IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

OCT Imaging for Pose Estimation and Feedback Control of an Articulated Magnetic Surgical Tool

Erik Fredin[1*], Nirmal Pol[2*], Anton Zaliznyi[3], Dmytro Fishman[3], Eric Diller[1,2,4] and Lueder A. Kahrs[2,4,5]

***Abstract*—Magnetically-driven surgical tools are a new class of millimetre-scale devices that could enable procedures such as minimally invasive neurosurgery due to their high dexterity at a small size. However, safe and effective control of these magnetic tools necessitates real-time observation of tool joint angles, which is challenging inside a surgical environment. Optical coherence tomography (OCT) is an emerging volumetric imaging technique offering 3D visualization of tissue and tools simultaneously, which we explore for joint angle estimation. While some previous studies have used OCT for estimating the pose of rigid instruments, those methods are specific to needle-like tools, and often have slow processing speed. In this work, we benchmark eight deep-learning models adapted from other 3D modalities to OCT data showing magnetic tools in a mock surgical environment. The models are tested in the presence of other objects, occlusion, noise, and the tool being partially outside of the OCT's field of view. The best performing model, VoxelNeXt, is adapted from 3D object detection in LiDAR scans, the first time a model of this kind is used on medical data. It infers tool pose with 0.6 mm position and 5° angular errors, with 40 ms inference time. We use this model to provide feedback for controlling a multi-jointed magnetic tool, demonstrating the robustness of OCT-based feedback control. Code and dataset are available at https://medcvr.utm.utoronto.ca/ral2025-oct-pose.html.**

***Index Terms*—Computer Vision for Medical Robotics, Machine Learning for Robot Control, Deep Learning for Visual Perception**

I. INTRODUCTION

Minimally-invasive surgery (MIS) offers the potential for surgical procedures to be performed with smaller incisions, shorter hospital stays, and reduced trauma, as compared to open surgery. Robot-assisted MIS is used in disciplines such as gynecology, urology and gastroterology [1], but robotic MIS remains relatively unexplored for neurosurgery, primarily due to challenges in miniaturizing conventional tools with sufficient dexterity and maintaining accurate observation of the tool and tissue for precise control.

Over the past two decades, magnetic microrobots have shown increased potential to perform minimally-invasive medical procedures [2]. Small magnets are embedded in the tip of tools, with magnetic fields generated outside the body being used to steer the tools wirelessly. Notable examples of such tools include magnetic capsules [3], needles [4] and catheters [5]. Applications such as neurosurgery, which require particularly small tools only a few millimeters across, provide a particularly challenging scenario. Hong et al. proposed a flexible needle for neurosurgery, capable of magnetically navigating narrow neurosurgical sites [6]. Forbrigger et al. proposed a magnetic gripper capable of grasping tissue and applying forces of up to 200 mN [7]. A variation on this magnetic gripper, with a diameter of 4 mm, is used for this study (Fig. 1). A delivery platform (here a straight metal tube) positions and orients the gripper base, while its joint angles are controlled by an external magnetic coil system. While precise tool joint control is possible with magnetic fields, effective control of these multi-jointed magnetic tools requires feedback on its joint angles, which is particularly challenging given the constrained environment of minimally invasive neurosurgery. Feedback can be obtained by using medical imaging systems like ultrasound, magnetic resonance imaging (MRI), endoscopy, or optical coherence tomography (OCT). However, ultrasound has limitations in image resolution, and MRI interferes with the tool's magnetic actuation. Endoscopy has been explored for estimating the pose of surgical tools [8], [9], but these endoscopic approaches face challenges from variable lighting, glare, and poor subsurface imaging capabilities. Furthermore, endoscopic feedback faces difficulties with determining the ground truth on the tool's

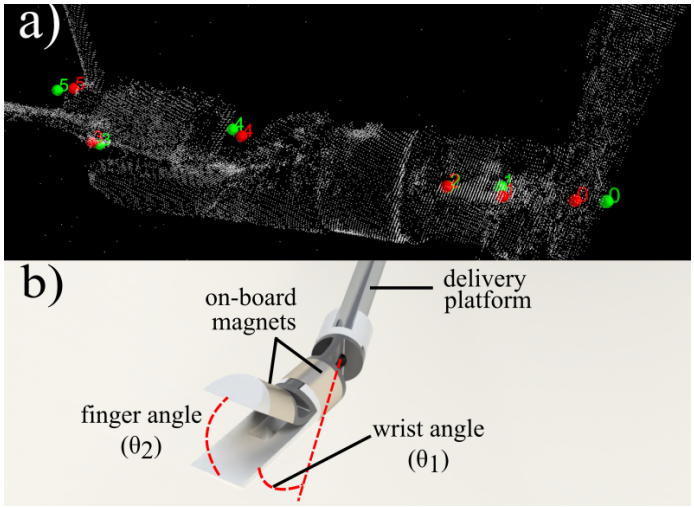

Fig. 1. An articulated magnetic surgical instrument in a) an OCT scan, with detected (red) and ground-truth (green) keypoints, and b) in illustrated concept showing the key component definitions.

Manuscript received: May, 28, 2025; Revised September, 27, 2025; Accepted November, 20, 2025

This paper was recommended for publication by Editor Jens Kober upon evaluation of the Associate Editor and Reviewers' comments.

*Authors contributed equally to this work.

[1]Department of Mechanical and Industrial Engineering, University of Toronto, Canada

[2]Institute of Biomedical Engineering, University of Toronto, Canada

[3]Institute of Computer Science, University of Tartu, Estonia

[4]Robotics Institute, University of Toronto, Canada

[5]Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Canada

Corresponding author: Erik Fredin (erik.fredin98@gmail.com)

3D pose. However, 3D OCT scans allow for ground truth annotation from just the scans themselves. Because of this, and their ability to generate high-resolution 3D images, we examine OCT for providing pose-feedback in this work.

OCT can generate high-resolution 3D images of small tools using low-coherence interferometry, and also possesses sub-surface tissue imaging capabilities. OCT has been previously explored for pose estimation of rigid surgical instruments in retinal surgery, though these methods lack real-time performance [10]–[12]. Gessert et al. investigated two deep-learning models to estimate the pose of a rigid marker using OCT scans [13]. While their models processed scans in real-time, their approach requires modifications of surgical tools using their markers.

In summary, while magnetic tools require pose feedback for effective control, current pose-estimation methods for OCT images have not been used for tracking markerless jointed surgical tools. This work proposes a deep learning framework for estimating the pose of an instrument—similar in functionality to [7], [14]—from OCT volumes acquired at 20 Hz, which is faster than current conventional OCT systems. The framework achieves 0.6 mm positional accuracy and 5° joint angle error under challenging conditions such as occlusion, the presence of other objects, and mirroring, and further demonstrates its efficacy for feedback control of a magnetic tool. We present the following contributions:

- We evaluate eight diverse model architectures for markerless keypoint detection without needing physical modification, uncovering that a sparse CNN — leveraging techniques from LiDAR and self-driving community — yield substantially better accuracy than dense CNNs.
- Demonstration of high-speed volumetric OCT feedback control by integrating our best-performing pose estimator into a closed-loop pipeline that drives our small, articulated magnetic tool.
- We provide our annotated volumetric OCT dataset for multi-jointed surgical-tool pose estimation, containing realistic imaging artifacts (e.g., speckle noise, shadowing, mirroring), occlusions, and partially out-of-view configurations, and uniquely supporting 8-DoF labels; prior OCT pose datasets are few [15] and limited to 6-DoF for rigid needle tools [11], [15].

II. RELATED WORKS

A. Pose Estimation in OCT Images

The use of OCT for pose estimation of surgical instruments has thus far been limited to rigid needles. Some works [10], [11] segment individual slices and then calculate the needle pose by composing the segmentation data and using additional algorithms, such as ICP. While these methods localize needles with micrometer precision, they have a slow inference time of 200-300 ms, and thus too slow for feedback control. Gessert et al. proposed the 3D CNNs Inception3D and ResNeXt3D to directly regress a marker's pose [13]. This approach was both accurate and fast, achieving a mean absolute error of 14-21 µm and an average speed of 21 ms for a resolution of 64x64x16. However, their models have yet to be tested on marker-free tools, warranting further investigation. Gessert et al. also showed that 3D OCT volumes can be compressed into 2D maximum intensity projection (MIP) and normalized depth images taken from canonical directions (axial, coronal, sagittal). Their best 2D results were obtained by pairing a single enface-view MIP with its corresponding depth-normalized image extracted from the same volume, yet this 2D approach lagged behind their 3D models.

B. Deep learning for Medical Image Segmentation

While the use of deep-learning (DL) for pose estimation on OCT images is still limited, it is more common for segmentation. The U-Net [16] was a major breakthrough for medical image segmentation and DL in general. It introduced an autoencoder structure, which is still used in both segmentation and pose estimation today. Lee et al. applied a modified U-Net architecture for segmentation on OCT images [17]. Many OCT segmentation studies between 2015-19 used variations of either the U-Net or the DenseNet [18] architecture [19]. Models for MRI or CT scans continued to build on the autoencoder structure of the U-Net [20], [21]. Notably, MedNeXt uses this structure and takes inspiration from transformers [22], and leads the AMOS22 benchmark. Autoencoder-based segmentation models can be adapted to pose estimation, since their structure can also be used to generate heatmaps to detect keypoints.

C. Hand-Pose Estimation

Deep learning-based pose estimation has been explored extensively for 3D DL domains outside of medical images. Hand pose estimation uses RGB-D images to train DL models, and requires estimating a high number of DoFs. To achieve this, RGB-D images are converted to 3D voxel volumes, making this approach similar to OCT imaging which generates voxel volumes directly. Moon et al. proposed the V2V-PoseNet, marking an important baseline for hand pose estimation [23], though its performance degrades under occlusion or severe viewpoint variation. To address these challenges, Cheng et al. developed a virtual view selection and fusion framework [24], where input depth is re-projected into multiple virtual views. Their method selects the most informative views using a learned confidence network and fuses the pose estimates, resulting in improved accuracy.

D. 3D Object Detection

3D object detection in autonomous driving estimates 3D bounding boxes of objects such as cars, pedestrians and bikes in LiDAR data. LiDAR data is typically very sparse, rendering dense convolutions quite inefficient for processing it. The method SECOND [25] introduced sparse convolutions [26] which skip empty voxels, and are significantly more efficient than dense convolutions. Sparse CNNs have become backbone networks for many 3D object detection approaches [27]–[29]. However, these rely on architectural features such as anchors or birds eye view (BEV) projection, which assume that objects lie on a known 2D plane. While this holds true for cars driving

on a road, it does not for surgical tools in OCT data. On the other hand, VoxelNeXt is a comparatively simple model for estimating 3D bounding boxes, yet still achieves state-of-the-art results [30]. It features a sparse backbone network, followed by a sparse head that classifies individual voxels and regresses bounding box locations from them. While it projects 3D data to BEV, the model can be modified to process the data entirely in 3D.

E. Model Selection

Based on this review, we select the following models for pose estimation on OCT images: Gessert et al. (ResNeXt-3D and Inception-3D) [13], V2V Pose-Net [23] and MedNeXt [22]. We also test a 3D implementation of ResNet-50 [31]. Inspired by [13] and [24], we implement a multi-view approach (Section IV-B). Finally, we adapt VoxelNeXt [30] to pose estimation in OCT images (Section IV-C).

III. DATASET GENERATION

A. Dataset Overview

A total of 793 OCT volumes were collected which contained the surgical tool in diverse joint configurations, orientations, and positions. The volumes were captured with a shape of (1096, 1936, 1152) where 1096 refers to the number of cross-sectional slices (i.e. B-scans). The voxel dimension was measured to be approximately 16 x 14 x 9 μm^3. Each of the volumes were also accompanied with their corresponding downsampled version of (28, 1936, 1152) which are captured at approximately 20 Hz from OptoRes GmbH OMES System. The larger volumes are helpful for annotations while their corresponding smaller volumes are necessary for real-time processing.

Our dataset contains OCT volumes of the gripper tool in ideal conditions where the tool was fully visible, with no occlusions, artifacts, external objects, or out of the field-of-view (FOV). In this *No Artifact* dataset, only one joint was actuated at a time and kept static during imaging. This group also contained some sequential volumes with continuous elbow or jaw motion which did not contain any artifacts. These sequential volumes were excluded from validation and testing due to their minimal inter-volume variation.

The dataset also includes artifacts such as external objects in the scene (e.g., scissors, rods, tissue phantoms), occlusions (via object placement or shadowing), partial tool visibility (out-of-FOV), and OCT-specific mirroring artifacts caused by the tool crossing the zero-delay line. The mirroring artifact, unique to Fourier Domain OCT systems, occurs when the object crosses the zero-delay line and appears flipped or partially flipped on itself. We included mirroring artifacts because they are common in current commercial OCT systems, making the dataset more representative of real-world use. A combination of these artifacts were also collected. Fig. 2 shows representative examples, and Table I summarizes the dataset distribution. Volumes were split into train/validation/test sets within each artifact category.

While many of the mentioned artifacts are not as prevalent or labelled in previous OCT related works [11]–[13], [15], their inclusion in our dataset allows us to examine potential failure cases that can exist for an OCT based pose estimation model. The training dataset was augmented by randomly translating the gripper to different positions, applying random intensity, contrast variations and also randomly moving the gripper partially out of view. This resulted in a total training dataset size of 2236 volumes.

TABLE I
VOLUMES PER CATEGORY FOR TRAIN, VALIDATION, AND TEST SETS

Grouped Category	Train	Validation	Test
No Artifacts	257	26	32
Occlusion	25	5	7
Partially Out of View	18	4	5
Mirror	24	6	8
Other Objects	34	8	11
2 Artifact	125	25	45
3 Artifact	64	15	29
4 Artifact	12	2	6
Total	559	91	143

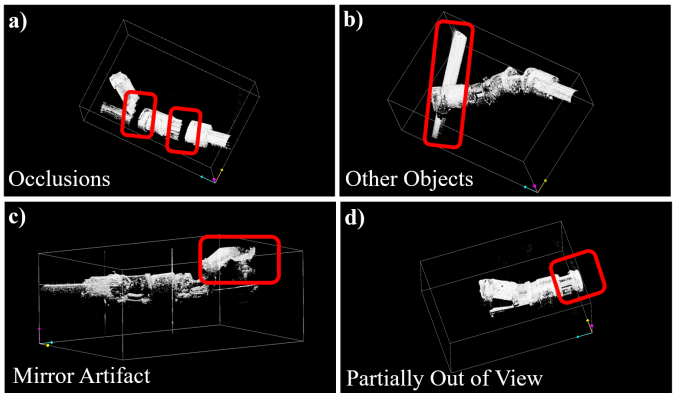

Fig. 2. Example OCT volumes and artifacts in dataset.

B. Dataset Annotation

Annotating a 3D volumetric dataset for pose estimation is a challenging and time consuming process. We developed a 3D annotation method for volumetric data using Mayavi [32]. The datasets were annotated by converting the voxels to point clouds and labelling the 3D point clouds in Mayavi. The annotators were required to position 3 bounding boxes (each containing 2 key points at the front and back surfaces of the boxes) at the appropriate locations on the gripper in the image. The bounding boxes were placed on the gripper using sliders or the annotators could input the voxel coordinates. The boxes were also fixed in size to indicate which bounding box must be placed on their respective links on the gripper. Fig. 3 show an example placement of these boxes on each link. This approach for annotating the 3D volumetric data resulted in an average inter-annotator error of 0.3 degrees for joints angles and a positional error of 19 μm for placement of keypoints. To ensure reliability, we minimized bias by cross-checking annotations across multiple annotators, reconciling disagreements, and flagging rare ambiguous cases for adjudication. The time to annotate varied between 2-5 minutes per volume depending

on the complexity of artifacts in the volume. We provide this annotation method in our repository.

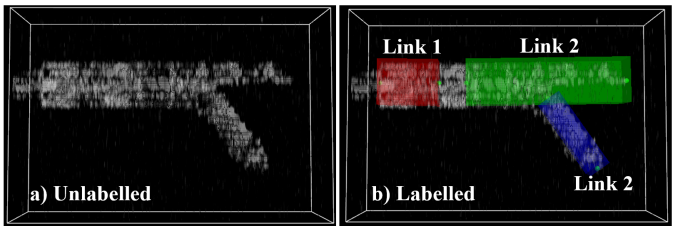

Fig. 3. (a) Unlabelled OCT volume, (b) Labelled OCT volume with bounding boxes for each link.

IV. 8-DOF POSE ESTIMATION

A. Overview

The goal of pose estimation in this work is to estimate the 6-DoF position and orientation of the tool, as well as its joint angles θ_1 and θ_2 (see Fig. 1). We use a keypoint detection approach to this end. Fig. 4(a) shows the placement of the keypoints on the tool. Six keypoints (shown in blue) are used to calculate the tool's joint angles. These keypoints are placed at the end of each link, which enables the calculation of link orientations and joint angles from them. The wrist features further keypoints (shown in green), which are used to determine the tool's orientation. This is achieved via a Kabsch-Umeyama algorithm performing registration between these detected keypoints and their modeled counterparts. Keypoints on links other than the wrist are not used, as errors in joint angle estimation would affect registration if they were.

The training and validation data from Table I was used for training and hyperparameter tuning. The final benchmarks in Table II are performed on the test data. Finally, Table III shows the performance of our best model on each category.

B. Multi-View MIP

Inspired by the results for 25 viewpoints in [24], we also adapted a multi-view selection approach using 25 uniformly spaced viewpoints distributed over the upper hemisphere of each volume. Then similar to [13], each view generates a 2D MIP and normalized depth of the volume. This is processed by a shared lightweight backbone to extract features. A dedicated confidence head then assigns a score to each view based on these features. Simultaneously, a pose head generates an initial pose prediction for each view using the same features. Finally, leveraging the confidence scores, the pose predictions from the top 15 most informative views are selected and combined through a weighted average. We also implement Inception-2D model for the single enface-view MIP and normalized depth approach from [13] as a baseline.

C. VoxelNeXt

Fig. 4(b) shows how VoxelNeXt is used for keypoint detection on OCT data. We apply thresholding to raw OCT scans to filter out low-intensity voxels, producing a sparse voxel volume. The raw OCT data collected is highly imbalanced,

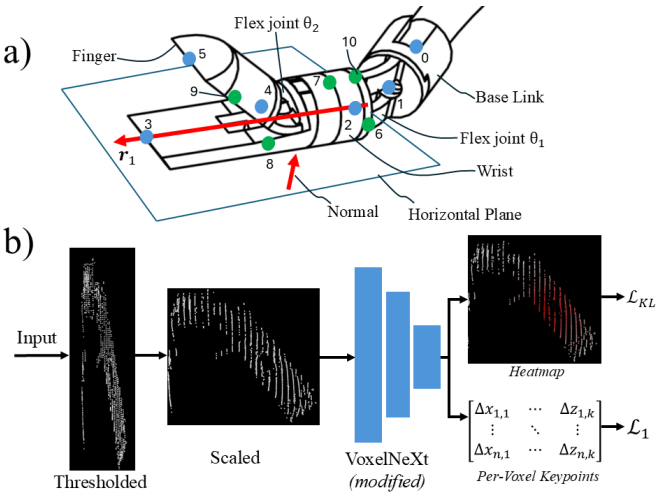

Fig. 4. a) Location of keypoints on surgical tool. Blue keypoints are used to compute joint angles, green keypoints to compute tool orientation and blue/green keypoints for both. b) Overview of our VoxelNeXt-based approach for keypoint detection.

having a dimension of (28, 242, 144) after downsampling along the y- and z-dimensions. VoxelNeXt regresses keypoint coordinates, and high imbalances between the dimensions in the volume can lead to sub-optimal training for this task. To address this, each axis is scaled to match the scan's real-world size as seen in Fig. 4(b). This is possible because sparse voxel volumes store voxel coordinates in a list (similar to point clouds), in contrast to dense volumes.

The main architectural change is that the present adaptation estimates only keypoints rather than bounding boxes. Further, it estimates a fixed number of keypoints, M, in contrast to the original, which detects a variable number of objects between different scans. For N voxels, our adaptation generates a $N \times M$ heatmap and a $N \times 3M$ per-voxel location estimate for each keypoint. During inference, the K highest scoring voxels for each keypoint are selected, and their weighted average is used as the final estimate, where K is a configurable hyperparameter. While the original architecture uses a Focal Loss [33] for training heatmaps, we empirically found a KL-divergence loss to work better. This models the heatmaps as a distribution, rather than as individual probabilities for each voxel. In addition, we place the heatmaps' centers at the keypoint's location, in contrast to the original, which places them at the closest voxel. Finally, we removed all components of the model that used 2D convolutional operations, treating the data as a 3D volume throughout all layers.

V. FEEDBACK CONTROL

A. Setup

To use OCT scanning as feedback, we mount the OptoRes OCT probe on top of the workspace where the tool is controlled (Fig. 5). The electromagnetic coil system (ECS) used to control the tool consists of 8 coils mounted below. The coil system therefore has no more space constraints than a table. The tool is attached to a delivery platform consisting of two

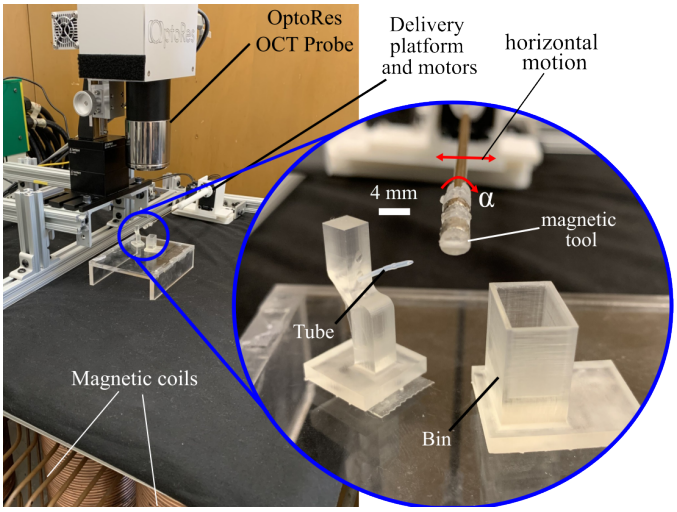

Fig. 5. Experimental setup for the control experiments, in particular the pick and drop-off experiment. An OCT probe is mounted above the workspace. The tool is attached to a delivery platform, which translates it horizontally and rotates it around it's centre axis.

motors. The first motor rotates the tool around its center axis (α), while the second motor moves it horizontally (Fig. 5). This adds 2 DoF to the tool's magnetically actuated joints, meaning the system has 4-DoF overall.

The OCT scans must be collected and processed with minimal latency to enable feedback control. The scans are collected in the (28, 1936, 1152) format, down-sampled to (28, 242, 144) and low intensity voxels are filtered out (as described in Section IV-C). We use 2 NVIDIA RTX 2080-Ti GPUs, allowing for scan collection at 20 Hz and pose estimation at 20-25Hz. These two processes are computed in parallel, meaning that the next scan is collected as the pose is estimated on the current scan. We use the best performing model (see Table II) to estimate tool pose. Due to hardware constraints, the coil system is controlled by a separate computer, with data transfer introducing an additional 1-5 ms latency.

Inspired by [7], we use a PID controller to determine the currents, $\mathbf{u}$, needed for controlling the tool:

$$\mathbf{u} = \mathbf{B}^{\dagger}(\mathbf{q}) \left(K_p \mathbf{e} + K_d \dot{\mathbf{e}} + K_i \int \mathbf{e}\, dt + K_r \tau_S(\mathbf{q_d}) \right), \quad (1)$$

where $\mathbf{B}^{\dagger}$ is the pseudo-inverse of the current actuation matrix, $\mathbf{q} = [\alpha, \theta_1, \theta_2]$ is the state of the tool, $\mathbf{e} = \mathbf{q}_d - \mathbf{q}$ is the state error. K_p, K_d, K_I and K_r are tuning parameters. Besides the standard PID terms, the controller contains $\tau_S = \tau_{int}(\mathbf{q_d}) + \mathbf{k q_d}$, where τ_{int} are the forces between the tool's on-board magnets and $\mathbf{kq}$ models the tool's flexure joints. These directional forces act unevenly on the tool, and accounting for them in the controller thus ensures smoother behavior and easier tuning. Feedback on $\mathbf{q}$ is provided via the OCT system. The tool's position is not included in $\mathbf{q}$, since we confine experiments to the coil system's centre. The centre features homogenous fields, meaning that $\mathbf{B}^{\dagger}$ is independent of the tool's position within this area. $\mathbf{B}^{\dagger}$ has been calibrated via the method described in [34].

B. Experiments

1) Step-Response: To assess the PID controller's ability to respond to a control input, we record several step responses for both joints. The tool is initially placed into a neutral configuration, and the PID controller receives setpoints of $(10°, 20°, 30°, 40°, 50°)$ for a given joint. Each input lasts 10 seconds before changing to the next. Step-responses are tested for each joint individually. To minimize rise-times and overshoots, we tune the PID controller using gain scheduling. New controller gains are chosen every $10°$ for both joints. The range for these controller gains are: $K_p = 0.7 - 2.5$ mN/rad, $K_d = 0 - 0.3$ mN/rad, $K_i = 0 - 0.28$ mN/rad, and $K_r = 0 - 0.45$ mN.

2) Pick and Drop-Off: We conduct a pick and drop-off experiment to test the tool's dexterity. The aim is to pick up a tube protruding out of a fixture and dropping this tube into a bin (Fig. 5). The tool is tele-operated using a gamepad controller. The time taken to pick up the tube and the time taken to place it are recorded separately. The tool picking up the tube and holding it for more than 1 second is considered a successful pickup, while dropping it into the bin is considered a successful drop-off. If a trial completes both successfully it is marked as a success, if only pickup is successful it is a partial success, and it is considered a failure if neither are successful. In total, we collected 18 trials. To demonstrate efficacy of closed-loop PID control, we compare it using the same experiment with an open-loop method.

3) Handover Experiment: We conduct a hand-over experiment in which OCT distinguishes two visually identical tubes—one empty and one containing 0.2% intralipid, via speckle-variance analysis. Fluid detection uses speckle-variance OCT, where N repeated volumes $I_i(x, y, z)$ are acquired at the same location, and the voxel-wise variance σ^2 highlights the dynamic molecular scattering in the fluid. Regions with motion exhibit elevated σ^2 relative to static tissue, providing intrinsic contrast and automatic differentiation of moving objects. The gripper selects the fluid-filled tube and passes it to a secondary tweezer tool, demonstrating OCT's dual role in diagnosis and control.

VI. RESULTS AND DISCUSSION

A. Pose Estimation

Table II shows the performance of all models discussed in Section IV on the test dataset. The position error is the mean keypoint detection errors in mm for all 11 points. θ_1 and θ_2 show the mean errors for the two joint angles. R_x, R_y and R_z show the orientation errors, defined as the Euler angles between the tool's estimated and ground-truth SE(3) rotations. VoxelNeXt leads by a wide margin, achieving a keypoint position error of 0.6 mm. Consequently, the orientation and joint angle errors of VoxelNeXt are also lower than all other models. The θ_1 error of just $3.5°$ is particularly important, since accurate feedback on the tool's wrist joint is instrumental for precise feedback control. θ_2 has a higher error at $6.6°$, which is acceptable since the finger requires less precise control as it is often either fully open or fully closed. VoxelNeXt's orientation errors range from around $2.4° - 6.1°$.

TABLE II
COMPARISON OF POSE ESTIMATION MODELS, VALUES SHOWN REPRESENT AVERAGE ERROR TO GROUND TRUTH

Method	x, y, z (mm)	R_x (°)	R_y (°)	R_z (°)	θ_1 (°)	θ_2 (°)
ResNet	2.3	26.2	13.1	16.8	12.8	12.0
ResNeXt3D	1.8	13.7	7.1	7.7	8.4	9.7
Inception3D	1.6	17.9	7.8	9.3	9.3	10.4
V2VPoseNet	3.1	28.3	12.5	35.6	35.6	43.0
MedNext	3.5	34.8	14.3	30.5	33.8	30.8
MIP	2.2	22.5	11.2	11.2	13.8	13.9
Multi-view MIP	1.9	16.1	7.7	9.8	9.8	9.9
VoxelNeXt	**0.6**	**6.1**	**2.4**	**3.9**	**3.6**	**6.5**

TABLE III
VOXELNEXT PERFORMANCE WHEN DEALING WITH DIFFERENT CHALLENGES IN OCT VOLUME. VALUES SHOWN REPRESENT AVERAGE ERROR

	No Artifacts	**Mirror Artifact**	**Occlusion**	**Out of View**	**Other Objects in FOV**	**Multiple Artifacts**
Position & Orientation Error						
Position (mm)	0.3	0.6	0.8	0.8	0.6	0.7
R_x (°)	1.5	4.2	4.9	8.0	7.3	8.0
R_y (°)	1.0	2.5	3.3	4.9	2.0	2.8
R_z (°)	1.0	2.6	2.4	3.4	3.3	5.4
Joint Error (°)						
θ_1	1.9	2.2	4.1	4.6	3.4	4.4
θ_2	1.6	13.1	2.9	16.4	5.9	7.6

While these errors show promise, they have the potential to disrupt safe and effective magnetic actuation. Lowering these errors, along with θ_2 error, is an important area of future work.

The second best performing model, Inception3D, has a position error over 260% higher, at 1.6 mm with similar performance as ResNeXt3D. CNNs directly regressing keypoint coordinates (Inception3D, ResNeXt3D and ResNet3D) outperformed those that generate heatmaps for keypoint detection (MedNext, V2V PoseNet). A potential explanation could be the large imbalance between the x and y, z axes of the scans. This does not affect regression based models, since their regressed target coordinates are normalized.

The Inception-2D MIP model performed worse than the multi-view MIP model. Diverse viewpoints enable certain MIP images to indicate the keypoint locations, whereas other viewpoints are more ambiguous to evaluate as the MIP image might show the gripper occluding itself. We also note that directly regressing the 8-DoF pose instead of the keypoints leads to significantly poorer performance. Although the multi-view strategy yielded modest improvements, its performance under occlusion and out-of-view conditions remained unsatisfactory, as the original volumes (already downsampled to 28 B-scans) are further compressed into 2D projections that obscure fine details and make tool pose identification difficult even upon visual inspection. This loss of depth information and inter-slice context limits the approach to outperforming VoxelNeXt.

We also analyze VoxelNeXt's performance in scans containing mirror artifacts, occlusions, gripper partially out of view, gripper with other objects in the scene, and any combination of these artifacts in Table III. The model performs well in position, orientation, and joint angle estimation when there are no artifacts. In the presence of mirror artifacts, there is an increase in θ_2 error. This is expected, because the jaw link 3 is most likely to reach the top of the volume workspace causing the jaw to mirror downwards into the volume. The model examines local regions and correctly identifies the link 3 tip where the keypoint would typically be present. But it fails to recognize that link 3 has partially mirrored downwards in the volume as shown in Fig. 6. A potential mitigation strategy could entail adding a detection head to detect mirroring and then reflecting the predicted keypoint 5 across the zero-delay plane. When there are occlusions present, there is a small increase in errors. Similarly, out of view scenarios present a larger increase in error for the second joint. The presence of other objects in the scene also shows an increase in errors if objects in the workspace contain features similar to the gripper. Finally, in the presence of multiple artifacts, we observe an increase of 2.5° and 6.0° in error for θ_1 and θ_2 respectively.

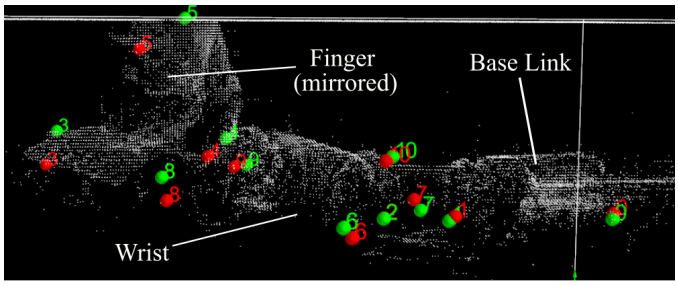

Fig. 6. Ground-truth keypoints (green) vs detected keypoints (red) using VoxelNeXt on a scan where the tool is partially out of view and mirrored.

B. Feedback Control

1) PID Step-Responses: Fig. 7 shows the step response behavior of the tuned PID controller. Overall, the controller achieved an average rise time of 0.75 seconds and an average overshoot of 6.5% across both joints. Both metrics are promising, and indicate similar step-response behavior as other magnetic tools [7]. However, at higher setpoints (especially

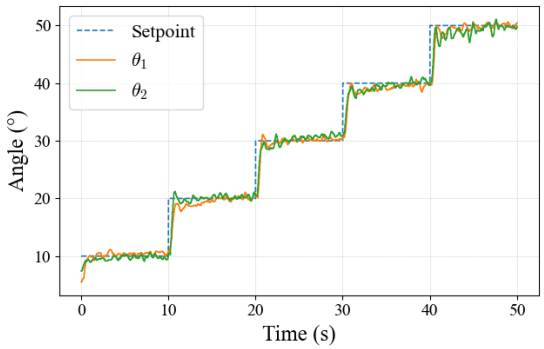

Fig. 7. Step response of the tool wrist angle under PID control

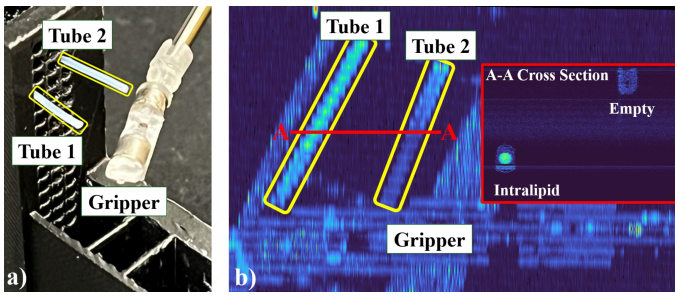

Fig. 8. a) Placement of tubes relative to gripper, b) Top view of speckle variance OCT volume with red line indicating region where cross section was taken. Brighter regions show more motion. Tubes are outlined in yellow.

50°) both joints experience minor increases in oscillation. This highlights the system's complex dynamics, where forces such as the inter-magnet force (τ_{int}) and joint stiffness potentially have a greater effect at higher setpoints. A limitation to our system is the slow rate of feedback. The system takes 50 ms to collect a scan and another 40-50 ms to estimate the tool's joint angles. Since these processes are parallelized, our OCT feedback has a sampling rate of 20 Hz, with each pose estimate having a delay of 100 ms. This reduces the system's ability to quickly adjust the torque during actuation. Studies in surgical latency report that greater delays lead to longer task completion times, increased tool motion, and a higher frequency of errors [35]. A key area of future work is to test with faster OCT hardware and optimize inference.

2) Pick and Drop-Off: Table IV summarizes the performance of the magnetic gripper when executing a pick-and-drop routine with and without OCT feedback. In the open-loop condition (no imaging or pose updates after the initial command), the gripper succeeded in lifting the target tube in only 50% of the trials. Failures arose because (i) the jaws did not close firmly around the tube, and (ii) the grasp weakened as Joint-1 bent to the drop-off position, causing premature release. Accurate drop-off was achieved in 5.6% of attempts.

TABLE IV
AVERAGE RESULTS FROM THE PICK UP AND DROP-OFF CONTROL EXPERIMENT

	Open-Loop	**Closed-Loop**
Average pick-up time (s)	37.1	**13.4**
Average pick-up successes (%)	50	**94.4**
Average drop-off time (s)	**12**	15.4
Average drop-off successes (%)	5.6	**83.3**

Introducing closed-loop control with OCT improved the task execution. Continuous pose estimates allowed the controller to maintain a tight grasp, yielding an 83.3% success rate for pick and placement. The combined pick and drop off time is 28.9 seconds for closed-loop control, and thus faster than open-loop (49.1 seconds). These results highlight the fragility of open-loop control for magnetic tools, as this relies on heavily accurate modeling, fabrication and lack of disturbances. These conditions are in practice infeasible to achieve in real-world settings, leading to the poor results for open-loop control. OCT-based closed loop control on the other hand enables robust control over gripping the tube and dropping it into the bin. Furthermore, the placement of the pick-up and drop-off locations required the tool to move partially out of the FOV of the OCT workspace. Despite this challenge, the feedback controller can complete the task majority of the time.

3) Handover Experiment: Fig. 8 illustrates a pick-and-pass scenario in which two micro-tubes are presented in the workspace—one filled with Intralipid solution, the other empty. The contents are indistinguishable to human vision and conventional RGB imaging (Fig. 8a).

To expose the sub-surface fluid, we acquire 30 consecutive OCT volumes of the static scene and compute the speckle-variance volume. Voxels that contain moving scatterers (Intralipid) exhibit larger temporal variance and appear bright in the speckle-variance rendering (Fig. 8b). An orthogonal slice in the figure identifies the fluid-filled tube, highlighting OCT's unique ability to reveal sub-surface details. The combination of speckle variance OCT, our pose-estimation model and the closed-loop controller enables the operator to grasp the correct tube and transfer it to a tweezer in 23 s, despite the presence of multiple objects, specular noise, and clutter in the scene.

VII. FUTURE WORK

This work used an OCT scanner based on a dual-axis scanning galvanometer mirror, which is impractical for MIS neurosurgery due to its size. However, the scanner could be replaced with smaller forward scanning endoscopic OCT probes [36]. Furthermore, our current system achieves a slow rate of feedback. Improving this rests on faster OCT scan acquisition via improved hardware and faster model inference. Regarding the former, OCT systems with faster volumetric acquisition rates have been demonstrated [37]. Finally, conducting more extensive control experiments for a wider array of surgical tasks is important future work, including more realistic surgical sites (e.g., cadaver brains). Success of such trials could benefit from improvements to the tool's controller design. Designing and testing more advanced controllers for magnetic tools is thus also subject to future work.

ACKNOWLEDGMENT

This research was enabled in part by support provided by Compute Ontario and the Digital Research Alliance of

Canada. The OCT system was purchased with support of the Canada Foundation for Innovation (CFI, Advanced Fetal Diagnosis and Therapy Program). LAK received funding from NSERC (RGPIN-2020-05833). This work is supported by the Faculty of Applied Science and Engineering at the University of Toronto through the EMHSeed & XSeed program.

REFERENCES

[1] N. Simaan, R. M. Yasin, and L. Wang, "Medical technologies and challenges of robot-assisted minimally invasive intervention and diagnostics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 465–490, 2018.

[2] B. J. Nelson, S. Gervasoni, P. W. Y. Chiu, L. Zhang, and A. Zemmar, "Magnetically actuated medical robots: An in vivo perspective," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 1028–1037, 2022.

[3] Y. P. Lai, T. Lee, D. Sieben, L. Gauthier, J. Nam, and E. Diller, "Hybrid hydrogel-magnet actuated capsule for automatic gut microbiome sampling," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 10, pp. 2911–2922, 2024.

[4] W. Pryor, Y. Barnoy, S. Raval, X. Liu, L. Mair, D. Lerner, O. Erin, G. D. Hager, Y. Diaz-Mercado, and A. Krieger, "Localization and control of magnetic suture needles in cluttered surgical site with blood and tissue," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 524–531, 2021.

[5] G. Pittiglio, P. Lloyd, T. D. Veiga, O. Onaizah, C. Pompili, J. H. Chandler, and P. Valdastri, "Patient-specific magnetic catheters for atraumatic autonomous endoscopy," *Soft Robotics*, vol. 9, pp. 1120–1133, Dec 2022.

[6] A. Hong, A. J. Petruska, A. Zemmar, and B. J. Nelson, "Magnetic control of a flexible needle in neurosurgery," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 616–627, 2 2021.

[7] C. Forbrigger, E. Fredin, and E. Diller, "Evaluating the feasibility of magnetic tools for the minimum dynamic requirements of microneurosurgery," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4703–4709, 2023.

[8] 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)*, pp. 4783–4789, 2021.

[9] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi, "Mbapose: Mask and bounding-box aware pose estimation of surgical instruments with photorealistic domain randomization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9445–9452, 2021.

[10] S. Dehghani, M. Sommersperger, P. Zhang, A. Martin-Gomez, B. Busam, P. Gehlbach, N. Navab, M. A. Nasseri, and I. Iordachita, "Robotic navigation autonomy for subretinal injection via intelligent real-time virtual ioct volume slicing," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4724–4731, IEEE, May 2023.

[11] M. Zhou, X. Hao, A. Eslami, K. Huang, C. Cai, C. P. Lohmann, N. Navab, A. Knoll, and M. A. Nasseri, "6dof needle pose estimation for robot-assisted vitreoretinal surgery," *IEEE Access*, vol. 7, pp. 63113–63122, 2019.

[12] E. Fredin, N. Pol, A. Zaliznyi, E. Diller, and L. A. Kahrs, "Estimating the joint angles of an articulated microrobotic instrument using optical coherence tomography," in *2024 International Symposium on Medical Robotics (ISMR)*, pp. 1–7, 2024.

[13] N. Gessert, M. Schlüter, and A. Schlaefer, "A deep learning approach for pose estimation from volumetric oct data," *Medical Image Analysis*, vol. 46, pp. 162–179, May 2018.

[14] S. Frieler, S. Misra, and V. K. Venkiteswaran, "Selectively tunable joints with variable stiffness for a magnetically-steerable 6-dof manipulator," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[15] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, vol. 121, pp. 393–412, Proceedings of Machine Learning Research, 06–08 Jul 2020.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[17] C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomedical Optics Express*, vol. 8, pp. 3440–3448, July 2017.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[19] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, "Oct segmentation via deep learning: A review of recent work," in *Computer Vision – ACCV 2018 Workshops*, vol. 11367 of *Lecture Notes in Computer Science*, pp. 316–322, Springer, 2019.

[20] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, 2022.

[21] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnformer: Volumetric medical image segmentation via a 3d transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 4036–4045, 2023.

[22] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jäger, and K. H. Maier-Hein, "Mednext: Transformer-driven scaling of convnets for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, pp. 405–415, Springer Nature Switzerland, 2023.

[23] J. Y. Chang, G. Moon, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5079–5088, 2018.

[24] J. Cheng, Y. Wan, D. Zuo, C. Ma, J. Gu, P. Tan, H. Wang, X. Deng, and Y. Zhang, "Efficient virtual view selection for 3d hand pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 419–426, 2022.

[25] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, p. Art. no. 3337, Oct 2018.

[26] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9224–9232, 2018.

[27] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10529–10538, 2020.

[28] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1201–1209, May 2021.

[29] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3144–3153, 2021.

[30] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21674–21683, 2023.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[32] P. Ramachandran and G. Varoquaux, "Mayavi: 3d visualization of scientific data," *Computing in Science Engineering*, vol. 13, no. 2, pp. 40–51, 2011.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[34] A. Schonewille, C. He, C. Forbrigger, N. Wu, J. Drake, T. Looi, and E. Diller, "Electromagnets under the table: an unobtrusive magnetic navigation system for microsurgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, no. 3, pp. 980–991, 2024.

[35] A. Nankaku, M. Tokunaga, H. Yonezawa, T. Kanno, K. Kawashima, K. Hakamada, S. Hirano, E. Oki, M. Mori, and Y. Kinugasa, "Maximum acceptable communication delay for the realization of telesurgery," *PLOS ONE*, vol. 17, pp. 1–12, 10 2022.

[36] A. Gunalan and L. S. Mattos, "Towards oct-guided endoscopic laser surgery—a review," *Diagnostics*, vol. 13, no. 4, 2023.

[37] J. P. Kolb, W. Draxinger, J. Klee, T. Pfeiffer, M. Eibl, T. Klein, W. Wieser, and R. Huber, "Live video rate volumetric oct imaging of the retina with multi-mhz a-scan rates," *PLOS ONE*, vol. 14, pp. 1–20, 03 2019.