2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Semantic-SuPer: A Semantic-aware Surgical Perception Framework for Endoscopic Tissue Identification, Reconstruction, and Tracking

Shan Lin[1], Albert J. Miao[1], Jingpei Lu[1], Shunkai Yu[1], Zih-Yun Chiu[1], Florian Richter[1], Michael C. Yip[1], *Senior Member, IEEE*

***Abstract*— Accurate and robust tracking and reconstruction of the surgical scene is a critical enabling technology toward autonomous robotic surgery. Existing algorithms for 3D perception in surgery mainly rely on geometric information, while we propose to also leverage semantic information inferred from the endoscopic video using image segmentation algorithms. In this paper, we present a novel, comprehensive surgical perception framework, Semantic-SuPer, that integrates geometric and semantic information to facilitate data association, 3D reconstruction, and tracking of endoscopic scenes, benefiting downstream tasks like surgical navigation. The proposed framework is demonstrated on challenging endoscopic data with deforming tissue, showing its advantages over our baseline and several other state-of-the-art approaches. Our code and dataset are available at https://github.com/ucsdarclab/Python-SuPer.**

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160746

I. Introduction

With the maturation of surgical robotic platforms like the da Vinci® Surgical System (Intuitive Surgical, Inc., Sunnyvale, CA) well underway, there has been growing interest for these robot platforms to incorporate increased intelligence towards understanding the anatomy of the operative scene. 3D scene understanding of both geometric and semantic features of anatomy enables more effective navigation in patients and is an important step towards the automation of surgical tasks in the future. Current surgical navigation systems generally utilize segmented preoperative images, such as CT/MRI scans, to provide 3D semantics information. Specifically, the 3D semantic segmentation estimated from the preoperative data serves as a map, unchanged throughout the surgery, and matched with the current scene through registration approaches [1], [2], [3], [4], [5]. However, whenever there is a large deformation due to surgical operations or different pre- and intra-operative body positioning, the map becomes less reliable, thereby limiting the applications of existing navigation systems to deformable surgical scenes.

In contrast, video semantic segmentation has the potential to extract precise anatomical information in real-time to update the navigation map. Existing works have shown the benefits of integrating semantic information extracted from videos with 3D geometry for perception in indoor dynamic environments and autonomous driving [6], [7], [8], [9], [10], [11], [12], [13]. Different from environments addressed in these works, surgical scenes usually involve textureless tissue surfaces and constantly deforming tissues. These features can cause difficulty in data association, which is a process to find the correspondences between the sensory inputs and the tracked models. Several works have investigated data association approaches, including pixel-level feature matching and dense photometric loss, but still show that this remains a challenge. We aim to alleviate this problem from a different perspective: using semantics to guide data association. We propose a cost function named "morphing loss" that explicitly encourages the border consistency between the semantic segmentation from the current video frame and the 3D model by projecting its semantic information to the image plane.

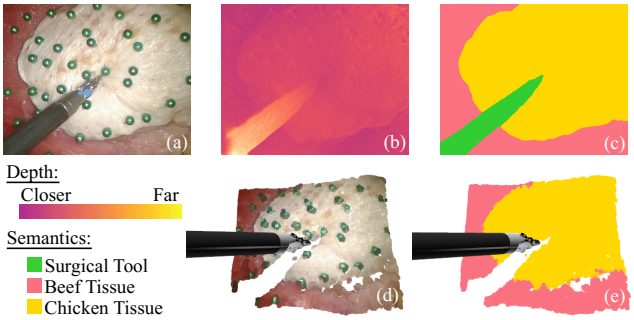


Fig. 1. **A demonstration of Semantic-SuPer.** (a) Input video frame. (b) Depth and (c) semantic segmentation map estimated from the input. (d) Scene rendered from the tracked surfels and surgical tool pose. (e) Scene rendered from surfels visualized by colors corresponding to their semantics.

In this work, we propose a novel semantic-aware surgical perception framework, Semantic-SuPer, that combines semantics from 2D videos with 3D deformation tracking and reconstruction. This framework builds upon SuPer [14], [15] by integrating kinematic model-based surgical tool tracking, model-free deformable tissue tracking, and a robotic arm control loop with semantics, increasing robustness and improving accuracy. The main contributions are as follows:

- A perception framework that integrates semantic information from videos with the 3D geometric properties of the surgical environment for better scene understanding.
- A novel morphing loss that explicitly constrains the semantic consistency between the new input data and the 3D model, showing the benefits of including semantic information for surgical perception.
- An investigation on the influence of semantic segmentation accuracy on the endoscopic video datasets.
- A released robotic tissue manipulation dataset for semantic-aware 3D tracking and reconstruction, collected using the da Vinci Research Kit (dVRK) [16].

[1]S. Lin, A.J. Miao, J. Lu, S. Yu, Z.Y. Chiu, F. Richter, and M.C. Yip are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA. (e-mail: {shl10, amiao, jil360, shyu, zchiu, frichter, yip}@ucsd.edu) This project was funded by the Telemedicine and Advanced Technology Research Center (TATRC) via award MTEC-21-06-MPAI-004 as well as NSF Award #2045803. F. Richter was supported on an NSF Graduate Research Fellowship.

II. Related Work

Endoscopic tissue deformation tracking and semantic segmentation are two critical components for intelligent assistance in surgery. Because the integration of these two tasks in endoscopic scenes remains under explored, we mainly focus on related work for each individual topic.

Surgical Scene Semantic Segmentation aims to divide a surgical image into regions of different tissues and tools. Motivated by the developments in deep learning, various networks, especially Convolutional Neural Networks (CNNs), have achieved state-of-the-art segmentation performance for endoscopic images [17], [18], [19], [20], [21], [22], [23]. Most existing works focus on developing segmentation algorithms, while the exploration of their potential applications is still limited. In contrast, we focus on integrating the semantic information with surgical scene tracking in the 3D space.

Endoscopic Tissue Tracking is a specific area of non-rigid tracking, where the low textured, deformable tissue is a significant challenge. Earlier works typically track the scenes based on rigidness assumptions [24], [25], [26], thus their performance degrades when dealing with large motions. MIS-SLAM [27] described the deformation using embedded deform nodes and achieved better tracking. Also, there is another trend of works that model deformations with models like mesh and spline, based on the assumption that the tissue surface is continuous [28], [29], [30], [31]. Yet, the complex surgical scenes still make the data association a challenge even for state-of-the-art methods [31], [32], which led to exploration on stronger features as well as photometric losses that impose dense constraints to this process [31], [32], [33]. Compared to the approaches mentioned above, we aim to build a comprehensive framework that integrates surgical scene tracking and reconstruction with semantic segmentation and show that semantic information can provide guidance for data association.

Semantic SLAM refers to approaches that combine the geometry of the world estimated by SLAM with object detection or semantic segmentation results. Semantic SLAM has been widely studied for areas including autonomous driving, where semantic information has been used to select features or regions of interest, improve data association, identify and remove dynamic points for mapping, and assist long-term localization [6], [8], [9], [10], [11], [13]. Yet, methods that combine 3D tracking and reconstruction with semantic segmentation is still very limited for endoscopic data. To the best of our knowledge, there are only two works involving endoscopic data, using a binary mask to segment surgical tools from tissue backgrounds [34], [35]. In contrast, we consider segmentation information of the whole surgical scene, including different types of tissues.

III. Methods

The proposed approach builds upon the surgical perception framework SuPer for tissue manipulation [14], [15] by including pixel-wise semantic labels of the endoscopic videos as inputs and building *surface elements (surfels)* [37], [38], [39] with semantic information, as shown in Figure 2. We investigate three popular deep segmentation models to provide the semantic labels, while other segmentation architectures can also be used. This semantic information is used to suppress erroneous data association between different classes, and therefore improves the robustness of tissue tracking. Furthermore, this allows us to build a 3D surfel map that contains semantic information of different anatomies in deformable surgical scenes, which could benefit endoscopic surgical navigation and other related tasks in the future.

A. SuPer Framework

SuPer tracks the geometric information of the entire surgical scene, including both the tools controlled by the surgical robot and the deforming tissues. The tool tracking is achieved by a particle filter that utilizes kinematic modeling and information from the endoscopic video. For more details on tool tracking, refer to Richter's *et al.* previous work [40]. Meanwhile, the surgical environment (i.e. soft tissue) is tracked using a model-free method based on surfel representation. Each surfel $\mathcal{S}$ is defined by a position $\mathbf{p}_i \in \mathbb{R}^3$, a normal $\mathbf{n}_i \in \mathbb{R}^3$, a color $\mathbf{c}_i \in \mathbb{R}^3$, a radius $\mathbb{r}_i \in \mathbb{R}$, a confidence score $\mathbb{c}_i \in \mathbb{R}$, and a time stamp $\mathbb{t}_i \in \mathbb{N}$ of its last update. One can refer to [38], [39] for details on adding, fusing, and deleting surfels.

The number of surfels is proportional to the number of image pixels, so tracking each surfel individually requires a large number of parameters. As inspired by [41], SuPer introduces the Embedded Deformation (ED) graph with vertices that are much sparser than the surfels to drive the motion of the surfel set. The ED graph is given by $\mathcal{G}_{ED} = \{\mathcal{V}, \mathcal{E}, \mathcal{P}\}$, where $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is the set of edges, and $\mathcal{P}$ is the set of parameters. Each vertex (ED node) contains $(\mathbf{g}_j, \mathbf{q}_j, \mathbf{b}_j) \in \mathcal{P}$, where $\mathbf{g}_j \in \mathbb{R}^3$ is its position, $\mathbf{q}_j \in \mathbb{R}^4$ and $\mathbf{b}_j \in \mathbb{R}^3$ are the quaternion and translation parameters, respectively. The position and normal of each surfel is then updated according to

$$\widetilde{\overline{\mathbf{p}}}_i = \mathbf{T}_g \sum_{j \in \mathcal{N}_i} \omega_j(\mathbf{p}_i)[T(\mathbf{q}_j, \mathbf{b}_j)(\overline{\mathbf{p}}_i - \vec{\mathbf{g}}_j) + \vec{\mathbf{g}}_j] \quad (1)$$

$$\widetilde{\vec{\mathbf{n}}}_i = \mathbf{T}_g \sum_{j \in \mathcal{N}_i} \omega_j(\mathbf{p}_i)[T(\mathbf{q}_j, 0)\vec{\mathbf{n}}_i] \quad (2)$$

where $\mathbf{T}_g \in SE(3)$ is the global homogeneous transformation matrix that is shared by all surfels, $T(\cdot) \in SE(3)$ is the local homogeneous transform matrix from a ED node, $\bar{\cdot}$ and $\vec{\cdot}$ are the homogeneous representations of a point and motion (*i.e.* $\overline{\mathbf{p}} = [\mathbf{p}, 1]^T$ and $\vec{\mathbf{g}} = [\mathbf{g}, 0]^T$), $\mathcal{N}_i$ is the set of k-nearest neighbors of $\mathbf{p}_i$ in $\mathcal{G}_{ED}$. $\omega_j(\mathbf{p}_i)$ is a weight that indicates the influence of $\mathbf{g}_i$ to $\mathbf{p}_i$ and is calculated as $\omega_j(\mathbf{p}_i) = e^{-\|\mathbf{p}_i - \mathbf{g}_j\|}$ and then normalized to sum to one within $\mathcal{N}_i$. (1) and (2) can be interpreted as transforming the surfels by averaging the motions of their neighboring ED nodes. For every new image frame received, a total of $7 \times (n+1)$ parameters ($\mathbf{q}_j$, $\mathbf{b}_j$, and $\mathbf{T}_g$), where n is the total number of ED nodes, will be estimated. The estimated parameters are then committed (*i.e.*, updating all ED nodes transformations and all surfels based on (1) and (2)) to track the deformations of the reconstructed tissue.

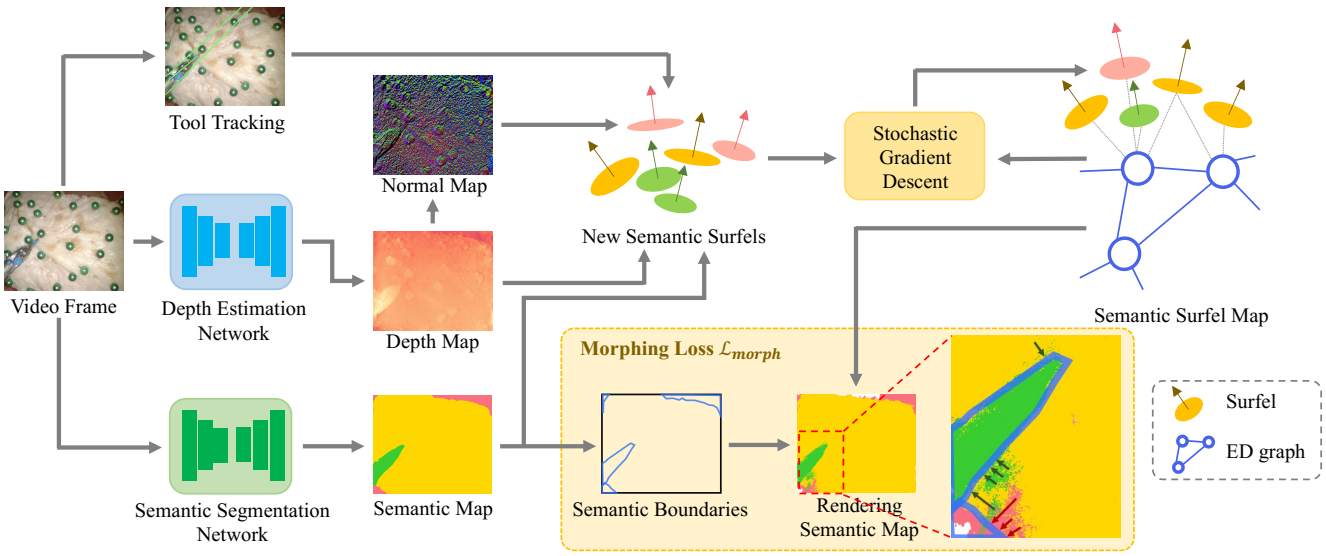


Fig. 2. **Overview of the proposed framework.** The depth, semantics, and tool poses are extracted from the input video frame. The transformations of ED nodes are optimized to match the observations under several constraints, including a semantics-based morphing loss. The optimization is implemented using stochastic gradient descent [36] with PyTorch's automatic differentiation. Surfel position and normal updates are then controlled by the ED nodes.

B. Inputs to the Framework: Depth, Normals, and Semantics

As we found previously [14], [15], the noise in the input depth map is a major factor that limits the performance of the proposed perception framework. Thus, in SuPer-Deep [15] we did a comprehensive comparison showing that deep learning depth estimation models pre-trained on other datasets, such as the KITTI benchmark [42], led to better tissue tracking performance than traditional stereo matching algorithms. Yet, without finetuning the deep learning models to the surgical dataset, their estimations are still noisy. For example, the SuPer-Deep framework with PSMNet pretrained on KITTI [43] crashes within the first 40 image frames from the datasets collected in this work due to the noisy depth observations. Furthermore, there is not a sufficient amount of surgical data with ground truth depth maps to train supervised stereo matching algorithms.

In this work, we use Monodepth2 which can be trained in a self-supervised fashion [44] . Self-supervision is achieved by maximizing the similarity between the real left image and the reconstructed left image, which is warped from the right view using the estimated depth and camera intrinsics and extrinsics. From the depth map, we estimate the surface normal at each pixel as the average of the cross products of all pairs of vectors that point to its 8 neighboring pixels [45]. The surfel set is initialized from the first depth and normal maps. ED nodes are initialized by sampling uniformly in a rectangle mesh grid in the image, similar to [46], and the corresponding node positions $\mathbf{g}$ are extracted from the point cloud. The ED graph edges are then initialized by connecting each node with its 8 neighbors.

Meanwhile, the deep learning based segmentation model (see Section IV-C.2) is utilized to predict the segmentation map of each frame. For each surfel, its semantic confidence scores $\mathbf{s}_i \in \mathbb{R}^C$ (C is the number of classes) are initialized as the corresponding softmax outputs of the network and its semantic label $\mathrm{y}_i \in \mathbb{R}$ is the class that corresponds to the highest softmax score.

C. Semantic-aware Registration

After the surfels and ED nodes are initialized, they are updated according to new observed depth, normal, and segmentation maps. The update is applied by solving

$$\underset{\mathbf{q},\mathbf{b},\mathbf{T}_g}{\arg\min}\, \lambda_s \mathcal{L}_{sim} + \lambda_m \mathcal{L}_{morph} + \lambda_r \mathcal{L}_{reg} \tag{3}$$

where $\mathcal{L}_{sim}$ the similarity metric that measures the similarity between the transformed model and input data based on data association, $\mathcal{L}_{morph}$ is the proposed semantic-aware morphing loss, $\mathcal{L}_{reg}$ is the regularization term, and $\lambda_m, \lambda_s, \lambda_r$ are hyper-parameters. The coming subsections detail explicit expressions for each loss term and (3) is solved in a similar manner as the original SuPer [14].

1) Data Association Approaches: Data association aims to find the correspondences between the tracked data and new observations (i.e. depth, normal, and semantic maps) which are then used to optimize the transformations. The first loss is the same point-to-plane ICP loss [47] as in SuPer [14]:

$$\mathcal{L}_{icp} = \sum_i \rho_{i,o} (\vec{\mathbf{n}}_o^T (\widetilde{\mathbf{p}}_i - \overline{\mathbf{p}}_o))^2 \tag{4}$$

where $\overline{\mathbf{p}}_o$ and $\vec{\mathbf{n}}_o$ are the observed positions and normals, bilinearly sampled [48] from the observations at the projected pixel coordinates of $\widetilde{\mathbf{p}}_i$ and $\rho_{i,o}$ is the weight for each term. The original SuPer uses the naive ICP loss, so $\rho_{i,o} = 1$ [14]. For Semantic-SuPer, the weight is computed based on the Jensen–Shannon divergence [49] between the semantic softmax confidences of the surfel and the observation, *i.e.*, $\rho_{i,o} = \exp^{-JSD(\mathbf{s}_i \| \mathbf{s}_o)}$. The second term utilizes Pulsar [50], a real-time differentiable renderer, to compute a loss directly between a rendering of the tracked soft tissue and the raw

image:

$$\mathcal{L}_{render} = \frac{1}{N}\sum_{i=0}^{N}\left\|\frac{1 - SSIM(I_i, \mathcal{R}(\mathcal{S}; \mathbf{q}, \mathbf{b}, \mathbf{T}_g, \mathbf{K})_i)}{2}\right\|^2 \tag{5}$$

where $SSIM(\cdot)$ is the structural similarity index measure [51], N is the number of pixels in the image, I_i denotes the ith pixel of image I, $\mathcal{R}(\cdot)$ is the Pulsar renderer, and $\mathbf{K}$ is the camera intrinsic parameters.

2) Semantic-aware Morphing Loss: The data-association losses, $\mathcal{L}_{sim}$, are restricted by the locality of their gradients, which can lead to incorrect associations when the scene is undergoing large deformations. Therefore, we propose a semantic aware morphing loss that provides longer-range hints by explicitly restricting the semantic edge consistency between the tracked surfel map and every new input semantic map. The morphing loss, $\mathcal{L}_{morph}$, minimizes the distance of surfels whose projections onto the image fall outside of their own semantic boundary (see Figure 2) and the semantic boundary:

$$\mathcal{L}_{morph} = \sum_{\pi(\mathbf{p}_i)\notin\mathcal{R}_i} \min_{\mathbf{o}\in\mathcal{B}_i} ||\pi(\mathbf{p}_i) - \mathbf{o}||^2 \tag{6}$$

where $\pi(\cdot)$ projects points from the 3D space to the image plane, $\mathcal{R}_i$ is the corresponding semantic region that the 2D projection of $\mathbf{p}_i$ should lie in, and $\mathcal{B}_i$ is the set of coordinates of boundary pixels of $\mathcal{R}_i$. Note that we minimize the distance in the image plane instead of in the 3D space because the estimated depths around the object boundaries are naturally quite noisy due to the background occlusion [52].

3) Regularization term: The regularization term consists of two terms. The first term, $\mathcal{L}_{face}$, ensures all ED nodes move as rigidly as possible, which helps the model resist incorrect guidance from noisy inputs, and more importantly, provides hints to control some ED nodes that do not receive enough information from observation (*e.g.*, due to occlusion) [53], [54], [14]. As inspired by [55], $\mathcal{L}_{face}$ is the l_2 norm of the difference between the area of all the transformed triangle surfaces in the ED graph:

$$\mathcal{L}_{face} = \sum_{e_{ij}\in\mathcal{E}, e_{ik}\in\mathcal{E}, j\neq k} \mathbb{I}_{tri}\left\|\frac{1}{2}|e_{ij}\times e_{ik}| - A_{ijk}\right\|^2 \tag{7}$$

where e_{ij} denotes the edge that connects the i-th and j-th ED nodes. For each group of three edges that form an triangle, the indicator is set to 1, *i.e.*, $\mathbb{I}_{tri} = 1$, and A_{ijk} is the initial area of this triangle; otherwise $\mathbb{I}_{tri} = 0$. The second term, $\mathcal{L}_{Rot}$, is the quaternion normalization term adopted from SuPer to ensure the quaternions hold $\|\mathbf{q}\|^2 = 1$:

$$\mathcal{L}_{Rot} = \sum_{k} \|1 - \mathbf{q}_k^T\mathbf{q}_k\|^2 \ . \tag{8}$$

IV. Experiments and Results

A. Experimental Setup

The proposed semantic-aware surgical perception framework was deployed on the da Vinci Research Kit (dVRK) [16], [56] for evaluation. As shown in Figure 3, we overlaid a piece of chicken meat across a piece of beef. The green pins

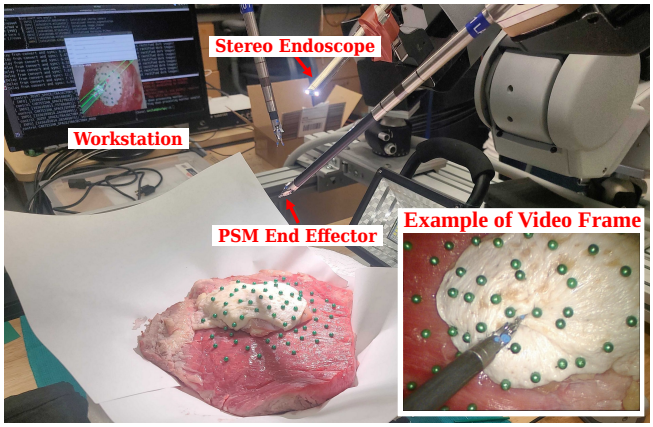


Fig. 3. Experimental setup with the dVRK system.

attached on the tissue surface were used to collect ground truth for quantitative evaluation of tracking (Section IV-B). To generate deformations in the scene, the beef was pushed up-and-down manually from below. Meanwhile, the dVRK was used to control a single surgical robotic arm, *i.e.*, the Patient Side Manipulator (PSM), to grasp and tug the chicken tissue. The Robot Operating System (ROS) was used to handle all communication between subsystems, which ran on the Intel i9-7940X and Nvidia RTX 2080 GPU. 4 trials were conducted, each consisting of 150 frames, named Lab 1, 2, 3, and 4 in the following sections. Raw stereo videos were recorded at 1080p, 30 fps, and then rectified to 640×480 resolution. For each trial, timestamped joint angles, gripper poses, and rectified video streams were saved.

B. Metrics and Ground Truth

To quantitatively evaluate the tracking results, we attached roughly 60 green pins onto the tissue surface (see Figure 3), covering important areas such as those with larger deformations and the boundaries of tissues, and extracted their trajectories in the image plane throughout each trial. The pin trajectories were obtained through a tracking-by-detection paradigm, where detection was achieved by identifying green circular regions in the Hue-Saturation-Value (HSV) color space of the image. Because the videos were captured at different distances to the scene in the 4 trials, the number of green pins that appear in the videos ranges from 30 to 60. The tracked surfels were then projected to the image plane for calculating the reprojection errors, i.e., the distances between the surfel reprojections and their corresponding ground truth.

C. Implementation Details

1) Depth Estimation: Monodepth2 is pre-trained with a larger surgical dataset, the Hamlyn dataset [57], and then finetuned with our data. We find that the generalizability of existing self-supervised deep depth estimation models including Monodepth2 is limited when applied to our data, which we believe is because 1) the domain gap between the Hamlyn dataset and our data is not ignorable and we do not have sufficient frames for finetuning, and 2) the distances from the scene to the camera vary a lot between different trials. Improving depth estimation is not our focus here, so

TABLE I

REPROJECTION ERROR COMPARISON ON OUR DATASET.

Method	Data			
	Lab1	Lab2	Lab3	Lab4
DefSLAM [30]	16.5(12.5), 14.5(11.3)	14.5(13.2), 15.6(13.8)	12.8(8.8), 13.0(8.5)	7.0(5.2), 7.3(5.4)
SD-DefSLAM [31]	8.5(8.5), **5.2(4.6)**	**8.0(9.4)**, 10.3(12.0)	8.4(9.1), 7.1(7.3)	**3.8(4.0)**, **3.1(3.6)**
SuPer [14]	10.8(8.8), 8.6(7.0)	10.8(10.1), 10.2(9.1)	8.1(6.7), 7.2(5.9)	4.8(4.3), 4.1(3.2)
NoSoftLabel-Semantic-SuPer	12.7(9.5), 9.1(7.4)	10.8(9.7),10.4(9.6)	8.9(7.3),7.3(6.0)	7.1(8.0),13.2(17.6)
Semantic-SuPer	**7.5(6.1)**, 6.7(5.7)	8.6(7.6), **9.2(7.8)**	**6.0(4.9)**, **5.9(4.8)**	4.3(3.8), 4.3(3.4)

* From left to right, the two metrics of each data are the reprojection errors averaged over all points and over points near object boundaries, respectively. The errors are formatted as 'mean(standard deviation)'. The best result in each row is in **bold**.

TABLE II

ABLATION STUDIES OF SEMANTIC-SUPER.

Method	Cost Functions		Data			
	$\mathcal{L}_{morph}$	$\mathcal{L}_{render}$	Lab1	Lab2	Lab3	Lab4
SuPer[14]	-	✗	10.8(8.8), 8.6(7.0)	10.8(10.1), 10.2(9.1)	8.1(6.7), 7.2(5.9)	4.8(4.3), 4.1(3.2)
	-	✓	10.9(9.1), 8.8(7.1)	9.9(9.0), **8.7(7.2)**	8.8(7.2), 8.3(6.9)	4.7(4.2), 4.0(3.4)
Semantic-SuPer	✗	✗	10.0(8.3), 8.5(7.1)	10.6(9.9), 10.1(9.0)	8.7(7.1), 8.6(7.1)	4.7(4.1), 4.0(3.3)
	✓	✗	7.7(6.3), **6.4(5.4)**	9.4(8.5), 9.6(8.4)	6.3(5.0), 6.1(5.0)	4.5(3.9), 4.4(3.6)
	✗	✓	10.2(8.3), 9.5(7.5)	10.2(9.4), 9.4(8.1)	8.2(6.6), 8.5(7.0)	4.7(4.1), **3.9(3.1)**
	✓	✓	**7.5(6.1)**, 6.7(5.7)	**8.6(7.6)**, 9.2(7.8)	**6.0(4.9)**, **5.9(4.8)**	**4.3(3.8)**, 4.3(3.4)

* Refer to the note of Table I.

we finetune the model without a train-test split. For both pre-training and finetuning, Monodepth2 is trained for 20 epochs using the Adam optimizer [58], with a batch size of 16, a learning rate of 10^{-4} for the first 15 epochs and 10^{-5} for the remainder. The estimated depth map is post-processed by a widely adopted method that fuses the depth maps for the target image and its horizontally flipped image [59].

2) Semantic Segmentation: We compare the influence of three segmentation algorithms, DeepLabv3+ [60], U-Net [61], and UNet++ [62], on the performance of Semantic-SuPer. These models are trained for 50 epochs using Adam optimizer, with a batch size of 16 and an initial learning rate of 10^{-4} which step decays by 0.1 for every 16 epochs. The models are trained under a K-fold cross-validation setup with $K = 4$. At each split, three trials were used to train the model, which was then directly applied to the remaining trial for evaluating Semantic-SuPer.

3) Deformable Tracking: The methods for initializing surfel radius and adding surfels and ED nodes are in the same manner as SuPer [14]. To initialize the ED graph, since the depths vary a lot between trials, instead of using a fixed step size to generate the mesh, we choose the step size for each trial by ensuring the average edge length of the graph is around 5mm. Finally, the hyperparameters that control the cost functions are chosen as $\lambda_m = 10$, $\lambda_s = 1$, and $\lambda_r = 10$.

D. Comparisons with the State-of-the-art Methods

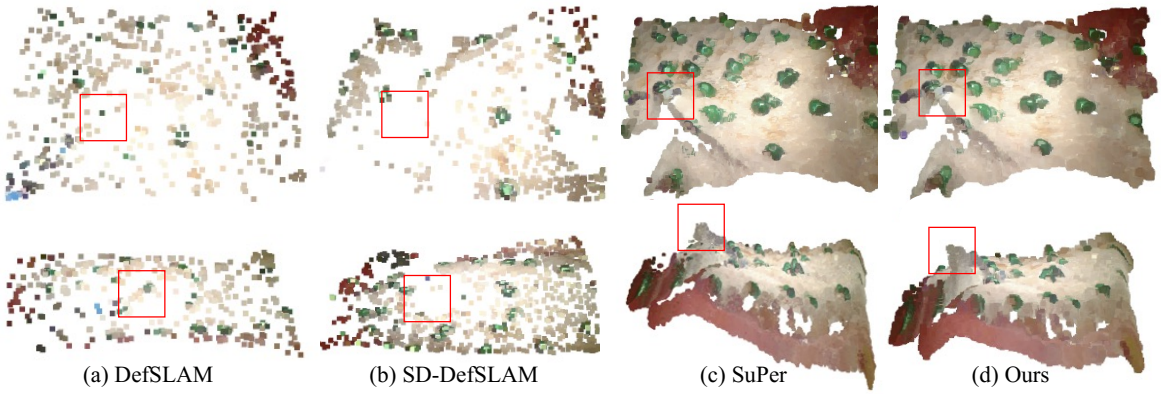
As shown in Table I, we compare Semantic-SuPer against two baselines: 1) SuPer [14], and 2) NoSoftLabel-Semantic-SuPer that does not consider the semantic confidence score, only connects surfels and ED nodes that belong to the same class, and uses naive ICP metric calculated from pairs of surfels from the same class. We also evaluate the performance of SOTA deforming surgical scene tracking and reconstruction algorithms DefSLAM [30] and SD-DefSLAM [31]. DefSLAM and SD-DefSLAM track the scene based on relatively sparse feature matching and may not track the labeled points, so we estimated the flow of a certain labeled point by averaging the flows of its 3 nearest neighbors. We also conduct an ablation study in Table II to evaluate the effectiveness of the proposed modules on top of SuPer.

Table I shows that Semantic-SuPer outperforms the baselines on Lab 1-3. Lab 4 has the furthest distance between the camera and the scene so it has minimum deformation and changes in the segmentation map, which leads to the small performance gap between Semantic-SuPer and SuPer on it. Moreover, our framework outperforms DefSLAM on each of the videos, while achieving either comparable or better performance than SD-DefSLAM. DefSLAM and SD-DefSLAM use matching algorithms based on sparse image features so they could achieve low reprojection errors by selecting more robust features. Yet, because the data was collected by a stationary camera, these two algorithms are unable to reconstruct the 3D surfaces well, while our approach uses monocular depth estimation techniques and thus can provide more accurate and dense tracking, as shown in Figure 4.

Figure 5 shows the average reprojection error of each video frame in Lab 1. The large deformation happens around the 60th to the 110th frames; during this period we can see larger gaps between DefSLAM, SuPer and Semantic-SuPer. SD-DefSLAM has a much parse detection in the region with a larger deformation, *i.e.*, the grasping point shown by the red rectangle in Figure 4, so it does not follow the same pattern as the other three methods. It should be noted that the detections near the grasping point are very important because the robot requires this information to perform autonomous manipulation, which is one of our final goals.

E. Influence of different segmentation methods

A comparison of the influence of different segmentation algorithms on Semantic-SuPer is shown in Table III. We measure the quality of the predicted segmentation maps using Hausdorff distance (HD), which indicates the largest



(a) DefSLAM (b) SD-DefSLAM (c) SuPer (d) Ours

Fig. 4. Comparison of the tracked point cloud with SOTA methods. The tissue grasping point is in the red rectangle.

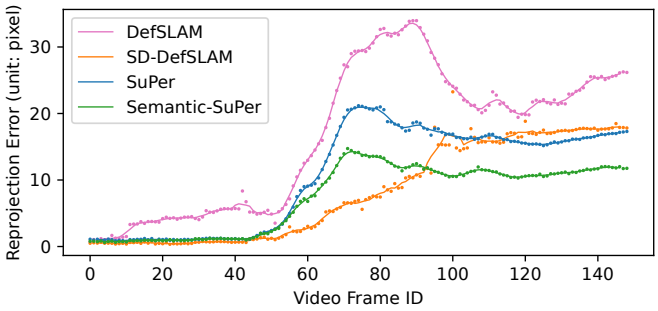


Fig. 5. Average reprojection error through time in Lab 1.

TABLE III

COMPARISON STUDIES OF THE INFLUENCE OF SEGMENTATION QUALITY ON SEMANTIC-SUPER.

Data		Segmentation Method		
		DeepLabV3+	UNet	UNet++
Lab1	HD(pixel)	**124.6**	182.7	205.3
	F_1(%)	96.3	96.6	**96.9**
	Reproj. Err.	7.5(6.1), **6.7(5.7)**	7.3(6.0), 7.2(5.9)	**7.3(5.9)**, 7.4(5.9)
Lab2	HD(pixel)	**155.6**	164.9	181.9
	F_1(%)	97.2	97.3	**97.6**
	Reproj. Err.	**8.6(7.6)**, 9.2(7.8)	11.3(10.9), 12.4(11.5)	9.0(8.1), **8.5(7.4)**
Lab3	HD(pixel)	**224**	369.6	313.3
	F_1(%)	97.5	97.7	**98.1**
	Reproj. Err.	**6.0(4.9)**, 5.9(4.8)	6.2(5.2), 5.7(4.8)	6.0(5.0), **5.5(4.6)**
Lab4	HD(pixel)	**107.3**	156.4	175
	F_1(%)	96.1	**96.8**	96.7
	Reproj. Err.	**4.3(3.8)**, 4.3(3.4)	**4.3(3.8)**, **3.7(2.9)**	4.6(3.9), 4.1(3.1)

* The best result in each row is in **bold**.
Refer to Table I for notes on the reprojection errors.

segmentation error, and F_1 score [63]. Table III shows a better performance is associated with a better segmentation, *i.e.*, lower HD or higher F1 score. Also, we find that it is more likely to observe bad trackings after a small region within a larger semantic area is incorrectly segmented as another class, which could be addressed by postprocessing the segmentation map using methods such as Conditional Random Fields (CRFs) [64].

V. DISCUSSION

Table I demonstrates the benefits of including semantics for tissue tracking, while Table II shows that the morphing loss augments the effectiveness of ICP and rendering loss. Specifically, the rendering loss may not improve ICP-based SuPer / Semantic-SuPer, but when using the morphing loss, the combination of ICP and rendering loss leads to better performance than ICP-based Semantic-SuPer. Moreover, it should be noted that the green pins, attached to tissues to allow quantitative evaluation, in fact, make tracking more challenging for our framework. With the pins, the tissue surfaces became uneven, but advanced self-supervised depth estimation methods usually rely on an assumption of the smoothness of the depth and lead to suboptimal depth maps on our data. In addition, we do not use image feature matching methods like the Lucas-Kanade tracker used in SD-DefSLAM [31] for data association and therefore do not take much advantage of the strong features from the green pins. Nevertheless, our results are still competitive.

The comparison between Semantic-SuPer and NoSoftLabel-Semantic-SuPer shows the benefits of considering the certainty of segmentation. NoSoftLabel-Semantic-SuPer achieves worse performance, because without using the soft semantic labels, the incorrect segmentations assign surfels to ED nodes that belong to other classes, and the estimations of the ED node transformations will be affected more by wrong associations between surfels belonging to different classes. Thus, adopting better uncertainty estimation methods for the semantics (also required for depth input) [65], [66], [67], as well as leveraging multi-task learning-based cross-task knowledge [68] to estimate uncertainty could lead to better tracking performance.

VI. CONCLUSIONS

In this paper, we present a novel surgical perception framework Semantic-SuPer that achieves better surgical scene 3D reconstruction and tracking by integrating semantic information, showing the benefits of including semantics in this task, which has not been well-explored in prior works. In the future, we will deploy our framework on endoscopic videos captured by moving cameras. We will also investigate multi-task learning to leverage useful information among depth, normal estimation, and semantic segmentation to improve these tasks, whose performance still limits our framework. Furthermore, we plan to extract cross-modality knowledge among these framework inputs to achieve better uncertainty estimation for more robust tracking. This perception uncertainty could further augment the combination of advanced control algorithms with our framework, aiming for surgical task automation.

REFERENCES

[1] S. Ieiri, M. Uemura, K. Konishi, *et al.*, "Augmented reality navigation system for laparoscopic splenectomy in children based on preoperative CT image using optical tracking device," *Pediatric Surgery Int.*, vol. 28, no. 4, pp. 341–346, 2012.

[2] M. Metzger, G. Bittermann, L. Dannenberg, *et al.*, "Design and development of a virtual anatomic atlas of the human skull for automatic segmentation in computer-assisted surgery, preoperative planning, and navigation," *Int. J. Comput. Assisted Radiol. Surgery*, vol. 8, no. 5, pp. 691–702, 2013.

[3] X. Chen, L. Xu, Y. Wang, *et al.*, "Development of a surgical navigation system based on augmented reality using an optical see-through head-mounted display," *J. Biomed. Inform.*, vol. 55, pp. 124–131, 2015.

[4] A. Teatini, E. Pelanis, D. L. Aghayan, *et al.*, "The effect of intraoperative imaging on surgical navigation for laparoscopic liver resection surgery," *Scientific Reports*, vol. 9, no. 1, pp. 1–11, 2019.

[5] H. Zhang, M. Shen, P. L. Shah, *et al.*, "Pathological airway segmentation with cascaded neural networks for bronchoscopic navigation," in *Proc. Int. Conf. Robot. Autom.*, pp. 9974–9980, 2020.

[6] J. Civera, D. Gálvez-López, L. Riazuelo, *et al.*, "Towards semantic SLAM using a monocular camera," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, pp. 1277–1284, 2011.

[7] A. Kundu, Y. Li, F. Dellaert, *et al.*, "Joint semantic segmentation and 3D reconstruction from monocular video," in *Proc. Europ. Conf. Comput. Vis.*, pp. 703–718, 2014.

[8] S. L. Bowman, N. Atanasov, K. Daniilidis, *et al.*, "Probabilistic data association for semantic slam," in *Proc. Int. Conf. Robot. Autom.*, pp. 1722–1729, 2017.

[9] L. Zhang, L. Wei, P. Shen, *et al.*, "Semantic SLAM based on object detection and improved octomap," *IEEE Access*, vol. 6, pp. 75 545–75 559, 2018.

[10] X. Chen, A. Milioto, E. Palazzolo, *et al.*, "SuMa++: Efficient LiDAR-based semantic SLAM," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, pp. 4530–4537, 2019.

[11] K. J. Doherty, D. P. Baxter, E. Schneeweiss, *et al.*, "Probabilistic data association via mixture models for robust semantic SLAM," in *Proc. Int. Conf. Robot. Autom.*, pp. 1098–1104, 2020.

[12] D. Menini, S. Kumar, M. R. Oswald, *et al.*, "A real-time online learning framework for joint 3D reconstruction and semantic segmentation of indoor scenes," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1332–1339, 2021.

[13] Y. Fan, Q. Zhang, Y. Tang, *et al.*, "Blitz-SLAM: A semantic SLAM in dynamic environments," *Pattern Recognition*, vol. 121, p. 108225, 2022.

[14] Y. Li, F. Richter, J. Lu, *et al.*, "SuPer: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2294–2301, 2020.

[15] J. Lu, A. Jayakumari, F. Richter, *et al.*, "SuPer Deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction," in *Proc. Int. Conf. Robot. Autom.*, pp. 4783–4789, 2021.

[16] P. Kazanzides, Z. Chen, A. Deguet, *et al.*, "An open-source research kit for the da Vinci® surgical system," in *Proc. Int. Conf. Robot. Autom.*, pp. 6434–6439, 2014.

[17] L. C. García-Peraza-Herrera, W. Li, L. Fidon, *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, pp. 5717–5722, 2017.

[18] A. A. Shvets, A. Rakhlin, A. A. Kalinin, *et al.*, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, pp. 624–628, 2018.

[19] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 412–420, 2019.

[20] M. Allan, A. Shvets, T. Kurmann, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[21] F. Qin, S. Lin, Y. Li, *et al.*, "Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6639–6646, 2020.

[22] T. Roß, A. Reinke, P. M. Full, *et al.*, "Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge," *Medical image analysis*, vol. 70, p. 101920, 2021.

[23] S. Lin, F. Qin, H. Peng, *et al.*, "Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6773–6780, 2021.

[24] O. G. Grasa, J. Civera, and J. Montiel, "EKF monocular SLAM with relocalization for laparoscopic sequences," in *Proc. Int. Conf. Robot. Autom.*, pp. 4816–4821, 2011.

[25] O. G. Grasa, E. Bernal, S. Casado, *et al.*, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 135–146, 2013.

[26] A. Marmol, A. Banach, and T. Peynot, "Dense-ArthroSLAM: Dense intra-articular 3-D reconstruction with robust localization prior for arthroscopy," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 918–925, 2019.

[27] J. Song, J. Wang, L. Zhao, *et al.*, "MIS-SLAM: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4068–4075, 2018.

[28] W.-K. Wong, B. Yang, C. Liu, *et al.*, "A quasi-spherical triangle-based approach for efficient 3-D soft-tissue motion tracking," *IEEE Trans. Mechatronics*, vol. 18, no. 5, pp. 1472–1484, 2012.

[29] K. L. Lurie, R. Angst, D. V. Zlatev, *et al.*, "3D reconstruction of cystoscopy videos for comprehensive bladder records," *Biomed. Opt. Exp.*, vol. 8, no. 4, pp. 2106–2123, 2017.

[30] J. Lamarca, S. Parashar, A. Bartoli, *et al.*, "DefSLAM: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Trans. Robot.*, vol. 37, no. 1, pp. 291–303, 2020.

[31] J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, *et al.*, "SD-DefSLAM: Semi-direct monocular SLAM for deformable and intracorporeal scenes," in *Proc. Int. Conf. Robot. Autom.*, pp. 5170–5177, 2021.

[32] D. Recasens, J. Lamarca, J. M. Fácil, *et al.*, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7225–7232, 2021.

[33] M. C. Yip, D. G. Lowe, S. E. Salcudean, *et al.*, "Tissue tracking and registration for image-guided surgery," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 2169–2182, 2012.

[34] H. Wu, J. Zhao, K. Xu, *et al.*, "Semantic SLAM based on deep learning in endocavity environment," *Symmetry*, vol. 14, no. 3, p. 614, 2022.

[35] Y. Wang, Y. Long, S. H. Fan, *et al.*, "Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 431–441, 2022.

[36] S. Sra, S. Nowozin, and S. J. Wright, "The tradeoffs of large scale learning," *Optimization*, pp. 351–368.

[37] H. Pfister, M. Zwicker, J. Van Baar, *et al.*, "Surfels: Surface elements as rendering primitives," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, pp. 335–342, 2000.

[38] M. Keller, D. Lefloch, M. Lambers, *et al.*, "Real-time 3D reconstruction in dynamic scenes using point-based fusion," in *Int. Conf. 3D Vision*, pp. 1–8, 2013.

[39] W. Gao and R. Tedrake, "SurfelWarp: Efficient non-volumetric single view dynamic reconstruction," *arXiv preprint arXiv:1904.13073*, 2019.

[40] F. Richter, J. Lu, R. K. Orosco, *et al.*, "Robotic tool tracking under partially visible kinematic chain: A unified approach," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1653–1670, 2022.

[41] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," in *ACM siggraph 2007 papers*, 2007, pp. 80–es.

[42] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3354–3361, 2012.

[43] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5410–5418, 2018.

[44] C. Godard, O. Mac Aodha, M. Firman, *et al.*, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3828–3838, 2019.

[45] Z. Yang, P. Wang, W. Xu, *et al.*, "Unsupervised learning of geometry from videos with edge-aware depth-normal consistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[46] Y. Li, A. Bozic, T. Zhang, *et al.*, "Learning to optimize non-rigid tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4910–4918, 2020.

[47] K.-L. Low, "Linear least-squares optimization for point-to-plane ICP surface registration," *Chapel Hill, University of North Carolina*, vol. 4, no. 10, pp. 1–3, 2004.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances Neural Inf. Proc. Syst.*, vol. 28, 2015.
[49] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.
[50] C. Lassner and M. Zollhofer, "Pulsar: Efficient sphere-based neural rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1440–1449, 2021.
[51] Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image process.*, vol. 13, no. 4, pp. 600–612, 2004.
[52] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 13 116–13 125, 2020.
[53] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Symp. Geometry Process.*, vol. 4, pp. 109–116, 2007.
[54] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 343–352, 2015.
[55] Y. Han, F. Liu, and M. C. Yip, "A 2D surgical simulation framework for tool-tissue interaction," *arXiv preprint arXiv:2010.13936*, 2020.
[56] F. Richter, E. K. Funk, W. S. Park, *et al.*, "From bench to bedside: The first live robotic surgery on the dVRK to enable remote telesurgery with motion scaling," in *Int. Symp. Med. Robot.*, pp. 1–7, 2021.
[57] M. Ye, E. Johns, A. Handa, *et al.*, "Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery," *arXiv preprint arXiv:1705.08260*, 2017.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[59] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 270–279, 2017.
[60] L. C. Chen, Y. Zhu, G. Papandreou, *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Europ. Conf. Comput. Vis.*, pp. 801–818, 2018.
[61] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 234–241, 2015.
[62] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, *et al.*, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.
[63] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, 2015.
[64] S. Zheng, S. Jayasumana, B. Romera-Paredes, *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1529–1537, 2015.
[65] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, *et al.*, "Efficient uncertainty estimation for semantic segmentation in videos," in *Proc. Europ. Conf. Comput. Vis.*, pp. 520–535, 2018.
[66] C. J. Holder and M. Shafique, "Efficient uncertainty estimation in semantic segmentation via distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3087–3094, 2021.
[67] M. Poggi, F. Aleotti, F. Tosi, *et al.*, "On the uncertainty of self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3227–3237, 2020.
[68] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.