deformation, and also suffers from poor alignment in the tangential directions on tissue surface [11]. Some works propose extracting feature descriptors like SIFT [24], SURF [25], or ORB [11] and then recovering the dense deformation field through interpolation. However, these methods often generate deformations with poor precision and low generalization, as they rely on sparse feature points that are noisy due to poor texture and boundaries in surgical videos. We propose to rely on dense correspondence which can be more robust to noisy pixels.

Diffeomorphic mapping. A diffeomorphism refers to a globally one-to-one smooth and continuous mapping with invertible derivatives. Diffeomorphic deformation guarantees the topology is preserved after deformation and also enforces consistency under compositions of the deformations. It has been widely applied in tasks including data augmentation [26], image registration [27], anatomical shape matching [28], and mesh reconstruction [29]. Therefore, it is also a potentially practical way to model soft-tissue deformation. In diffeomorphic deformation, the deformation field is parameterized by an underlying stationary velocity field, and the deformation is calculated by velocity integration, which can be approximated by scaling and squaring [30]. In this paper, we propose to use a diffeomorphic deformation to model the soft tissue movement, so the recovered deformation can be smooth and inherently satisfy several desired physical properties. This is crucial for many downstream tasks such as manipulation and biomechanical modeling.

Cycle consistency. Cycle consistency has been a classic idea in tracking [31] and recently has been extended beyond its original use as an evaluation metric [32] or uncertainty measure [33] to become a learning objective for various tasks such as registration [34], optical flow [35], image-to-image translation [36], depth estimation [37], and video segmentation [38]. One of its major forms, forward-backward consistency, is commonly used for determining the occlusion region during unsupervised optical flow estimation [35] and self-supervised correspondence learning [39]. Wang et al. [39] first employed cycle consistency across multiple steps in

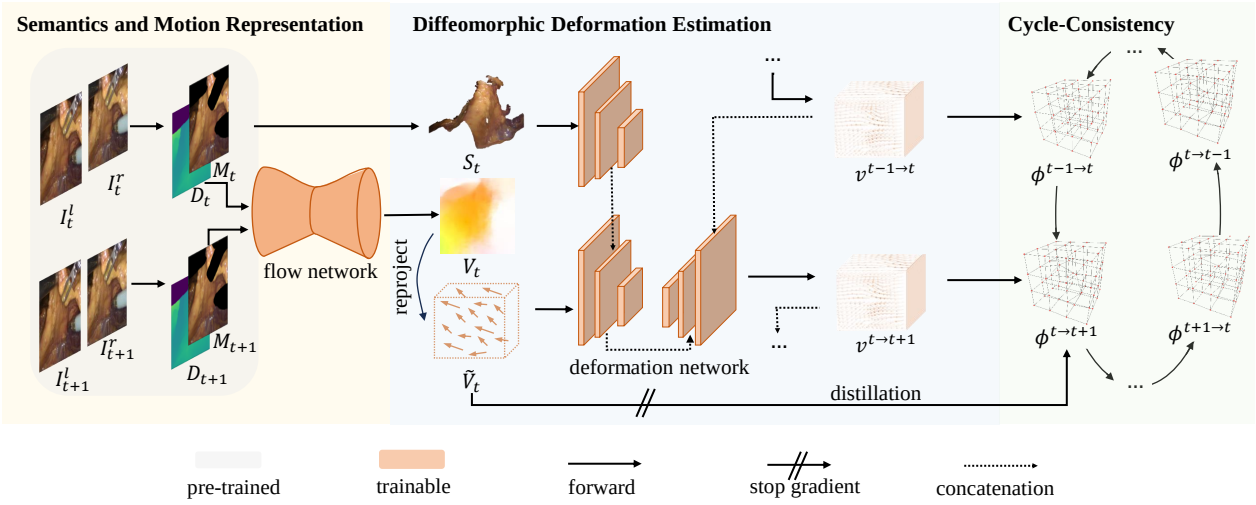

Fig. 2. The pipeline of our proposed soft tissue deformation recovery. Given two adjacent frames $\boldsymbol{I_t}$ and $\boldsymbol{I_{t+1}}$, we first estimate their depth $\boldsymbol{D_t}/\boldsymbol{D_{t+1}}$ and the tool masks $\boldsymbol{M_t}/\boldsymbol{M_{t+1}}$ so as to establish the semantic representation $\boldsymbol{S_t}$ and flow representation $\boldsymbol{V_t}$. A deformation network aggregates these representations together with temporal context and predicts a velocity field $\boldsymbol{v^{t\to t+1}}$, which is later integrated into deformation field $\boldsymbol{\phi^{t\to t+1}}$. Cyclic consistency is utilized as self-supervision.

time and proposed a framework that can solve multiple tasks including mask, texture, and pose propagation. Li et al. [38] found cycle consistency can alleviate error propagation and enhance temporal stability in video object segmentation. In this paper, we adapt cycle consistency to our specific problem setting. We incorporate both the entire forward-backward circle consistency loss and skip cycle loss to calculate the consistency at the intermediate level. We also calculate the consistency from both a 2D perspective and a 3D perspective.

III. Methods

A. Overview of self-supervised framework

This paper presents a self-supervised learning framework to tackle the task of soft tissue deformation recovery (see overview in Fig. 2). The input to the model consists of a sequence of 2D image pairs with both left and right views, $I_t, \cdots, I_{t+k} \in \mathbb{R}^{2\times H\times W\times 3}$ revealing tissue movements. The goal is to learn a series of physical-plausible deformation fields $\phi^{t\to t+1}, \cdots, \phi^{t+k-1\to t+k} \in \mathbb{R}^{H'\times W'\times D'\times 3}$ that accurately represent the actual deformation of the soft tissue in the 3D mesh, where $H\times W$ is the size of 2D image and $H'\times W'\times D'$ is the size of the scene in world coordinates. Given two adjacent frames, we first initialize their 3D representation and estimate the pixel-correspondence flow between them, which serves as the semantic and cross-frame flow representation. A deformation network is used to aggregate the semantic and flow representation and the long-range temporal information extracted from the deformation field of previous frames. The deformation network produces a stationary velocity field that could be integrated into deformation fields, therefore guaranteeing the diffeomorphism. As a learning goal, the cycle-consistency loss is calculated based on the similarity between the warped pixels before and after an entire forward-backward cycle, which forces the model to recover deformation fields that are in agreement with the tissue movement depicted in the image sequence throughout the cycle. Our framework is designed to be applicable to various deformation patterns in different procedures by fully exploiting both semantics and pixel-wise flow information from the image. At inference time, the deformation is predicted in a seamless and continuous manner, thereby supporting real-time applications.

B. Temporally-consistent semantics and motions

The first step in soft tissue deformation recovery involves a critical process of establishing the semantic representation of the tissue and estimating its pixel-wise motion flow. Semantics provides information in terms of where the deformation appears and flow representation can inform the direction and distance of the displacement. Combining both representations helps to accurately determine displacement vectors and thereby enhances the temporal-consistency of the recovered deformation. We expect the semantic and flow representation to well reveal the 3D dynamic scene of the tissue so that recovered deformation can be consistent with the tissue motion in real-world space rather than 2D image space. However, the inputs to our pipeline are all 2D images, where only 2D semantics and pixel-wise flow on the plane can be directly obtained. To this end, 3D reconstruction is needed to restore the 3D semantic and flow representation before plugging into the deformation network.

To reconstruct the 3D representation, depth estimation techniques are needed to re-project the pixel value on the 2D image back to the 3D space. Due to the lack of ground truth in the surgical scenarios, we use a pre-trained STereo TRansformer (STTR) [40] for the stereo depth estimation. It uses a transformer to formulate the depth estimation problem as a matching problem, which can generalize well to different domains without fine-tuning. We experiment with the model pre-trained on Scene Flow dataset [41] provided by the original work, and find it works well on our dataset. However, directly applying a dense depth estimation network can result in noisy

estimation due to uneven illumination, ambiguous boundaries, and motion blurs. Therefore, we propose a gradient-based method to detect and remove the outliers. To be more specific, for each pixel, we compare its depth value with its four neighbors. If at least two neighbors have similar depth values (defined to be the difference is less than 1 mm), then it is assumed to be an inlier. Otherwise, this pixel is an outlier and thereby should be removed. Note this may also remove some points at the boundaries that are not necessarily outliers, however, this will only affect a few pixels without influencing too much in terms of global deformation estimation. We use D_t to denote the estimated depth at frame t.

To avoid interruption by the intensive and rigid motion of instruments, which may contaminate the semantic and flow representation of the soft tissue and therefore degenerate the temporal-consistency, we also pre-train an instruments segmentation model to mask the instrument region (denoted as M_t) and isolate instrument motion. Combining depth estimation and tool mask prediction enables us to faithfully reconstruct the scene of soft tissue in the 3D world coordinate.

For flow representation, we first obtain the pixel-wise dense flow from the 2D images and then re-project the 2D displacement back to 3D space based on the estimated depth. The motion flow is estimated through dense correspondence learning techniques. The training of the flow network follows a similar unsupervised schema as the UFlow [42]. Given input images of two adjacent frames I_t and I_{t+1} (both are left view), the network predicts an optical flow map $V_t \in \mathbb{R}^{H\times W\times 3}$ denoting the motion of each pixel from I_t to I_{t+1}. We denote the reprojected 3D flow field to be $\tilde{V}_t \in \mathbb{R}^{H'\times W'\times D'\times 3}$. The unsupervised loss consists of an occlusion-aware photometric consistency term $\mathcal{L}_{photo}$ and an edge-aware smooth term $\mathcal{L}_{smooth}$. The photometric consistency term is defined as

$$\mathcal{L}_{photo}(I_t, I_{t+1}, V_t) = \frac{1}{HW}\sum O_t \odot M \odot \rho(I_t, w(I_{t+1}, V_t)),$$

where $\frac{1}{HW}\sum$ is a shorthand notation for the mean over all pixels. The function $w(\cdot,\cdot)$ inverse warps an image with a flow field. The function $\rho(\cdot,\cdot)$ measures the photometric difference between two images based on a soft Hamming distance on the Census-transformed images and applies the generalized Charbonnier function. The occlusion masks $O_t \in \{0,1\}^{H\times W}$ is estimated through a range-map-based occlusion estimation [43] with gradient stops. M is the instrument mask generated by the multiplication of masks of both frames $M = M_t \odot w(M_{t+1}, V_t)$, whose gradient is also stopped. $\mathcal{L}_{smooth}$ is selected to be first-order edge-aware smoothness,

$$\mathcal{L}_{smooth}(V_t, I_t) = \frac{1}{HW}\sum(\exp(-\frac{\lambda}{3}\sum_c \left|\frac{\partial I_{t_c}}{\partial x}\right|) \odot \left|\frac{\partial V_t}{\partial x}\right| + \exp(-\frac{\lambda}{3}\sum_c \left|\frac{\partial I_{t_c}}{\partial y}\right|) \odot \left|\frac{\partial V_t}{\partial y}\right|),$$

where λ is a hyper-parameter controlling the sensitivity to visual edges and c denotes three color channels. and the overall loss function for the optical flow estimation is a weighted combination of photometric loss and smooth loss $\mathcal{L}_{flow} = \mathcal{L}_{photo} + w_{smooth}\mathcal{L}_{smooth}$. For certain flow estimation methods that contain intermediate predictions (e.g., RAFT [44]), we apply the losses to both the final output and intermediate predictions, with an exponentially decayed weight schema $\mathcal{L}_{flow} = \sum_{i=1}^{n}\gamma^{n-i}\mathcal{L}^i_{flow}$, where n is the number of flow iterations, γ is the decay factor and L^i_{flow} is the loss at iteration i. In our experiments, we set n to be 12 and γ to be 0.8. For the original UFlow, the intermediate predictions have different scales from the final output. Therefore, we only apply the loss to the final prediction, but the loss is multiplied by 5 to balance with the subsequent loss terms.

For simplicity, we skip the $\mathcal{L}_{self}$ of the original UFlow. It is designed to improve the performance on the margin of the images, while we care more about the tool-tissue interaction at the center of the images. Note that our framework is highly adaptable, and designed to be used with a variety of networks or methods for estimating depth or flow. We interpolate the recovered semantics and flow into a grid volume with a predefined size as the final representation, making it easier for subsequent processing.

C. Physically-plausible diffeomorphic mapping

Re-projecting 2D motion flow directly onto a 3D space may produce deformation fields that are not physically-plausible. Firstly, unconstrained dense correspondence is semantic unaware and topology agnostic, which may result in rough and intersecting correspondences. Moreover, 2D flow estimation is depth-unaware. A reasonable correspondence on the 2D plane may be unrealistic from a 3D perspective. Furthermore, the estimated depth contains inherent noise due to multiple factors such as ambiguous boundaries, uneven illumination, and motion blur, which may cause the reconstructed pixel to deviate far away from its actual location. This flow also hardly satisfies long-term temporal-consistency, as it is estimated based on two adjacent frames without incorporating long-range temporal information, and can easily fail when pixels are masked by instruments midway. As a result, a separate deformation network is required to further aggregate the semantics and flow representation, together with the long-range temporal context, making it both temporal-consistent and physical-plausible. As a solution to encourage plausible deformation fields, we leverage diffeomorphism to model the deformation. Diffeomorphism is an invertible mapping where the forward and backward transformations are both smooth. In this regard, the topology of the tissue can be preserved after the deformation, which agrees better with the practical deformation. Moreover, as the composition of multiple diffeomorphisms is also diffeomorphism, it is inherently proper to model sequential deformations over time.

To this end, we rely on a new deformation network that takes both semantic and flow representation as inputs and produces the deformation field $\phi^{t\to t+1} \in \mathbb{R}^{H'\times W'\times D'\times 3}$. The deformation field is parameterized with a stationary velocity field $v^{t\to t+1} \in \mathbb{R}^{H'\times W'\times D'\times 3}$, which is the direct output of the deformation network, so that the diffeomorphism can be guaranteed. The path of the diffeomorphic deformation field ϕ parameterized by $s \in [0,1]$ is generated by the velocity fields:

$$\frac{\partial \phi}{\partial s} = v(\phi^{(s)}) = v \circ \phi^{(s)},$$

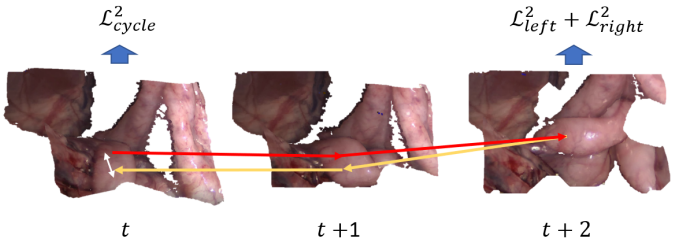

Fig. 3. Cyclic-consistency enforces the points to re-arrive at their initial position after an entire forward-backward cycle.

where v denotes the velocity field that remains the same alongside s, $\phi^{(0)}$ is the identity transformation, and $\circ$ is the composition operator. The deformation at time $s = 1$ (i.e., $\phi^{(1)}$) is the final deformation field we need (i.e., $\phi^{t\to t+1}$). With such parameterization, the deformation field is a member of a Lie group and can force the mapping to be diffeomorphic and invertible. The deformation field can be obtained through the integration of the velocity field via the scaling and squaring methods [30]. Moreover, to utilize the temporal context, we concatenate the previous step's velocity field to the feature maps of the current frame's decoder. Thereby, the prediction of the current velocity field can take previous fields as references and further promote temporal coherence and smoothness. To facilitate the convergence of the deformation network, we design a consistency loss $\mathcal{L}_{cons}$ to distillate the displacement from the flow representation to the deformation field, which is defined as the $l1$-norm distance between the flow representation and the deformation field. As shown in Fig. 2, we stop the gradient of the consistency loss with respect to flow representation. This is because the flow network is pre-trained, while the deformation network is initialized with an identity transformation. By stopping this gradient, we can prevent the pre-trained flow network from collapsing, resulting in a more stable training process. We also encourage a smooth velocity field with a diffusion regularizer on the spatial gradients of velocity $\mathcal{L}_{spatial}$:

$$\mathcal{L}_{spatial} = \frac{1}{H'W'D'}\sum\left(\left\|\frac{\partial v}{\partial x}\right\|^2 + \left\|\frac{\partial v}{\partial y}\right\|^2 + \left\|\frac{\partial v}{\partial z}\right\|^2\right),$$

where $\frac{1}{H'W'D'}$ is a normalization factor for the mean over all grids, which helps to avoid a discontinuous velocity field.

D. Cycle-consistency leaning objectives

To complete the self-supervised workflow, we rely on cycle-consistency as our learning objective. Consider as inputs a sequence of video frames, the estimated deformation fields can gradually warp the images and generate a deformed image. Note the deformation happens in the 3D space and the deformed image is the projection in the 2D image space. We expect the image to re-arrive to its original state after an entire forward-backward circle, and we also enforce the image to be similar to the true frames along the path, as depicted in Fig. 3. The former enables the learning objective to be robust to discontinuous semantics caused by partial observable scenes midway, while the latter guarantees the recovered deformation is temporal-consistent all the way along the path. Combining both terms contributes to a stronger learning signal that can prevent the learning process from taking shortcuts [45].

Through accumulating the step-wise deformation, we can derive the total forward deformation, $\phi^{t\to t+i} = \phi^{t\to t+1} \circ \cdots \circ \phi^{t+i-1\to t+i}$ and the total backward deformation in a similar way. The cycle loss of i steps $\mathcal{L}^i_{cycle}$ is defined as the $l1$-norm distance between the original pixel coordinates and the deformed pixel coordinates after the entire forward-backward circle, both in the world coordinates. The skip cycle loss $\mathcal{L}^i_{left}$ and $\mathcal{L}^i_{right}$ computes the photometric similarity between warped 2D views at the end of the forward path and their corresponding true image views. For the left view, the photometric loss is based on a soft Hamming distance on the Census-transformed images that applies the generalized Charbonnier function [46]. For the right view, we minimize the angle between RGB vectors:

$$\mathcal{L}^i_{right} = \frac{1}{HW}\sum M \odot cos^{-1}\left(\frac{I^r_{t+i}\cdot w(I^r_t, \phi^{t\to t+i})}{\|I^r_{t+i}\|\|w(I^r_t, \phi^{t\to t+i})\|}\right),$$

where $\frac{1}{HW}\sum$ is shorting for the mean over all pixels, M is the inverse tool masks, $w(I^r_t, \phi^{t\to t+i})$ is the warped image. Here cos^{-1}, $\cdot$, and $\|\cdot\|$ are all calculated per pixel. The reason for the difference is that census transformation relies on relative intensity, but the light source of the surgical scene makes the brightness inconsistent for different views. By incorporating photometric loss of both views, we ensure the recovered deformation is temporal-consistent in a 3D perspective. During each training iteration, we take $k+1$ frames as inputs. The overall cycle-consistency loss sums over the k possible cycles:

$$\mathcal{L}_{cycle} = \sum_{i=1}^{k} \mathcal{L}^i_{cycle} + w_{left}\mathcal{L}^i_{left} + w_{right}\mathcal{L}^i_{right}.$$

The overall loss is a weighted combination:

$$\mathcal{L} = \mathcal{L}_{flow} + w_{cons}\mathcal{L}_{cons} + w_{spatial}\mathcal{L}_{spatial} + w_{cycle}\mathcal{L}_{cycle}.$$

In this regard, we can train the flow network and deformation network simultaneously, so as to achieve both temporal-consistency and physical-plausibility.

E. Implementation details

The segmentation network is based on the Swin Transformer [47], which contains a Swin transformer backbone, a upernet decoder, and an auxiliary FCN decoder. We collected the clinical trial data from the same source for training deformation recovery, to form our training dataset. Sequences in which significant instrument interactions were observed from independent videos are selected. The data are labeled by a trainee with a three-year related experience and verified by a team of engineering students, where the instruments are labeled as foreground. We labeled only the left frames to reduce the workload. The annotations are performed with "LabelMe" [48] toolbox. The implementation of networks is based on mmSegmentation [49]. The input images are cropped to 1024 $\times$ 1024 and recovered to the original size after the network. Data augmentation is adopted to improve the robustness of the algorithm with horizontal and vertical flips,

TABLE I

QUANTITATIVE EVALUATION OF OUR METHOD COMPARED WITH EXISTING METHODS. WE TRAIN THE MODEL WITH ALL TRAINING DATA AND REPORT THE SOFT TISSUE MANIPULATION-LEVEL METRIC ON THE TEST SET. THE BEST NUMBER FOR EACH CATEGORY IS HIGHLIGHTED IN BOLD.

Models	$\%\|J_\phi\| \leq 0 \downarrow$				$l1$-norm $\downarrow$				PSNR $\uparrow$				SSIM $(\%) \uparrow$			
	push	dissect	retract	total	push	dissect	retract	total	push	dissect	retract	total	push	dissect	retract	total
NICP	3.53	6.58	3.87	4.65	22.78	24.04	26.34	24.46	17.85	17.92	16.90	17.53	64.92	63.92	58.34	62.25
CPD	8.14	7.21	7.44	7.56	22.31	21.74	24.21	22.82	17.84	18.42	17.54	17.89	61.21	63.88	58.46	60.90
SIFT	0.15	0.14	0.41	0.24	21.03	20.85	25.31	22.53	18.78	18.92	17.65	18.41	76.46	76.64	71.47	74.73
Harris–Laplace	0.15	0.14	0.41	0.26	21.02	20.94	25.23	22.56	18.79	18.87	17.67	18.40	76.46	76.59	71.58	74.64
RAFT	3.64	3.86	3.37	3.61	13.04	12.00	13.61	12.89	21.92	23.08	21.60	22.19	84.29	85.59	82.19	83.97
UFlow	3.63	3.83	3.35	3.59	12.98	12.03	13.69	12.92	21.97	23.04	21.57	22.18	84.52	85.82	82.23	84.13
Ours (+RAFT)	**0.01**	**0.01**	**0.03**	**0.02**	13.03	**11.94**	**13.43**	**12.80**	22.02	23.09	**21.73**	22.27	84.61	85.97	**82.82**	84.42
Ours (+UFlow)	**0.01**	**0.01**	**0.03**	**0.02**	**12.86**	12.00	13.56	12.83	**22.11**	**23.12**	21.67	**22.28**	**84.70**	**86.03**	82.69	**84.43**

image shifts, and rotations [50]. The segmentation network is initialized with a pre-trained model on [49]. Binary Cross-Entropy loss is adopted to optimize both the decoder branch and the auxiliary FCN branch, which sets the instrument regions as foreground and the tissues as background. The loss ratio of the auxiliary branch is set as 0.4. We used the Adam optimizer with a learning rate of 1e-6 and the StepLR scheduler with a decay rate of 0.1 for every 10 epochs. The maximum epoch is set to be 30.

For the flow network, we use RAFT [44] and UFlow [42], with both pre-trained on FlyingThings3D [41]. The deformation network structure is borrowed from VoxelMorph-dif [51], whose encoder is duplicated doubly to adapt to the multi-modal setting, and skip connections across encoders are used for cross-modality fusion. The velocity field is also concatenated to the last layer of the decoder. The original size of the rectified stereo image is 740×540, which is then cropped to 512×512 before plugging into the flow network. The size of the input to the deformation network is $64 \times 64 \times 64$, where each pixel equals $1\ mm$. Both flow and deformation networks are trained by the Adam optimizer with an initial learning rate of $2e-4$ which linearly decays to $1e-5$ in 20 epochs. The coefficients of each loss term are as follows in our experiments: $w_{left} = w_{right} = w_{cons} = 1.0$, $w_{spatial} = w_{cycle} = 0.1$. All three models are implemented with PyTorch 1.12.1 using one Nvidia GeForce RTX 3090.

IV. EXPERIMENTAL RESULTS

A. Dataset and evaluation metrics

To evaluate the efficacy of our method, we collect an in-house dataset to evaluate the model performance. The dataset is collected during the clinical trials conducted by expert surgeons using a private surgical robot system. Ethics approval of the study protocol performed at Multi-Scale Medical Robotics Center (MRC) has been granted by IACUC of the Hong Kong Science and Technology Parks Corporation (HKSTP) with a reference number (HKSTP IACUC ref. no.: 2021-010). There are two types of surgical procedures in the dataset, Right Hemi-colectomy (RHC) [52] and Total Mesorectal Excision (TME) [53]. Distinct from existing public surgical datasets which only contain monocular videos [54]–[56] or only provide video images at low frame rate (1 or 2 Hz) [57]–[59], we recorded synchronized stereo videos at 25 fps and calibrate the stereo camera parameters for precise and fine-grained deformation recovery. Specifically, we sort out three common surgical actions involving typical and obvious tissue deformations from the video frames for evaluation: (1) **Pushing** [59], [60]: push and then manipulate parts of the tissues to provide a better operational space for other surgical action; (2) **Dissection** (blunt) [61], [62]: separate tissues along tissue planes using blunt parts of the instruments to avoid harming sensitive tissue structure; (3) **Retraction** [63], [64]: grasp and then lift up parts of the tissues to expose the area of interest (e.g., tumor and vessel). The dataset consists of 180 sequences of stereo video frames, equally distributed for each surgical action and procedure. The average length for all sequences is around 26 frames (1 second), ranging from 13 frames to 47 frames (lasting for 0.52-1.88 seconds), which is the exact time that is needed to complete one movement of the surgical action. The dataset was randomly split into a 70% training set, a 10% validation set, and a 20% testing set, maintaining the proportional distribution of surgical actions and procedure types in each dataset.

Validating in vivo results is challenging without ground truth. As commonly adopted by the community, performance can be evaluated by comparing the photometric similarity between original and deformed images. The reference frame is used to reconstruct a point cloud and its points are deformed based on recovered temporal deformation. The deformed point cloud is projected onto the 2D image plane and similarity with the original image is measured using $l1$-norm, peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM). These metrics can reflect whether the recovered deformation is temporal-consistent. The percentage of positive determinants for the Jacobian matrix is reported to reflect whether the deformation is topology-preserving and physical-plausible.

B. Experimental setting

We compare our results with three types of deformation estimation methods that can be adapted to address our task: (1) **Point Cloud Based.** We apply NICP [65] and Coherent

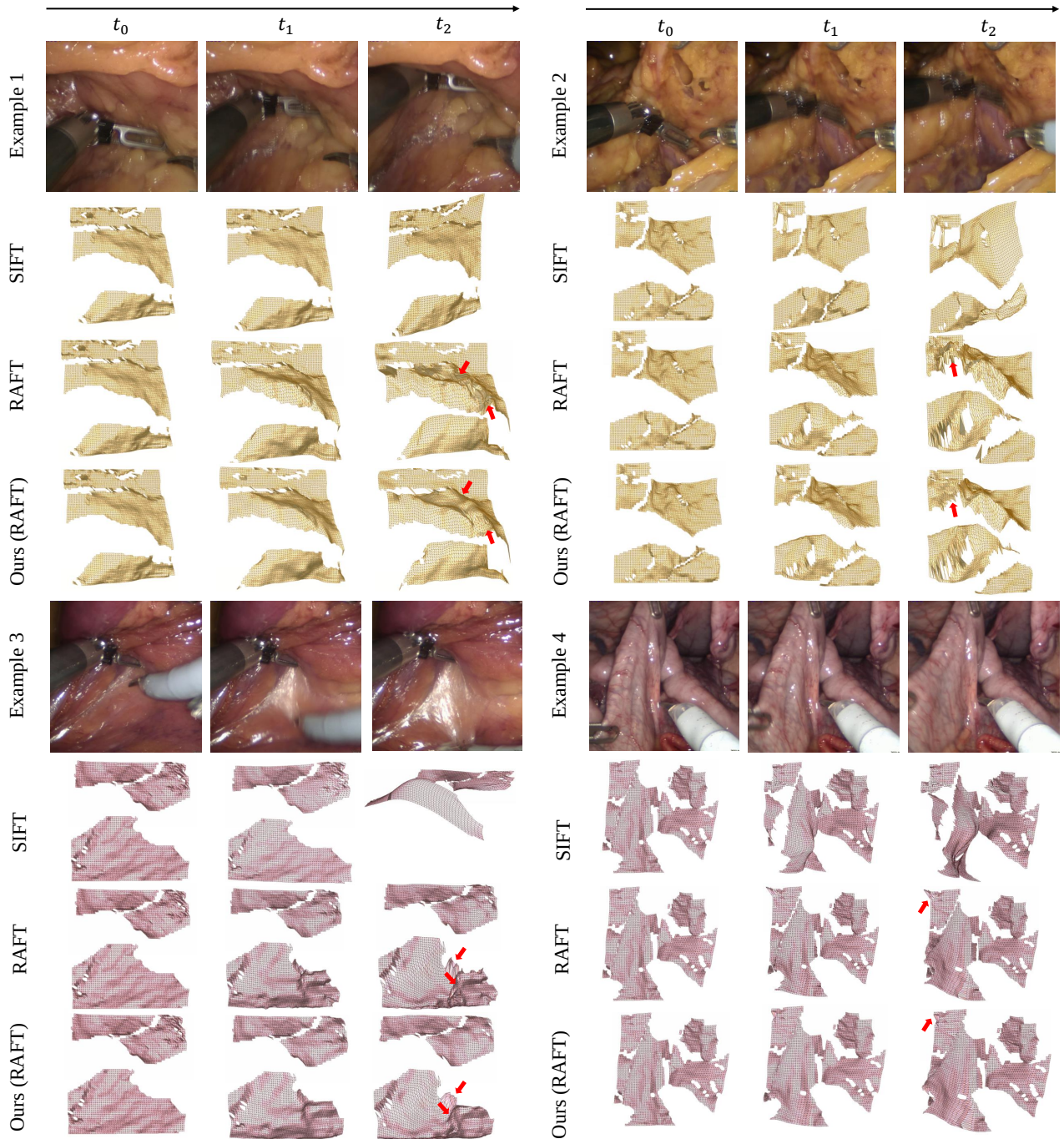

Fig. 4. Qualitative comparison of the deformation recovery results with state-of-the-art methods. Our methods can accurately reflect the movement of the soft tissue, and the region with the right arrow shows our recovered deformation is smoother and more realistic than dense flow based methods. (Blank regions correspond to regions sheltered by the instruments during the first frame, regions with noisy depth estimation, or margins where sudden depth change happens.)

Point Drift (CPD) [66] as comparison methods based on point cloud registration. The deformation is optimized to align the reconstructed point cloud of two adjacent frames; (2) **Sparse Feature Based.** Following [10], we use Harris-Laplace [10] and SIFT [24] as sparse feature descriptors and then apply feature points match and thin-plate interpolation to recover the deformation field; (3) **Dense Flow Based.** We also compare our method with several dense corresponding learning techniques including UFlow [42] and RAFT [44], [67]. Both models are pre-trained on FlyingThings3D and fine-tuned on our data using the scheme mentioned in Sec. III-B.

We test generalization and conduct ablation studies to verify components in our pipeline. We train and test our model on different actions/procedures and compare results. Ablation studies include training a deformation network with flow representation only, removing temporal aggregation, removing $\mathcal{L}_{cons}$, removing the cycle loss while only keeping the skip cycle loss, and comparing separate/joint training of flow and

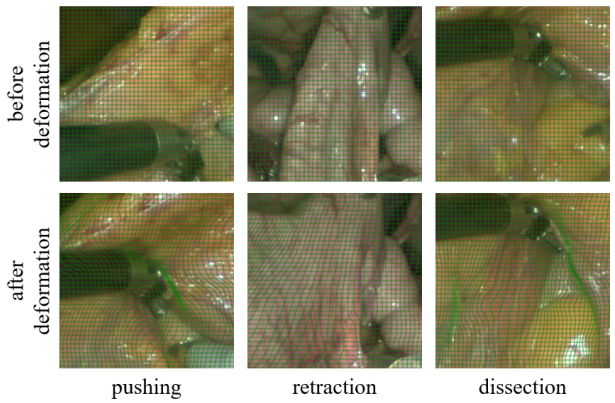

Fig. 5. Visualization of deformation field on 2D plane. We draw 2D grids, unproject to 3D space to move with the deformation field, and finally project to the 2D plane.

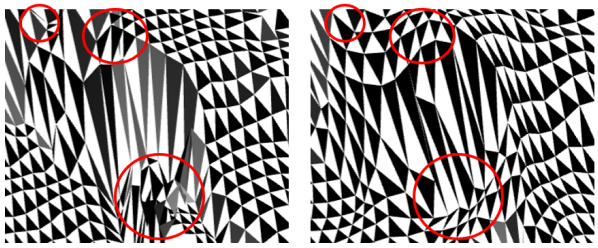

Fig. 6. Directly re-projecting the dense matching results in rough deformation fields and may cause topology changes with large displacement. Diffeomorphic mapping is topology-preserving and spatially smooth.

TABLE II
GENERALIZABILITY ACROSS VARIOUS MANIPULATIONS. WE TRAIN THE MODEL ON ONE MANIPULATION AND TEST IT ON THE OTHER.

Data	$\lvert J_\phi \rvert \downarrow$	$l1\downarrow$	PSNR↑	SSIM ↑
Pushing→Pushing	0.01	13.23	21.91	84.48
Dissection→Pushing	0.01	13.42	21.78	84.10
Retraction→Pushing	0.02	13.45	21.89	84.34
Pushing→Dissection	0.01	12.42	22.87	85.06
Dissection→Dissection	0.02	12.17	23.00	85.80
Retraction→Dissection	0.02	12.64	22.76	85.25
Pushing→Retraction	0.01	13.98	21.52	81.79
Dissection→Retraction	0.02	14.16	21.37	81.55
Retraction→Retraction	0.01	14.08	21.50	81.93

deformation networks.

C. Comparison with state-of-the-art methods

From Table I, we can see our method outperforms all other methods, with both overall and action-level performance. We also visualize the deformation by constructing surface mesh based on the reconstructed point cloud of the first frame and visualize its dynamic morphology changes (Fig. 4). Besides, we draw a 2D visualization of the deformation fields for a few cases for a clearer illustration (Fig. 5). In terms of temporal-consistency, our method is much better than point cloud based deformation strategy and sparse feature matching methods, with the overall performance uplift of more than 43.19%, 21.02%, and 12.98% for $l1$-norm, PSNR, and SSIM respectively. The reason is that the point cloud based deformation is semantic unaware. The sparse feature matching may easily fail because there are only a few feature points due to homogeneous textures. And many feature points are located at the margin, where depth estimation is inaccurate. Moreover, both methods suffer from huge performance drops with partial observable scenes masked by instruments. While our method fully utilizes the semantic information and is more robust to occlusions through the cyclic mechanism. Compared with dense correspondence-based methods, our method only has marginal gains in terms of temporal-consistency metrics. The performance uplift is 0.70%, 0.41%, and 0.43%. However, from the qualitative visualization, the mesh deformed by dense flow will gradually become rough, especially for regions with large displacement, while meshes generated by our method continue exhibiting high quality over time. This is because the dense flow based methods seek pixel correspondence purely on two adjacent frames while our method takes long-range temporal dependencies into consideration, making it perform better in the long run.

We also find our method can greatly enhance the physical-plausibility of the deformation, whose percentage of non-positive determinants of the deformation Jacobian matrix is less than 0.02%, while that of dense flow based methods can be greater than 3.59%. We further compare the deformation recovered by our method and pure dense flow by drawing the local pattern of the deformed triangle mesh (Fig.6). Our method guarantees spatial smoothness and topology-preserving, with the triangles in the deformed triangle mesh perfectly aligned with their neighborhoods. While purely applying optical flow can result in topology changes, represented by zig-zag patterns and the intersection of adjacent triangles.

Our pipeline (without depth estimation and instrument segmentation) achieves an inference speed of 20 fps. With many lightweight depth estimations and instrument segmentation candidates running in parallel with flow estimation, our method has the potential to support real-time application.

D. Analytical results for our approach

1) Generalizability: Our method can achieve good generalization under action shift (Table II) or procedure shift (Table III). The physical-plausibility is well-maintained as expected because distribution shift has no influence on diffeomorphic mapping. In terms of temporal-consistency, there is almost no performance drop when adapting a pre-trained model to another dataset with different actions, and the performance drop across procedures is less than 0.90% in terms of SSIM. The reason for the robustness is that our model can make the utmost use of the semantic information and pixel-level flow information, both of which are action-agnostic. Although procedure shift may result in a certain gap in terms of pixel values, the performance drop is still within a reasonable range. Therefore, our method has the potential to help robotic surgeries under variable environments.

TABLE III
GENERALIZABILITY ACROSS DIFFERENT PROCEDURES. WE TRAIN THE MODEL ON ONE PROCEDURE AND TEST IT ON THE OTHER.

Data	$\lvert J_\phi\rvert\downarrow$	$l1\downarrow$	PSNR↑	SSIM↑
TME →TME	0.02	12.52	22.73	85.75
RHC →TME	0.02	12.49	22.65	85.40
TME →RHC	0.01	14.06	21.47	81.84
RHC →RHC	0.01	13.61	21.66	82.62

TABLE IV
MODEL PERFORMANCE ON LONG VIDEO CLIPS (5S), WITH A COMPARISON OF DENSE FLOW-BASED BASELINE.

Methods	$\lvert J_\phi\rvert\downarrow$	$l1\downarrow$	PSNR↑	SSIM↑
RAFT	3.32	19.11	18.98	76.50
Ours (+RAFT)	**0.01**	**17.45**	**19.84**	**78.35**

2) Long video robustness: We conducted additional experiments on longer video clips to assess the performance of our methods in recovering deformations over a longer time horizon. Longer clips typically involve camera movements, movements of surgical tools, and the appearance or disappearance of new tissues that cannot be accurately modeled through deformation. To address this, we carefully selected clips from the testing phase videos that better reflect tissue deformation, resulting in 15 video clips, each approximately 5 seconds long. However, each clip still contains multiple different actions and may include snippets that cannot be modeled through deformation, particularly movements of vacant surgical tools. We evaluated the performance of our model on these longer snippets and compared it with dense flow-based methods, using RAFT as the flow estimation network for both methods. The results, shown in Table IV, indicate that although both methods experience a drop in performance as the clips become longer due to the complexity of the scene dynamics, our method outperforms the dense-correspondence-based method. This is because our method incorporates temporal information during modeling and employs a cyclic mechanism, which enhances the robustness of deformation estimation to temporal evolution. Additionally, the use of diffeomorphism modeling ensures that the reconstructed deformation fields are physically plausible, preventing excessively large deformations or discontinuous flickers at intermediate time frames and making our method more robust in the long run.

3) Ablation study: Our ablation study results in Table V reveal the importance of several of our components from the pipeline. Specifically, we perform both removal and incremental analysis. For removal analysis, we remove each component from the full model separately and compare their performance with the full model, while for incremental analysis, we fix a base model that only plugs the motion representation to the deformation network and a single-step photometric loss is used to train the deformation network. We then add different components to the base model to observe their benefits. All major elements of the deformation network, namely semantic representation, inter-frame flow information, and long-range temporal context, contribute to better model performance, which shows our deformation network can truly integrate information from all sources. This observation is especially true in the incremental experiments, which show improvements of 0.61%, 1.77%, and 1.45% in terms of SSIM when adding these three representations to the base model separately. The cyclic mechanism is an indispensable component of our model training, which can significantly improve the model performance. Moreover, we find that jointly training the flow network and the deformation network can also bring out performance gains compared with separate training.

V. DISCUSSION

Our pipeline relies on multiple pre-trained models (i.e., depth estimation, mask generation, and flow estimation) to obtain the semantic and motion representations. Their accuracies can somehow affect the results of our pipeline. In this section, we briefly analyze how their error will influence the results of our pipeline and how we can mitigate the negative influence.

1) Depth estimation: Accurate depth estimation is essential for determining tissue position in our framework. However, if the depth contains significant errors, the transformed world coordinates may not accurately represent the tissue's actual position, rendering the recovered deformation meaningless for downstream tasks. To ensure the reliability of depth estimation, we employ multiple mechanisms. Firstly, existing methods like STTR have demonstrated excellent generalization to surgical scenes and have been widely used in previous works. Secondly, we verify the accuracy of depth estimation by comparing the original image with its reprojected version, which involves unprojecting the image to 3D space based on the estimated depth and then rendering it back to 2D space. This independent examination has shown high consistency in our case. Thirdly, we employ self-supervised fine-tuning techniques to improve depth estimation accuracy for specific datasets. Additionally, we address potential inaccuracies at object boundaries by using a gradient-based method to remove noisy points and minimize their impact on the 3D representation. Empirically, we have observed that this scheme only removes a small proportion of points, which has minimal effect on the final performance. Moreover, for most deformation estimation tasks, the accuracy of boundary deformation is less critical compared to that of object centers. Therefore, errors resulting from inaccurate depth estimation remain within an acceptable range.

2) Tool mask generation: Accurately estimating deformation in soft tissue requires accounting for tool movement, which typically undergoes rigid displacement compared to the soft tissue. If the tools are not properly masked, their movement will be incorporated into the estimated deformation, compromising accuracy. Fortunately, tool segmentation is a well-established task with high performance. The model is less affected by false positives (tools not fully masked) compared to false negatives (tissue mistakenly masked). In the case of false positives, where some parts of the tools are not masked, their remaining movement can still impact deformation estimation. However, false negatives can be compensated for

TABLE V
ABLATION STUDIES ON INPUTS OF THE DEFORMATION NETWORK, LOSS TERMS, AND TRAINING SCHEMA.

Methods	$\lvert J_\phi \rvert \downarrow$	$l1\downarrow$	PSNR↑	SSIM ↑	Methods	$\lvert J_\phi \rvert \downarrow$	$l1\downarrow$	PSNR↑	SSIM ↑
Full model	0.02	**12.80**	**22.27**	**84.42**	Base Model	**0.01**	14.59	21.54	81.87
W/o semantics	0.02	13.07	22.18	84.23	W/ semantics	**0.01**	14.57	21.54	82.37
W/o temporal dependency	0.03	13.10	22.18	83.76	W/ temporal dependency	**0.01**	13.55	21.87	83.06
W/o distillation	**0.01**	13.48	21.98	83.69	W/ distillation	**0.01**	**13.49**	**21.99**	**83.32**
W/o cycle consistency	0.02	13.38	21.99	83.55	W/ cycle consistency	**0.01**	13.75	21.88	82.96
W/o joint training	**0.01**	13.02	22.20	84.20	W/ joint training	**0.01**	13.83	21.84	82.85

by employing smooth regularization and ensuring temporal consistency. To minimize the influence of tool movement on deformation estimation, we expand the predicted segmentation masks by 20 pixels to thoroughly mask the tools.

3) Flow estimation: Pre-trained flow estimation usually performs poorly on surgical videos due to the homogeneous textures of the tissues. As a result, this module is trainable in our framework through photometric consistency. Moreover, we refine the initial flow estimation with diffeomorphic constraints and cyclic-consistency. Even if the initial estimation of the flow field is inaccurate, the final deformation field will still accurately reflect the deformation of the soft tissue.

VI. CONCLUSION

This work shows a novel self-supervised learning framework that can recover soft tissue deformation in high quality from stereo surgical videos. Our method extracts visual features to yield the 3D deformation field with both temporal-consistency and physical-plausibility. Experiments present that our model can achieve state-of-the-art performance and can generalize well to different tissue manipulation and procedure types. We hope that this work serves as the first step forward to reliable data-driven solutions for visual-based deformation modeling for soft tissues, which holds enormous potential for robotic surgery applications.

There are still several limitations of our work. First, our model can hardly handle extremely large deformation with the appearance of a large amount of new tissue. Second, a good-quality depth estimation can be the bottleneck of our model performance. Third, the pipeline of our training network entails the explicit usage of the camera parameters, affecting its generalizability to videos caught by different cameras.

As for the evaluation scheme, the current approach is still not perfect as it is based on the photometric similarity between the deformed image sequence and the true image sequence, which can only be regarded as a heuristic evaluation. A more scientific evaluation scheme would require 3D ground truth labels which is extremely difficult to retrieve due to the constraint of the surgical setting. But as we continue promoting this research, it can be overcome, potentially by incorporating simulators or animal trials.

For future work, we will try to overcome these limitations, In addition, based on the recovered deformation, we will establish a mapping from robot kinematics to the tissue deformation which can further be incorporated into a planning and control pipeline to achieve automated soft-tissue manipulation.

REFERENCES

[1] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 14–24, 2010.

[2] A. K. Golahmadi, D. Z. Khan, G. P. Mylonas, and H. J. Marcus, "Tool-tissue forces in surgery: A systematic review," *Annals of Medicine and Surgery*, vol. 65, p. 102268, 2021.

[3] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1–8, 2019.

[4] E. Tagliabue, M. Piccinelli, D. Dall'Alba, J. Verde, M. Pfeiffer, R. Marin, S. Speidel, P. Fiorini, and S. Cotin, "Intra-operative update of boundary conditions for patient-specific surgical simulation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 373–382.

[5] 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.

[6] E. Tagliabue, A. Pore, D. Dall'Alba, E. Magnabosco, M. Piccinelli, and P. Fiorini, "Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 3261–3266.

[7] S. Giannarou, Z. Zhang, and G.-Z. Yang, "Deformable structure from motion by fusing visual and inertial measurement data," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4816–4821.

[8] H. J. Marcus, C. J. Payne, A. Hughes-Hallett, G. Gras, K. Leibrandt, D. Nandi, and G.-Z. Yang, "Making the leap: the translation of innovative surgical devices from the laboratory to the operating room," *Annals of surgery*, vol. 263, no. 6, p. 1077, 2016.

[9] R. Lagneau, A. Krupa, and M. Marchal, "Active deformation through visual servoing of soft objects," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8978–8984.

[10] S. Giannarou, M. Ye, G. Gras, K. Leibrandt, H. 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, 03 2016.

[11] H. Zhou and J. Jagadeesan, *Real-Time Surface Deformation Recovery from Stereo Videos*, 10 2019, vol. 11764, pp. 339–347.

[12] D. Ghosh, A. Sharf, and N. Amenta, "Feature-driven deformation for dense correspondence," in *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*, vol. 7261. SPIE, 2009, pp. 985–994.

[13] G. A. Puerto-Souza and G.-L. Mariottini, "Wide-baseline dense feature matching for endoscopic images," in *Image and Video Technology: 6th Pacific-Rim Symposium, PSIVT 2013, Guanajuato, Mexico, October 28-November 1, 2013. Proceedings 6*. Springer, 2014, pp. 48–59.

[14] 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.

[15] A. Mendizabal, E. Tagliabue, J.-N. Brunet, D. Dall'Alba, P. Fiorini, and S. Cotin, "Physics-based deep neural network for real-time lesion tracking in ultrasound-guided breast biopsy," in *Computational Biomechanics for Medicine: Solid and Fluid Mechanics for the Benefit of Patients 22*. Springer, 2020, pp. 33–45.

[16] Y. Salehi and D. Giannacopoulos, "Physgnn: A physics-driven graph neural network based model for predicting soft tissue deformation in image-guided neurosurgery," *arXiv preprint arXiv:2109.04352*, 2021.

[17] A. Bozic, M. Zollhofer, C. Theobalt, and M. Nießner, "Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7002–7012.

[18] R. Sundararaman, R. Marin, E. Rodolà, and M. Ovsjanikov, "Reduced representation of deformation fields for effective non-rigid shape matching," *arXiv preprint arXiv:2211.14604*, 2022.

[19] C. N. Riviere, J. Gangloff, and M. De Mathelin, "Robotic compensation of biological motion to enhance surgical accuracy," *Proceedings of the IEEE*, vol. 94, no. 9, pp. 1705–1716, 2006.

[20] R. Richa, A. P. Bo, and P. Poignet, "Motion prediction for tracking the beating heart," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 3261–3264.

[21] T. Ortmaier, M. Groger, D. H. Boehm, V. Falk, and G. Hirzinger, "Motion estimation in beating heart surgery," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 10, pp. 1729–1740, 2005.

[22] A. Schoob, D. Kundrat, L. A. Kahrs, and T. Ortmaier, "Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery," *Medical image analysis*, vol. 40, pp. 80–95, 2017.

[23] T. Collins, A. Bartoli, N. Bourdel, and M. Canis, "Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos," 10 2016, pp. 404–412.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[26] C. Chen, C. Qin, C. Ouyang, Z. Li, S. Wang, H. Qiu, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Enhancing mr image segmentation with realistic adversarial data augmentation," *Medical Image Analysis*, vol. 82, p. 102597, 2022.

[27] H. Qiu, C. Qin, A. Schuh, K. Hammernik, and D. Rueckert, "Learning diffeomorphic and modality-invariant registration using b-splines," in *Medical Imaging with Deep Learning*, 2021.

[28] P. A. Yushkevich, A. H. Aly, J. Wang, L. Xie, R. C. Gorman, A. M. Pouch, and L. Younes, "Diffeomorphic medial modeling," *CoRR*, vol. abs/1902.02371, 2019. [Online]. Available: http://arxiv.org/abs/1902.02371

[29] S. Sun, K. Han, D. Kong, H. Tang, X. Yan, and X. Xie, "Topology-preserving shape reconstruction and registration via neural diffeomorphic flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 845–20 855.

[30] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-euclidean framework for statistics on diffeomorphisms," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 924–931.

[31] H. Wu, A. C. Sankaranarayanan, and R. Chellappa, "In situ evaluation of tracking algorithms using time reversed chains," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[32] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2756–2759.

[33] P. Pan, F. Porikli, and D. Schonfeld, "Recurrent tracking using multifold consistency," in *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, vol. 3, 2009.

[34] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros, "Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1191–1200.

[35] 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.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[37] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[38] Y. Li, N. Xu, J. Peng, J. See, and W. Lin, "Delving into the cyclic mechanism in semi-supervised video object segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1218–1228, 2020.

[39] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.

[40] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6197–6206.

[41] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[42] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, "What matters in unsupervised optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 557–572.

[43] 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.

[44] 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.

[45] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3d-guided cycle consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 117–126.

[46] D. Rivoir, M. Pfeiffer, R. Docea, F. Kolbinger, C. Riediger, J. Weitz, and S. Speidel, "Long-term temporally consistent unpaired video translation from simulated surgical 3d data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3343–3353.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[48] A. Torralba, B. C. Russell, and J. Yuen, "Labelme: Online image annotation and applications," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1467–1484, 2010.

[49] M. Contributors, "Openmmlab semantic segmentation toolbox and benchmark," 2020.

[50] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[51] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 729–738.

[52] A. L. deSouza, L. M. Prasad, J. J. Park, S. J. Marecik, J. Blumetti, and H. Abcarian, "Robotic assistance in right hemicolectomy: is there a role?" *Diseases of the colon & rectum*, vol. 53, no. 7, pp. 1000–1006, 2010.

[53] M. Penna, C. Cunningham, and R. Hompes, "Transanal total mesorectal excision: why, when, and how," *Clinics in colon and rectal surgery*, vol. 30, no. 05, pp. 339–345, 2017.

[54] 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.

[55] C. I. Nwoye, D. Alapatt, T. Yu, A. Vardazaryan, F. Xia, Z. Zhao, T. Xia, F. Jia, Y. Yang, H. Wang *et al.*, "Cholectriplet2021: A benchmark challenge for surgical action triplet recognition," *Medical Image Analysis*, vol. 86, p. 102803, 2023.

[56] B. van Amsterdam, I. Funke, E. Edwards, S. Speidel, J. Collins, A. Sridhar, J. Kelly, M. J. Clarkson, and D. Stoyanov, "Gesture recognition

in robotic surgery with multimodal attention," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1677–1687, 2022.

[57] 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.

[58] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo *et al.*, "The saras endoscopic surgeon action detection (esad) dataset: challenges and methods," *arXiv preprint arXiv:2104.03178*, 2021.

[59] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," 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. 627–636.

[60] R. Ma, E. B. Vanstrum, J. H. Nguyen, A. Chen, J. Chen, and A. J. Hung, "A novel dissection gesture classification to characterize robotic dissection technique for renal hilar dissection," *The Journal of Urology*, vol. 205, no. 1, pp. 271–275, 2021.

[61] R. Elek, T. D. Nagy, D. Á. Nagy, T. Garamvölgyi, B. Takács, P. Galambos, J. K. Tar, I. J. Rudas, and T. Haidegger, "Towards surgical subtask automation—blunt dissection," in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*. IEEE, 2017, pp. 000 253–000 258.

[62] D. Á. Nagy, T. D. Nagy, R. Elek, I. J. Rudas, and T. Haidegger, "Ontology-based surgical subtask automation, automating blunt dissection," *Journal of Medical Robotics Research*, vol. 3, no. 03n04, p. 1841005, 2018.

[63] 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.

[64] A. Pore, E. Tagliabue, M. Piccinelli, D. Dall'Alba, A. Casals, and P. Fiorini, "Learning from demonstrations for autonomous soft-tissue retraction," in *2021 International Symposium on Medical Robotics (ISMR)*. IEEE, 2021, pp. 1–7.

[65] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on graphics (TOG)*, vol. 23, no. 3, pp. 399–405, 2004.

[66] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[67] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski, "Smurf: Self-teaching multi-frame unsupervised raft with full-image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3887–3896.