Semantic-SuPer: Employing Semantic Perception for Endoscopic Tissue Identification, Reconstruction, and Tracking

Shan Lin[1], Jingpei Lu[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.**

I. Introduction

As surgical robots advance, there is a growing interest in equipping them with intelligence to better understand the surgical environment. 3D scene understanding of both geometric and semantic features of anatomy enables more effective navigation in patients and paves the way for automated surgical tasks. Current surgical navigation relies on segmented, fixed preoperative images like CT/MRI scans to serve as a map to provide 3D geometry and semantics information [1], [2], [3], [4], [5], but it becomes less reliable in cases of significant deformation, limiting its use in deformable surgical scenes.

In contrast, video semantic segmentation can extract precise anatomical information in real-time to update the navigation map. While integrating semantics and 3D geometry has shown benefits in indoor and autonomous driving scenarios [6], [7], [8], [9], surgical scenes pose unique challenges with deformable, textureless tissues. These features complicate data association, despite attempts to address it with advanced approaches like photometric loss [10]. Our approach takes a different angle by using semantics to guide data association, introducing a "morphing loss" to ensure border consistency between semantic segmentation and the 3D model.

In summary, we introduce a novel surgical perception framework Semantic-SuPer, which merges semantic insights extracted from videos with 3D deformation tracking and reconstruction. This framework enhances the established SuPer framework [11], [12] by integrating model-free deformable tissue tracking with semantic information, increasing robustness and improving accuracy for surgical perception.

[1]S. Lin, J. Lu, 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: {shl102, jil360, 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.

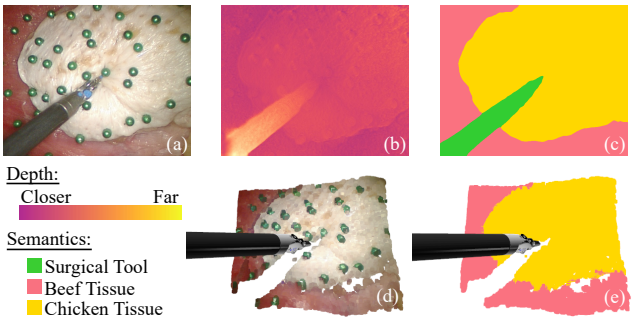


Fig. 1. **A demonstration of Semantic-SuPer.** (a) Input video frame. (b) Depth and (c) semantic segmentation map estimated from the input. (d) Scene rendered from the tracked surfels and surgical tool pose. (e) Scene rendered from surfels visualized by colors corresponding to their semantics.

II. Related Work

Surgical Scene Semantic Segmentation aims to segment surgical images into tissue and tool regions. Deep models have shown advanced performance for endoscopic image segmentation [13], [14], [15], [16]. While many works focus on developing segmentation algorithms, we emphasize the integration of semantics into 3D surgical scene tracking.

Endoscopic Tissue Tracking is a specific area of non-rigid tracking, where the low textured, deformable tissue is a significant challenge. Existing methods often rely on as-rigid-as-possible [17], [18], [19], [11] or spline-based [20], [21], [22], [10] assumptions for tracking in complex surgical scenes, which remains a challenge for even the latest techniques [10], [23]. To address this, we introduce a comprehensive framework that combines surgical scene tracking and reconstruction with semantic segmentation, demonstrating how semantic information can aid data association.

Semantic SLAM leverages semantic information to enhance SLAM. It is widely used in autonomous driving for various purposes, like feature selection, data association improvement, dynamic region identification, and long-term localization [24], [6], [7], [9]. Yet, such methods are 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 [25], [26]. 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 [11], [12] by including pixel-wise semantic labels as inputs and building *surface elements (surfels)* [27], [28] with semantic information, as shown in Figure 2. This semantic information is

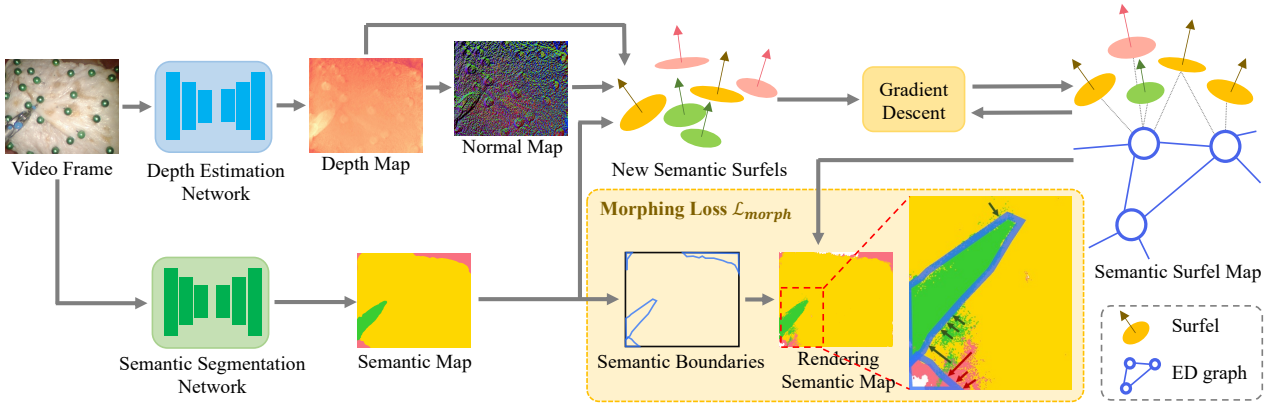


Fig. 2. **Overview of the proposed framework.** The depth and semantics are extracted from video. The transformations of ED nodes are optimized with PyTorch's automatic differentiation to match the observations. Surfel position and normal updates are then controlled by the ED nodes.

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 geometry of the entire surgical scene, including both the tools controlled by the surgical robot and the deforming tissues. Semantic-SuPer is primarily built upon SuPer's model-free tissue tracking method and employs surface elements (surfels) [27] to represent the surgical scene. 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 [29], [28] for more details about SuPer.

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 [30], 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 as the average motions of their adjacent ED nodes

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

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

where $\mathbf{T}_g \in SE(3)$ is the global homogeneous transformation shared by all surfels, $T(\cdot) \in SE(3)$ is the local homogeneous transform 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 $\mathcal{N}_i$.

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

Since the quality of the depth map significantly impacts tracking performance, previously we conducted a comprehensive study showing that pre-trained deep learning depth estimation models led to better tissue tracking than traditional stereo matching algorithms [12]. Yet, without finetuning the deep models on the surgical data, their predictions are still noisy and can result in early tracking failures. Thus, we use Monodepth2 and tune it in a self-supervised fashion with stereo training data [31]. 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 [32]. 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 [33], 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, we use a popular segmentation model DeepLabv3+ [34] 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 $\mathbb{y}_i \in \mathbb{R}$ is the class that corresponds to the highest softmax score.

C. Semantic-aware Registration

At each new frame, transformations of ED nodes are estimated according to new observations by solving

$$\arg\min_{\mathbf{q}, \mathbf{b}, \mathbf{T}_g} \lambda_s \mathcal{L}_{sim} + \lambda_m \mathcal{L}_{morph} + \lambda_r \mathcal{L}_{reg} \quad (3)$$

where $\mathcal{L}_{sim}$ 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. (3) is solved using gradient descent [35] with PyTorch's automatic differentiation.

1) Similarity Term $\mathcal{L}_{sim}$: The first loss is the same point-to-plane ICP loss [36] as in SuPer [11]:

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

where $\overline{\mathbf{p}}_o$ and $\vec{\mathbf{n}}_o$ are the observed positions and normals, bilinearly sampled [37] from the observations at the projected pixel coordinates of $\widetilde{\overline{\mathbf{p}}}_i$, and $\rho_{i,o}$ is the weight for each term. The original SuPer, $\rho_{i,o} = 1$ [11]. For Semantic-SuPer, the weight is computed based on the Jensen–Shannon divergence [38] 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 [39], 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 [40], N is image pixel number, I_i is 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 $\mathcal{L}_{morph}$: $\mathcal{L}_{sim}$ is known to suffer from gradient locality, thereby $\mathcal{L}_{morph}$ is proposed to provide longer-range hints. It minimizes the distance of surfels whose projections onto the image fall outside of their own semantic region (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 3D points to the image plane, $\mathcal{R}_i$ is the 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 $\mathcal{R}_i$.

3) Regularization term $\mathcal{L}_{reg}$: The regularization term consists of two terms. The first term, $\mathcal{L}_{face}$, ensures that all ED nodes move as rigidly as possible and is calculated based on the changes in area of all the transformed triangle surfaces in the ED graph [41]:

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

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) [42], [43] for evaluation (see Figure 3). We overlaid a piece of chicken meat across a piece of beef. The green pins attached on the tissue were used to collect ground truth for evaluation (Section IV-B). The beef was pushed up-and-down manually from below, and the dVRK was used to control a surgical robotic arm to grasp and tug the chicken tissue. 4 trials were conducted, each consisting of 150 rectified 640×480 frames at 30 fps, named Lab 1, 2, 3, and 4 in the following sections.

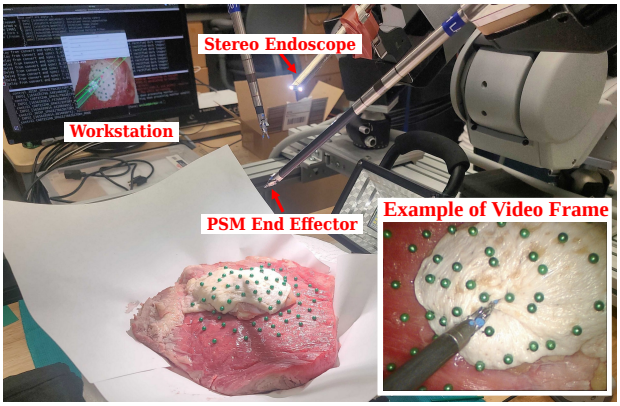


Fig. 3. Experimental setup with the dVRK system.

B. Metrics and Ground Truth

For quantitatively evaluation, we attached roughly 60 green pins onto the tissue surface (see Figure 3), and extracted their trajectories in the image plane throughout each trial by finding green region in the Hue-Saturation-Value (HSV) color space. 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 [44], and then finetuned with our data. Since the generalizability of existing 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 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 [45], with a batch size of 16, a learning rate of 10^{-4} for the first 15 epochs and 10^{-5} for the remainder.

2) Semantic Segmentation: DeepLabv3+ [34] is 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.

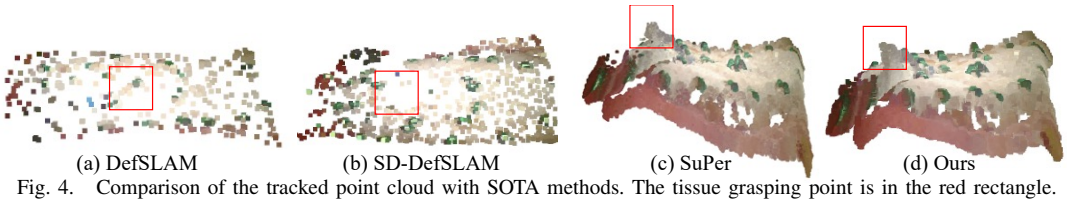
3) Deformable Tracking: The surfels are initialized and added in the same manner as SuPer [11]. To initialize the ED graph, since the depths vary a lot between trials, we choose the step size for each trial by ensuring the average edge length of the graph is around 5mm. The hyperparameters for the cost functions are $\lambda_m = 10$, $\lambda_s = 1$, and $\lambda_r = 10$.

TABLE I

REPROJECTION ERROR COMPARISON ON OUR DATASET.

Method	Data			
	Lab1	Lab2	Lab3	Lab4
DefSLAM [22]	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 [10]	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 [11]	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**.



(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.

TABLE II

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)

Refer to Table I for notes on the reprojection errors.

D. Results

As shown in Table I, we compare Semantic-SuPer against two baselines: 1) SuPer [11], 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 [22] and SD-DefSLAM [10]. 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.

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 presents minimum deformation, causing a minor performance difference between Semantic-SuPer and SuPer on it. Moreover, our framework outperforms DefSLAM, 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.

Moreover, a comparison of the influence of different segmentation algorithms on Semantic-SuPer is shown in Table II. We measure the quality of the predicted segmentation maps using Hausdorff distance (HD), which indicates the largest segmentation error, and F_1 score [46]. Table II 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) [47].

V. DISCUSSION

Table I demonstrates the benefits of including semantics for tissue tracking with morphing loss. Further, 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 [48], [49], [50], as well as leveraging multi-task learning-based cross-task knowledge [51] to estimate uncertainty could lead to better tracking performance.

VI. CONCLUSIONS

We present a novel surgical perception framework Semantic-SuPer that achieves better surgical scene 3D reconstruction and tracking by integrating semantic information, 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 input data to achieve better uncertainty estimation for more robust tracking.

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] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances Neural Inf. Proc. Syst.*, vol. 28, 2015.

[38] 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.

[39] C. Lassner and M. Zollhofer, "Pulsar: Efficient sphere-based neural rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1440–1449, 2021.

[40] 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.

[41] Y. Han, F. Liu, and M. C. Yip, "A 2D surgical simulation framework for tool-tissue interaction," *arXiv preprint arXiv:2010.13936*, 2020.

[42] 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.

[43] 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.

[44] 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.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.
[51] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.