

Robust Surgical Tool Tracking with Pixel-based Probabilities for Projected Geometric Primitives

Christopher D'Ambrosia¹, Florian Richter², Zih-Yun Chiu², Nikhil Shinde²,
Fei Liu², Henrik I. Christensen³ and Michael C. Yip²

Abstract—Controlling robotic manipulators via visual feedback requires a known coordinate frame transformation between the robot and the camera. Uncertainties in mechanical systems as well as camera calibration create errors in this coordinate frame transformation. These errors result in poor localization of robotic manipulators and create a significant challenge for applications that rely on precise interactions between manipulators and the environment. In this work, we estimate the camera-to-base transform and joint angle measurement errors for surgical robotic tools using an image based insertion-shaft detection algorithm and probabilistic models. We apply our proposed approach in both a structured environment as well as an unstructured environment and measure to demonstrate the efficacy of our methods.

I. INTRODUCTION

Surgical automation has the potential to increase access to and quality of surgical care [1]. For this reason, automation of surgical subtasks remains an active area of research [2]. Prior work includes autonomous surgical cutting [3], [4], hemostasis [5], peg transfer [6], and suturing [7], [8]. Surgical automation, however, requires robust and accurate end effector localization [9]. Localization typically involves camera or image-based feedback [10]. However, uncertainty in the camera-to-robot transforms [11] as well as imprecision due to mechanical factors including cable-stretch and hysteresis [12] limits image-based localization accuracy.

A. Related Works

Prior work in estimating the base-to-camera transform includes solving homogeneous linear systems based on multiple images that capture markers affixed to robotic links [14], [15], Solve-PnP approaches using detected markers on robotic manipulators [16] or marker-less keypoint detection [17], [18], [19], [20]. In the space of surgical robotics, previous works have calibrated remote center of motion (RCM) robotic systems [21] and kinematic remote center coordinate systems (KCS) that accounts for error in the RCM camera frame-to-base transform [22]. Markers, learned

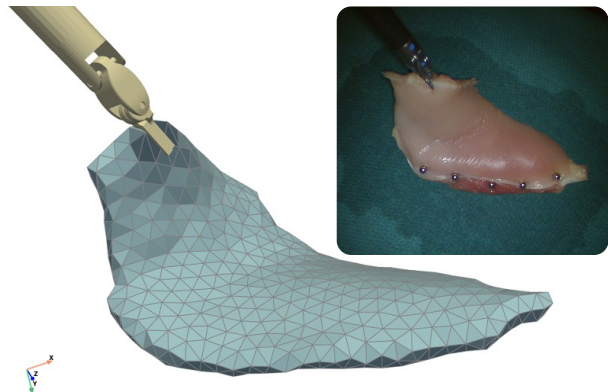


Fig. 1: The insertion-shaft on surgical robotic tools is an excellent feature for localization. In this work, we present a novel approach for detecting the insertion-shaft and incorporating it into Bayesian filtering to probabilistically localize and track surgical robotic tools. We stress test our method in a challenging deformable tissue dataset where a surgical robotic tool is in the back of the scene with low light and commanded to deform tissue. The tissue is registered to a simulated scene using a separate approach in the camera frame [13] and notice how well aligned the grasp between the localized surgical robotic tool and the tissue is in the 3D rendering.

features, silhouette matching, and online template matching from images have been used for KCS tracking [23], [24], [25], [26].

To estimate joint angle measurement errors in the context of cable-driven surgical robotics, prior work has applied data-driven approaches, real-time inverse kinematics, and physical modeling of cable transmission friction and hysteresis [27], [28]. Learning-based approaches for estimating cable stretch and Gaussian processes for measurement error compensation have also been used [12], [29], [30]. Image-based methods have used unscented Kalman filters and deep learning tools to accurately estimate cable-driven joint angles [31], [32].

Estimating both camera-to-base transform and joint angle offsets simultaneously using joint calibration techniques have been proposed [33], [34]. These methods, however, can not account for dynamic uncertainties such as non-constant joint angle errors. Methods to achieve real-time estimation of dynamic joint angle errors have used iterative closest point matching based on 3D point clouds and Kalman filters [35]. Probabilistic approaches using observation models parameterized by physical parameters have also been explored [36]. While these works prioritize integrating additional sensors such as depth cameras into real-time estimation, we rely solely on endoscopic images which is available in surgical

¹Christopher D'Ambrosia is with the College of Physicians and Surgeons, Columbia University and the Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093, U.S.A. (cdambros@ucsd.edu)

²Florian Richter, Zih-Yun Chiu, Nikhil Shinde, Fei Liu and Michael C. Yip are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, California 92093, U.S.A. (frichter@ucsd.edu; zchiu@ucsd.edu; nshinde@ucsd.edu; f4liu@ucsd.edu; yip@ucsd.edu)

³Henrik I. Christensen is with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093, U.S.A. (hichristensen@ucsd.edu)

robotic scenes. Furthermore, surgical robotics has the unique challenge of a partially visible kinematic chain which creates a one-to-many mapping between the detected images and the parameter set describing the camera-to-base transform and joint angle errors [11].

B. Contribution

To address the challenge of estimating the camera-to-base transform and joint angle errors, our prior work introduced the concept of lumped error. Lumped error accounts for uncertainty in the camera-to-base transform, camera calibration, and joint transforms in a forward-kinematic model for partially visible kinematic chains [11]. Estimating this lumped error has enabled more accurate, probabilistic tool tracking for surgical systems [37] and utilized for autonomous blood suction [5] and suture needle manipulation [7].

One key observation we made is the importance of the insertion-shaft as an observation for estimating the lumped error. The insertion-shaft is how the surgical robotic manipulator is inserted into the surgical scene through port entries and a standard feature for any laparoscopic tool hence using it for estimation will help constrain the RCM location with respect to the camera frame [21], [38]. Furthermore, the insertion-shaft can be modelled as geometric primitive, a cylinder, which has an analytical camera projection equation [39]. The insertion-shaft was incorporated into our previous work in a probabilistic fashion to estimate the lumped error via a detection algorithm and Bayesian filtering. While using the insertion-shaft as an observation was crucial for surgical robotic tool tracking, there are two limitations with our previous implementation: 1) detection of the insertion-shaft parameters from the endoscopic data was inconsistent in unstructured scenes and 2) the probabilistic observation model to update the lumped error belief was limited to a line parameterization space. To address these challenges we contribute a laparoscopic tool insertion-shaft detection approach based on a Deep Neural Network (DNN) with multiple observation models for Bayesian filtering. The approach is integrated into a Particle Filter and a comparative study is done to evaluate the different observation models on two datasets collected on the da Vinci Research Kit (dVRK) [40].

II. BACKGROUND

Any joint-link on a surgical robot can be localized in the camera frame by combining the robots forward kinematics and the base-to-camera transform. Written explicitly, the homogeneous transform matrix describing the j -th joint-link in the camera frame is

$${}^c\mathbf{T}_j = {}^c\mathbf{T}_b \prod_{i=1}^j {}^{i-1}\mathbf{T}_i(q_i) \quad (1)$$

where ${}^c\mathbf{T}_b \in SE(3)$ is the base-to-camera transform and ${}^{i-1}\mathbf{T}_i(q_i) \in SE(3)$ is the i -th joint transform with joint angle q_i . In an ideal case, the joint transforms are provided by the robot manufacturer and the joint angles are measured by encoders hence only the base-to-camera transform needs to be solved for.

Due to the unique workspace challenges faced in surgical environments (e.g. narrow corridors), surgical robotic systems often rely on cable-drive designs without encoders at the joint-link location such as the dVRK [40] and RAVEN robotic systems [41], [42]. Instead joint angles are read at the motor which leads to inaccurate joint readings from mechanical phenomena such as cable stretch, hysteresis and cable stretch [31], [12]. Furthermore, the kinematic chains of surgical robots are only partially visible in the camera frame. For example, the endo-wrist and insertion-shaft from da Vinci @surgical platform are often the only joint-links visible in the endoscopic frame.

Localizing the surgical robot in the camera frame from endoscopic images by estimating the hand-eye transform and joint angle errors is not feasible since the parameters are not identifiable from the camera frame due to the partially visible robotic chain [11]. In our previous work, we derive a smaller parameter set that is identifiable from the camera frame and enables complete localization of the surgical robotic links visible in the camera frame. The forward kinematics in (1) can be written with the new parameter set as

$${}^c\mathbf{T}_j = \mathbf{E} \prod_{i=1}^j {}^{i-1}\mathbf{T}_i(\tilde{q}_i) \quad (2)$$

where $\mathbf{E} \in SE(3)$ is the lumped error transform and \tilde{q}_i are the noisy joint angle readings. The lumped error transform captures both the camera-to-base transform and errors in joint angle readings. Please see [11] for a complete derivation.

To estimate the lumped error, features are detected on the surgical robotic tool. The scope of this work focuses on using the insertion-shaft as a feature which can be modelled as a cylinder with the following parameters: a position on the center-line, $\mathbf{p}_0^j \in \mathbb{R}^3$, direction vector of the centerline, $\mathbf{d}^j \in \mathbb{R}^3$, and radius $r \in \mathbb{R}^+$. Through the kinematics equation in (2), the insert-shaft cylinder on the j -th joint can be transformed to the camera frame: $\mathbf{p}_0^c = {}^c\mathbf{T}_j\mathbf{p}_0^j$ and $\mathbf{d}^c = {}^c\mathbf{T}_j\mathbf{d}^j$. Finally, a point, $\mathbf{p} \in \mathbb{R}^3$, is defined to be on the insertion-shaft in the camera frame if it satisfies the following equations

$$\begin{cases} (\mathbf{p} - \mathbf{p}_a^c)^\top (\mathbf{p} - \mathbf{p}_a^c) - r^2 = 0 \\ (\mathbf{d}^c)^\top (\mathbf{p} - \mathbf{p}_a^c) = 0 \end{cases} \quad (3)$$

where \mathbf{p}_a^c is a point on the center-line

$$\mathbf{p}_a^c = \mathbf{p}_0^c + \lambda \mathbf{d}^c \quad (4)$$

and $\lambda \in \mathbb{R}$. Chaumette showed that projecting (3) with the camera pin-hole model results in two line segments that represent the two edges of the projected cylinder [39],

$$a_1X + b_1Y + c_1 = 0 \quad a_2X + b_2Y + c_2 = 0 \quad (5)$$

where (X, Y) are pixel coordinates on a unit camera (i.e. $(X, Y) = (\frac{u-c_u}{f_x}, \frac{v-c_v}{f_y})$ where $(c_u, c_v), (f_x, f_y)$ are the

principle point and focal length of a non-unit camera),

$$\begin{aligned} A_{1,2} &= \frac{r(x_0^c - a^c(\mathbf{p}_0^c)^\top \mathbf{d}^c)}{\sqrt{(\mathbf{p}_0^c)^\top \mathbf{p}_0^c - (\mathbf{p}_0^c)^\top \mathbf{d}^c - r^2}} \pm (c^c y_0^c - b^c z_0^c) \\ B_{1,2} &= \frac{r(y_0^c - b^c(\mathbf{p}_0^c)^\top \mathbf{d}^c)}{\sqrt{(\mathbf{p}_0^c)^\top \mathbf{p}_0^c - (\mathbf{p}_0^c)^\top \mathbf{d}^c - r^2}} \pm (a^c z_0^c - c^c x_0^c) \\ C_{1,2} &= \frac{r(z_0^c - c^c(\mathbf{p}_0^c)^\top \mathbf{d}^c)}{\sqrt{(\mathbf{p}_0^c)^\top \mathbf{p}_0^c - (\mathbf{p}_0^c)^\top \mathbf{d}^c - r^2}} \pm (b^c x_0^c - a^c y_0^c) \end{aligned} \quad (6)$$

and $\mathbf{p}_0^c = [x_0^c \ y_0^c \ z_0^c]^\top$, $\mathbf{d}^c = [a^c \ b^c \ c^c]^\top$. By combining the forward kinematic model in (2) with a the insertion-shaft projection (5), estimation methods can be used to find the lumped error, \mathbf{E} , from insertion-shaft feature detections hence localizing the surgical robotic tools in the camera frame.

III. METHODS

To estimate the lumped error, a particle filter tracks its probability distribution temporally using observations from the endoscopic camera. The lumped error is parameterized with a translation and axis-angle vector, $\mathbf{E}(\mathbf{b}_t, \mathbf{w}_t)$ where $\mathbf{b}_t, \mathbf{w}_t \in \mathbb{R}^3$. In the coming sub-sections, the motion model, insertion-shaft line detection algorithm, and observation model are described to fully define the necessary components for a Bayesian Filter to track the lumped error hence localizing the surgical robotic tool in the camera frame.

A. Motion Model

The lumped error's motion model is defined with additive Gaussian noise due to its ability to generalize over a large number of random processes. Written explicitly, the motion model is

$$[\mathbf{b}_{t+1}, \mathbf{w}_{t+1}]^\top = \mathcal{N}([\mathbf{b}_t, \mathbf{w}_t]^\top, \Sigma_{\mathbf{b}, \mathbf{w}, t+1}) \quad (7)$$

where $\Sigma_{\mathbf{b}, \mathbf{w}, t+1} \in \mathbb{R}^{6 \times 6}$ is a covariance matrix.

B. Insertion-Shaft Line Detection

In our previous work, we manually tuned a canny edge detection algorithm [45] followed by a hough transform to detect lines [46]. From there, a greedy association strategy is employed to find which detected lines are associated with the edges of the insertion-shaft. In this work, we use SOLD2, a self-supervised occlusion-aware line description and detection DNN [43], for more reliable, robust insertion-shaft line detection in unstructured environments. The SOLD2 model at inference is provided a reference image of the surgical robotic tool and outputs the detected insertion-shaft lines by finding which of the detected lines are associated with the insertion-shaft from the reference image. SOLD2 also provides a heatmap of all the detected pixels associated with the detected lines. Written explicitly, SOLD2 outputs

$$\mathbf{e}_{1,2}^a, \mathbf{e}_{1,2}^b, \mathbf{H} = f(\mathcal{I}; \mathcal{I}_{ref}) \quad (8)$$

where $\mathbf{e}_{1,2}^a, \mathbf{e}_{1,2}^b \in \mathbb{R}^2$ are the endpoints of the detected insertion-shaft line segments, $\mathbf{H} \in \mathbb{I}^{H \times W}$ is the heatmap,

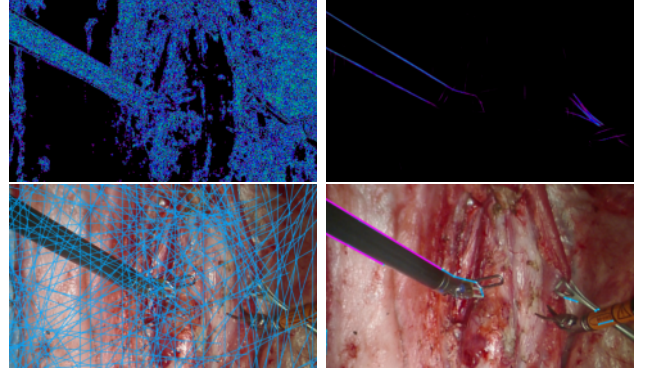


Fig. 2: From left to right, the columns show our previous approach, based on Canny Edge detection [11], and proposed approach, based on SOLD2 [43], for detecting the insertion shaft when deployed in a live surgery [44]. The top row shows a heatmap of the pixels potentially associated with a line segment and the bottom row shows all the detected line segments in light blue from both approaches. Additionally, the purple line segments for the SOLD2 approach highlight the detected lines associated with the insertion shaft of the surgical tool which was not possible with the previous approach.

$\mathcal{I} \in \mathbb{I}^{H \times W \times 3}$ is the input image, and $\mathcal{I}_{ref} \in \mathbb{I}^{H \times W \times 3}$ is the reference image. An example insertion-shaft detection is shown in Fig. 2.

SOLD2 relies on similarity-based image features to evaluate candidate line endpoints. Surgical endoscope images, however, may have barrel distortion and lens vignetting [47] which significantly increases image feature variance between the center and periphery of the image. Lines with one endpoint located at the center of the image and one endpoint located at the periphery of the image (e.g. insertion-shaft edges) are often falsely eliminated due to the distortion and artifact-based dissimilarities in the neighborhoods of proposed endpoints. We augment our surgical endoscope images through cropping and downsampling to reduce the number of false negative endpoint detections. While any line detection and association algorithm could be used in our tool-tracking pipeline, we choose to use SOLD2 rather than other learning-based line detection algorithms [48] because SOLD2 provides line associations across video frames which are a necessary component for our downstream filtering of false line detections.

C. Observation Model

From the insertion-shaft line detection, we present four different observation models and compare them experimentally. The first two are similar to our previous work where the uncertainty is modelled on line parameters, distance to the origin and slope angle. Meanwhile the second two observation models work with pixels directly by deriving a random variable that describes how well the pixel fits on to the projected insertion-shaft.

1) *Endpoint Intensities to Polar*: SOLD2 architecture directly outputs endpoint candidates and then completes the line segments though the heatmap [43]. Similarly, the first observation model uses the detected endpoints, $\mathbf{e}_{1,2}^a, \mathbf{e}_{1,2}^b$, to derive detected line parameters, $\rho_{1,2}^e, \theta_{1,2}^e$, which represents

the distance to origin and slope angle. To increase the robustness, we derive the detected line parameters from a set of points near the end-points that have a high heat-map value hence should be part of a line. Written explicitly, the endpoint set is

$$\mathcal{E}_e(\mathbf{e}) = \{\mathbf{p} \mid \|\mathbf{p} - \mathbf{e}\| \leq \alpha_e \cap \mathbf{H}(\mathbf{p}) \geq \beta\} \quad (9)$$

where α_e is the end-point search radius and β is the heatmap threshold. From the end-point sets, $\mathcal{E}_e(\mathbf{e}_{1,2}^a)$ and $\mathcal{E}_e(\mathbf{e}_{1,2}^b)$, sequential RANSAC [49] is employed to robustly detect two line segments parameterized by $\rho_{1,2}^e, \theta_{1,2}^e$. The probabilistic observation model for time t is defined in the line parameter space and written as a summation of Gaussians, similar to our previous work,

$$P(\rho_{1,2,t}^e, \theta_{1,2,t}^e | \mathbf{b}_t, \mathbf{w}_t) \propto \sum_{i=1,2} e^{-\gamma_\rho |\rho_{i,t}^e - \rho_i(\mathbf{b}_t, \mathbf{w}_t)| - \gamma_\theta |\theta_{i,t}^e - \theta_i(\mathbf{b}_t, \mathbf{w}_t)|} \quad (10)$$

where γ_ρ and γ_θ describe the standard deviation of the $\rho_{1,2}^e$ and $\theta_{1,2}^e$ detections and $\rho_{1,2}(\mathbf{b}_t, \mathbf{w}_t), \theta_{1,2}(\mathbf{b}_t, \mathbf{w}_t)$ are the projected insertion-shaft lines from the estimated lumped error. The projected insertion-shaft lines from the estimated lumped error come from (2) and (5) to generate line parameters $a_{1,2}(\mathbf{b}_t, \mathbf{w}_t), b_{1,2}(\mathbf{b}_t, \mathbf{w}_t), c_{1,2}(\mathbf{b}_t, \mathbf{w}_t)$ and subsequently converted as follows

$$\begin{aligned} \theta_{1,2}(\mathbf{b}_t, \mathbf{w}_t) &= -\tan^{-1} \left(\frac{a_{1,2}(\mathbf{b}_t, \mathbf{w}_t)}{b_{1,2}(\mathbf{b}_t, \mathbf{w}_t)} \right) \\ \rho_{1,2}(\mathbf{b}_t, \mathbf{w}_t) &= -\frac{c_{1,2}(\mathbf{b}_t, \mathbf{w}_t)}{b_{1,2}(\mathbf{b}_t, \mathbf{w}_t)} \sin(\theta_{1,2}(\mathbf{b}_t, \mathbf{w}_t)) \end{aligned} \quad (11)$$

2) *Line Intensities to Polar*: The Endpoint Intensities to Polar observation model is extended in this observation model by expanding the set of points to include points along the line segment rather than just near the detected endpoints to improve robustness by including more candidate points associated with the detected insertion-shaft line. Written explicitly, the new set of points is defined as

$$\mathcal{E}_l(\mathbf{e}^a, \mathbf{e}^b) = \{\mathbf{p} \mid \min_{l \in L(\mathbf{e}^a, \mathbf{e}^b)} \|\mathbf{p} - l\| \leq \alpha_l \cap \mathbf{H}(\mathbf{p}) \geq \beta\} \quad (12)$$

where $L(\mathbf{e}^a, \mathbf{e}^b)$ is the set of points on the line segment between $\mathbf{e}^a, \mathbf{e}^b$ and α_l is the search radius about the line. Similar to the previous observation model, line segments parameters $\rho_{1,2}^l, \theta_{1,2}^l$ are derived from the line point sets, $\mathcal{E}_l(\mathbf{e}_{1,2}^a, \mathbf{e}_{1,2}^b)$, using sequential RANSAC. The same probability model in (11) is used to define the observation model for the detected insertion-shaft lines, $\rho_{1,2,t}^l, \theta_{1,2,t}^l$, at timestep t .

3) *Endpoint Intensities*: The previous observation models use a noise distribution on derived line parameters. Instead, the pixel based observation works directly in the pixel space by deriving a random variable $R(\cdot)$ which describes how well the detected pixel fits to the projected insertion-shaft,

$$R(x, y | \theta, \rho) = \cos(\theta)x + \sin(\theta)y - \rho \quad (13)$$

where x, y is an input pixel coordinate. The derived variable $R(\cdot)$ is inspired from [37] where a point is fitted to an ellipse projected from a suture needle. The fitting-based

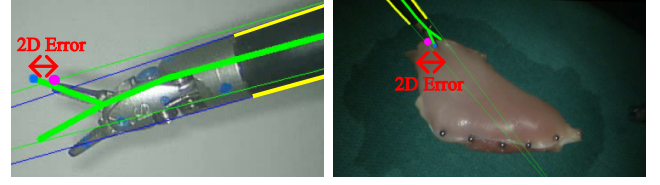


Fig. 3: Figures from our structured and deformable tissue datasets, from left to right respectively, where a green skeleton is overlaid to show the tracked surgical robotic tool in the scene and the yellow lines correspond to insertion-shaft line detections from our proposed approach. As shown in red, the 2D error metric for both datasets is calculated by comparing the projected and manually labelled inferior jaw tool tip. Note that the surgical tool in the deformable tissue dataset is intentionally far away to simulate a challenging surgical tool tracking scenario.

derived random variables allow for a pixel-based probability distribution without requiring pixel-to-pixel association rather a pixel-to-line association. We assume the noise of the pixels associated with the insertion-shaft's edges, \mathbf{p} , are independent and normally distributed from the projected line, $\mathbf{p} \sim \mathcal{N}([X, Y]^\top, \sigma^2)$ where X, Y come from the projected insertion-shaft, i.e (5). Therefore, the derived random variable is Gaussian whose distribution is

$$R(\mathbf{p} | \theta, \rho) \sim \mathcal{N}(0, \cos(\theta)^2 \sigma^2 + \sin(\theta)^2 \sigma^2) \quad (14)$$

hence providing a probability distribution for a detected insertion-shaft pixel without requiring explicit pixel-to-pixel association. This observation model applies the derived random variable over the endpoint sets of the detected insertion-shaft, $R(\mathbf{p}_t | \theta_{1,2}(\mathbf{b}_t, \mathbf{w}_t), \rho_{1,2}(\mathbf{b}_t, \mathbf{w}_t)) \forall \mathbf{p}_t \in \mathcal{E}_e(\mathbf{e}_{1,2,t}^a), \mathcal{E}_e(\mathbf{e}_{1,2,t}^b)$, for a timestep t .

4) *Line Intensities*: This observation model also uses the derived random variable $R(\cdot)$ defined in (13) with the same uncertainty distribution from (14). The only difference with the previous observation model is the point sets which the derived random variable is applied to now covers the entire line, $R(\mathbf{p}_t | \theta_{1,2}(\mathbf{b}_t, \mathbf{w}_t), \rho_{1,2}(\mathbf{b}_t, \mathbf{w}_t)) \forall \mathbf{p}_t \in \mathcal{E}_l(\mathbf{e}_{1,2,t}^a, \mathbf{e}_{1,2,t}^b)$, for a timestep t .

IV. EXPERIMENTS AND RESULTS

A comparison experiment is ran to test the efficacy of the proposed methods by implementing the different observation models into a Particle Filter to track a surgical robotic tool. The tracking algorithm compares across all the presented observation models in Sections III-C.1 to III-C.4 and the following baselines:

- *Canny*: previous surgical tool tracking approach [11]
- *Endpoint to Polar*: derived from Section III-C.1 where $\alpha = 0$ hence only relies on the endpoint detection algorithm from the DNN

In our implemented particle filter, we also include point feature detections replicated from our previously established approaches [23], [11].

A. Implementation Details

Motion model parameters are retained from our previous work [11]. The line parameter observation models used

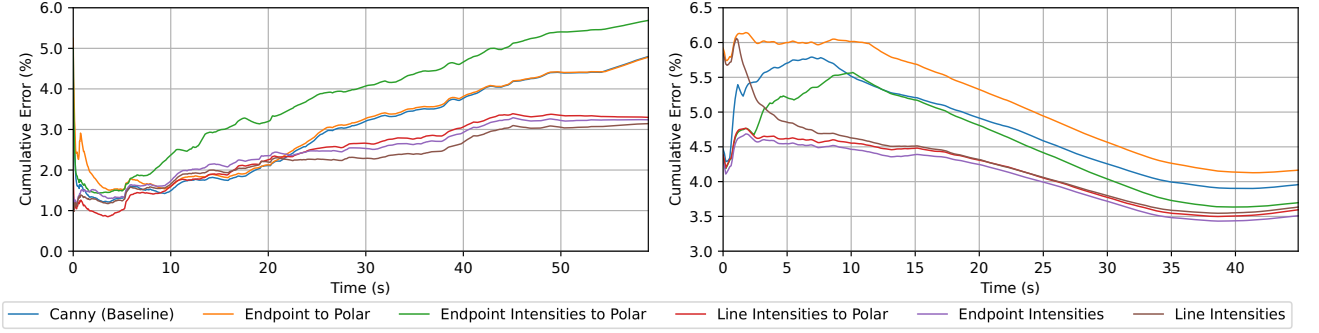


Fig. 4: From left to right, the plots show the accumulated 2D error for Structured and Deformable Tissue Datasets, respectively. Our proposed line intensities to polar, endpoint intensities, and line intensity observation models are measured to provide consistent tracking results as seen in the plots.

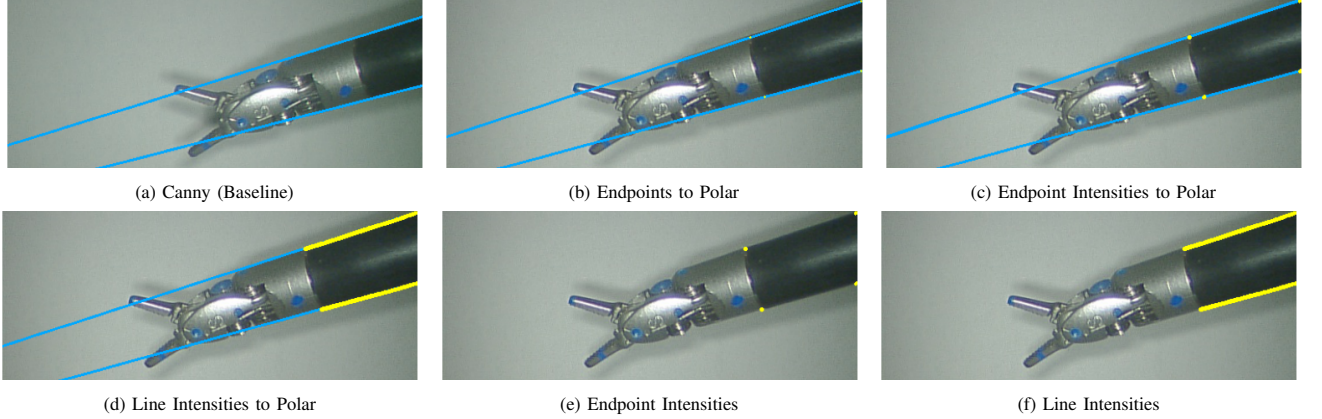


Fig. 5: Illustration of the different observation models for detecting the insertion shaft in the deformable tissue datasets. The blue lines highlight the parameterized lines (i.e. polar parameterization) and the yellow points represent the pixel set used in the observation model. In such a structured scene, all insertion-shaft detection algorithms work well. This figure is best viewed in color.

TABLE I: Surgical tool tracking results from structured environment and deformable tissue datasets.

Method	Structured 2D Error (%)	Deformable Tissue 2D Error (%)
Canny Edge Detector [11]	4.8 (3.0)	4.0 (1.2)
Endpoint to Polar	4.8 (2.9)	4.2 (1.3)
Endpoint Intensities to Polar	5.7 (2.9)	3.7 (1.4)
Line Intensities to Polar	3.3 (2.0)	3.6 (0.9)
Endpoint Intensities	3.2 (2.0)	3.5 (0.9)
Line Intensities	3.1 (2.1)	3.6 (0.9)

sequential RANSAC parameters of 5 iterations, 3.0 samples, a residual threshold of 0.75, and maximum number of trials of 100. The pixel observation models used a radius of $\alpha = 10$ pixels and a heatmap threshold of $\beta = 0.90$. The previous canny edge detection schema with hough accumulator used hyperparameters $\rho_{acc} = 5.0$, $\theta_{acc} = 0.09$, a vote threshold of $v = 100$, and cluster distances of $\rho_{cd} = 5.0$, $\theta_{cd} = 0.09$ as per [11].

B. Datasets

1) *Structured Environment Dataset*: To create our Structured Environment Dataset, we used a stationary dVRK [40] stereo endoscope to record a single PSM arm tool at various, teleoperated jaw and wrist configurations in 1080×1920 pixel resolution video at 30 frames-per-second (fps) over 60 seconds. This dataset was collected for experiments in our

prior work [11] that established our benchmark tool tracking algorithm. We chose to use a white background, a narrow angle field-of-view (FOV), and high intensity illumination to ensure adequate capture of surgical tool motion as well as tool tracking features. The pixel coordinates of the inferior jaw tool tip is labelled in each frame of the recorded video to create a ground truth reference for performance evaluation. Performance is evaluated as a pixel distance error between the projected and manually labelled point as shown in Fig. 3. Percentage error is calculated as the pixel distance error divided by the diagonal length of the image frame in which the pixel error was calculated.

2) *Deformable Tissue Dataset*: We used the same hardware and video specifications as noted in the Structured Environment Dataset. The tool trajectory is from an autonomous trajectory for manipulating deformable soft tissue [13]. This scenario has lower intensity lighting, more surgical tool occlusions, and a wider FOV to mimic a more realistic surgical environment for automation. The total length of this recording was 45 seconds. Similar to the Structured Environment Dataset, we manually labeled the pixel coordinates of the inferior jaw tool tip in each frame of the recorded video for a 2D error metric as shown in Fig. 3. Pixel error and percentage error were calculated as in the Structured Dataset.

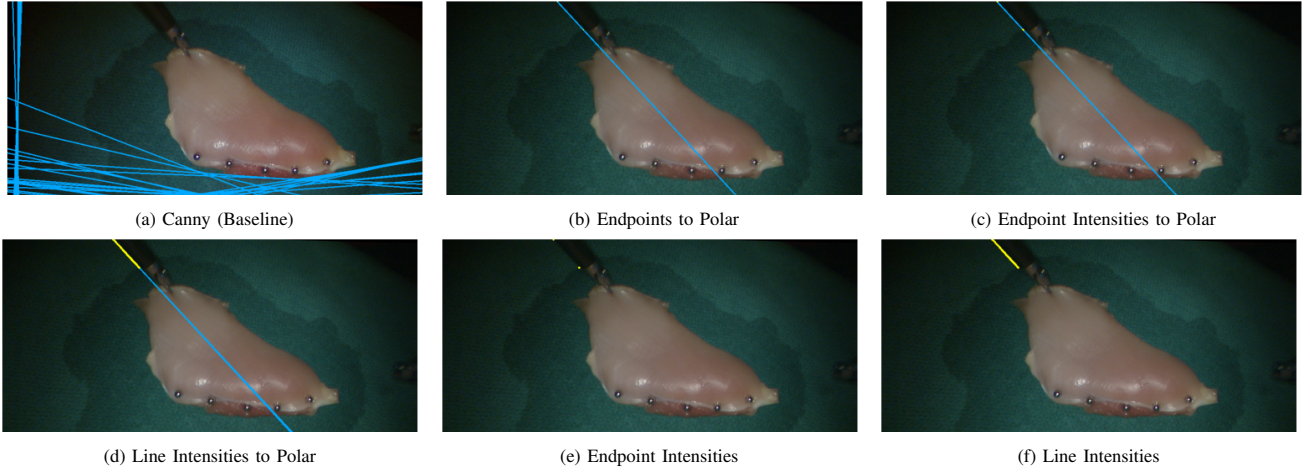


Fig. 6: Illustration of the different observation models for detecting the insertion shaft in the deformable tissue datasets. The blue lines highlight the parameterized lines (i.e. polar parameterization) and the yellow points represent the pixel set used in the observation model. The surgical robotic tool in the deformable tissue dataset was intentionally made challenging to see to create a challenging low light scenario. Even in such a challenging scene, the proposed approach is still able to detect the insertion shaft. This figure is best viewed in color.

C. Experimental Results

Our Structured Environment Dataset results (Table 1) demonstrated that our Line Intensities algorithm had the lowest percentage error (3.1%) relative to the other algorithms. The Endpoint Intensities algorithm had the second lowest percentage error (3.2%) and the lowest standard error (2.0%). With the exception of the Endpoint Intensities to Polar algorithm, all of our algorithms had lower error and standard error than the benchmark Canny algorithm.

We expected the benchmark canny edge detector algorithm to outperform given the extensive hyperparameter tuning for this specific dataset outlined in [11]. However, only the Endpoint Intensities to Polar algorithm demonstrated worse performance than the benchmark on the Structured Dataset, likely due to the challenge of fitting multiple shaft lines to endpoint intensity points in a well-lit, high-quality image. In this environment, additional potential fitted lines from endpoints offer little information gain relative to the lines detected by the Canny benchmark.

Our Deformable Tissue Dataset results demonstrated that 2D error and standard error were lowest for our Endpoint Intensities algorithm: 3.5% and 0.9%, respectively. All of our algorithms, with the exception of Endpoint to Polar, had similar error performance and also outperformed the Canny benchmark. We expected the Endpoint to Polar algorithm to underperform in this environment given the difficulty in identifying endpoints in this more challenging environment. When an endpoint is not detected, the potential endpoint pair that defines a line is removed, and a potential shaft line is ignored. This decreased the performance of both endpoint-to-line fitting algorithms: Endpoint to Polar and Endpoint Intensities to Polar. Endpoint Intensities, however, does not require a line-fitting approach so is able to compensate for missing shaft lines.

We also compute the accumulated error, which is normalized by total number of image data frames hence can decrease when performance improves, and this is shown in

Fig. 4. The accumulated error results for both datasets show that our algorithms outperform the benchmark in both 2D and 3D error metrics. These results are particularly important given the potential of tracking error to compound over time in autonomous surgery applications. In both datasets, our pixel-based methods had the lowest accumulated error. In the Structured Dataset, Line Intensities, Endpoint Intensities, and Line Intensities to Polar had 2D cumulative errors of 3.1%, 3.3%, and 3.2%, respectively. In the Deformable Tissue Dataset, 2D cumulative errors for these algorithms were 3.6%, 3.5%, and 3.6%, respectively.

V. DISCUSSION AND CONCLUSION

Our experimental results demonstrate that our tool tracking algorithms are not only more accurate in structured environments than the current state-of-the-art approach but are also more generalizable to less structured environments. The baseline approach, canny edge detection, has significant difficulty performing well in less structured environments such as low light scenarios. Through our proposed insertion-shaft line detection approach, we are able to more robustly detect the insertion shaft in challenging scenarios as seen in Fig. 5 and 6. Furthermore, our proposed endpoint and line intensity based observation models are able to take full advantage of the predicted heatmap which is why we believe they performed the best of all the comparisons.

Our previous tool tracking approach has already been used for significant surgical automation efforts including autonomous blood suction [5], [50], suture needle tracking [8], [37] and manipulation [7], real-to-sim registration [13], and more general surgical perception [23], [26], [10]. For future work, we intend to continue these research efforts in less structured environments through our proposed novel surgical tool tracking approach. Limitations that will need to be addressed include inference and computation time as the SOLD2 line detection module performs at 1-2 fps.

REFERENCES

- [1] M. Yip and N. Das, "Robot autonomy for surgery," in *The Encyclopedia of MEDICAL ROBOTICS: Volume 1 Minimally Invasive Surgical Robotics*. World Scientific, 2019, pp. 281–313.
- [2] A. Wilcox, J. Kerr, B. Thananjeyan, J. Ichnowski, M. Hwang, S. Paradis, D. Fer, and K. Goldberg, "Learning to localize, grasp, and hand over unmodified surgical needles," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9637–9643.
- [3] A. Murali, S. Sen, B. Kehoe, A. Garg, S. McFarland, S. Patil, W. D. Boyd, S. Lim, P. Abbeel, and K. Goldberg, "Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1202–1209.
- [4] B. Thananjeyan, A. Garg, S. Krishnan, C. Chen, L. Miller, and K. Goldberg, "Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2371–2378.
- [5] F. Richter, S. Shen, F. Liu, J. Huang, E. K. Funk, R. K. Orosco, and M. C. Yip, "Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1383–1390, 2021.
- [6] M. Hwang, D. Seita, B. Thananjeyan, J. Ichnowski, S. Paradis, D. Fer, T. Low, and K. Goldberg, "Applying depth-sensing to automated surgical manipulation with a da vinci robot," in *2020 International Symposium on Medical Robotics (ISMR)*. IEEE, 2020, pp. 22–29.
- [7] Z.-Y. Chiu, F. Richter, E. K. Funk, R. K. Orosco, and M. C. Yip, "Bimanual regrasping for suture needles using reinforcement learning for rapid motion planning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7737–7743.
- [8] Z.-Y. Chiu, A. Z. Liao, F. Richter, B. Johnson, and M. C. Yip, "Markerless suture needle 6d pose tracking with robust uncertainty estimation for autonomous minimally invasive robotic surgery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5286–5292.
- [9] J. Lu, F. Richter, and M. C. Yip, "Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 296–21 306.
- [10] S. Lin, A. J. Miao, J. Lu, S. Yu, Z.-Y. Chiu, F. Richter, and M. C. Yip, "Semantic-super: A semantic-aware surgical perception framework for endoscopic tissue classification, reconstruction, and tracking," *arXiv preprint arXiv:2210.16674*, 2022.
- [11] F. Richter, J. Lu, R. K. Orosco, and M. C. Yip, "Robotic tool tracking under partially visible kinematic chain: A unified approach," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1653–1670, 2021.
- [12] M. Hwang, B. Thananjeyan, S. Paradis, D. Seita, J. Ichnowski, D. Fer, T. Low, and K. Goldberg, "Efficiently calibrating cable-driven surgical robots with rgbd fiducial sensing and recurrent neural networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5937–5944, 2020.
- [13] F. Liu, Z. Li, Y. Han, J. Lu, F. Richter, and M. C. Yip, "Real-to-sim registration of deformable soft tissue with position-based dynamics for surgical robot autonomy," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 328–12 334.
- [14] I. Fassi and G. Legnani, "Hand to sensor calibration: A geometrical interpretation of the matrix equation $ax = xb$," *Journal of Robotic Systems*, vol. 22, no. 9, pp. 497–506, 2005.
- [15] F. C. Park and B. J. Martin, "Robot sensor calibration: solving $ax = xb$ on the euclidean group," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 5, pp. 717–721, 1994.
- [16] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep n p: An accurate o (n) solution to the p n p problem," *International journal of computer vision*, vol. 81, pp. 155–166, 2009.
- [17] J. Lambrecht and L. Kästner, "Towards the usage of synthetic data for marker-less pose estimation of articulated robots in rgb images," in *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 2019, pp. 240–247.
- [18] J. Lambrecht, "Robust few-shot pose estimation of articulated robots using monocular cameras and deep-learning-based keypoint detection," in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*. IEEE, 2019, pp. 136–141.
- [19] T. E. Lee, J. Tremblay, T. To, J. Cheng, T. Mosier, O. Kroemer, D. Fox, and S. Birchfield, "Camera-to-robot pose estimation from a single image," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9426–9432.
- [20] J. Lu, F. Richter, and M. Yip, "Robust keypoint detection and pose estimation of robot manipulators with self-occlusions via sim-to-real transfer," *arXiv preprint arXiv:2010.08054*, 2020.
- [21] F. Zhong, Z. Wang, W. Chen, K. He, Y. Wang, and Y.-H. Liu, "Hand-eye calibration of surgical instrument for robotic surgery using interactive manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1540–1547, 2020.
- [22] T. Zhao, W. Zhao, B. D. Hoffman, W. C. Nowlin, and H. Hui, "Efficient vision and kinematic data fusion for robotic surgical instruments and other applications," Mar. 3 2015, uS Patent 8,971,597.
- [23] 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.
- [24] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15*. Springer, 2012, pp. 592–600.
- [25] —, "Appearance learning for 3d tracking of robotic surgical tools," *The International Journal of Robotics Research*, vol. 33, no. 2, pp. 342–356, 2014.
- [26] J. Lu, A. Jayakumari, F. Richter, Y. Li, and M. C. Yip, "Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4783–4789.
- [27] P. Pastor, M. Kalakrishnan, J. Binney, J. Kelly, L. Righetti, G. Sukhatme, and S. Schaal, "Learning task error models for manipulation," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 2612–2618.
- [28] F. Wang, K. Chen, and X. Chen, "An online calibration method for manipulator with joint clearance," *Robot*, vol. 35, no. 5, pp. 521–526, 2013.
- [29] H. Peng, X. Yang, Y.-H. Su, and B. Hannaford, "Real-time data driven precision estimator for raven-ii surgical robot end effector position," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 350–356.
- [30] J. Mahler, S. Krishnan, M. Laskey, S. Sen, A. Murali, B. Kehoe, S. Patil, J. Wang, M. Franklin, P. Abbeel, *et al.*, "Learning accurate kinematic control of cable-driven surgical robots using data cleaning and gaussian process regression," in *2014 IEEE international conference on automation science and engineering (CASE)*. IEEE, 2014, pp. 532–539.
- [31] M. Haghighipanah, M. Miyasaka, Y. Li, and B. Hannaford, "Unscented kalman filter and 3d vision to improve cable driven surgical robot joint angle estimation," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4135–4142.
- [32] D. Seita, S. Krishnan, R. Fox, S. McKinley, J. Canny, and K. Goldberg, "Fast and reliable autonomous surgical debridement with cable-driven robots using a two-phase calibration procedure," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6651–6658.
- [33] V. Pradeep, K. Konolige, and E. Berger, "Calibrating a multi-arm multi-sensor robot: A bundle adjustment approach," in *Experimental Robotics: The 12th International Symposium on Experimental Robotics*. Springer, 2014, pp. 211–225.
- [34] Q. V. Le and A. Y. Ng, "Joint calibration of multiple sensors," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3651–3658.
- [35] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and object tracking for in hand model acquisition," in *Proceedings, IEEE International Conference on Robots and Automation*, 2010, pp. 1817–1824.
- [36] C. G. Cifuentes, J. Issac, M. Wüthrich, S. Schaal, and J. Bohg, "Probabilistic articulated real-time tracking for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 577–584, 2016.
- [37] Z.-Y. Chiu, F. Richter, and M. C. Yip, "Real-time constrained 6d object-pose tracking of an in-hand suture needle for minimally invasive robotic surgery," *arXiv preprint arXiv:2210.11973*, 2022.
- [38] B. Lu, B. Li, Q. Dou, and Y. Liu, "A unified monocular camera-based and pattern-free hand-to-eye calibration algorithm for surgical robots with rcm constraints," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 6, pp. 5124–5135, 2022.

- [39] F. Chaumette, "La relation vision-commande: théorie et application à des tâches robotiques," Ph.D. dissertation, L'Université de Rennes I, 1990.
- [40] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci® surgical system," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 6434–6439.
- [41] M. J. Lum, D. C. Friedman, G. Sankaranarayanan, H. King, K. Fodero, R. Leuschke, B. Hannaford, J. Rosen, and M. N. Sinanan, "The raven: Design and validation of a telesurgery system," *The International Journal of Robotics Research*, vol. 28, no. 9, pp. 1183–1197, 2009.
- [42] B. Hannaford, J. Rosen, D. W. Friedman, H. King, P. Roan, L. Cheng, D. Glozman, J. Ma, S. N. Kosari, and L. White, "Raven-ii: an open platform for surgical robotics research," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 954–959, 2012.
- [43] R. Pautrat, J.-T. Lin, V. Larsson, M. R. Oswald, and M. Pollefeys, "Sold2: Self-supervised occlusion-aware line description and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 368–11 378.
- [44] F. Richter, E. K. Funk, W. S. Park, R. K. Orosco, and M. C. Yip, "From bench to bedside: The first live robotic surgery on the dvkr to enable remote telesurgery with motion scaling," in *2021 International Symposium on Medical Robotics (ISMR)*. IEEE, 2021, pp. 1–7.
- [45] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.
- [46] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.
- [47] W. E. Smith, N. Vakil, and S. A. Maislin, "Correction of distortion in endoscope images," *IEEE Transactions on Medical Imaging*, vol. 11, no. 1, pp. 117–122, 1992.
- [48] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.
- [49] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [50] J. Huang, F. Liu, F. Richter, and M. C. Yip, "Model-predictive control of blood suction for surgical hemostasis using differentiable fluid simulations," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 380–12 386.