2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Realistic Data Generation for 6D Pose Estimation of Surgical Instruments

Juan Antonio Barragan[1], Jintan Zhang[1], Haoying Zhou[2],
Adnan Munawar[1], and Peter Kazanzides[1]

***Abstract*— Automation in surgical robotics has the potential to improve patient safety and surgical efficiency, but it is difficult to achieve due to the need for robust perception algorithms. In particular, 6D pose estimation of surgical instruments is critical to enable the automatic execution of surgical maneuvers based on visual feedback. In recent years, supervised deep learning algorithms have shown increasingly better performance at 6D pose estimation tasks; yet, their success depends on the availability of large amounts of annotated data. In household and industrial settings, synthetic data, generated with 3D computer graphics software, has been shown as an alternative to minimize annotation costs of 6D pose datasets. However, this strategy does not translate well to surgical domains as commercial graphics software have limited tools to generate images depicting realistic instrument-tissue interactions. To address these limitations, we propose an improved simulation environment for surgical robotics that enables the automatic generation of large and diverse datasets for 6D pose estimation of surgical instruments. Among the improvements, we developed an automated data generation pipeline and an improved surgical scene. To show the applicability of our system, we generated a dataset of 7.5k images with pose annotations of a surgical needle that was used to evaluate a state-of-the-art pose estimation network. The trained model obtained a mean translational error of 2.59 mm on a challenging dataset that presented varying levels of occlusion. These results highlight our pipeline's success in training and evaluating novel vision algorithms for surgical robotics applications.**

I. Introduction

In minimally invasive robotic surgery, automation of time-consuming and repetitive surgical subtasks has the potential to reduce the surgeon's mental demands and improve the overall efficiency of surgery [1]. Automation of surgical subtasks has been extensively studied by the research community, leading to autonomous algorithms for suturing [2]–[5], blood suction [6], [7], and tissue retraction [8], among others. One key challenge of surgical automation is developing perception algorithms to compensate for the robot's kinematic inaccuracies and execution failures. This requires estimating the 6D pose of rigid and articulated instruments from endoscopic video to modify autonomous motions based on visual feedback.

In the task of 6D pose estimation, the goal is to estimate the translation and rotation of the object of interest with respect to the camera coordinate frame. This task has traditionally been approached by extracting 2D visual features from RGB images and then matching them with corresponding 3D features on the object's model. These 2d-3d correspondences can then be used as an input to a Perspective-n-Point [9] solver to retrieve the object's pose.

More recently, end-to-end deep neural networks have demonstrated superior performance for 6D object pose estimation tasks than traditional point-pair feature approaches [10]. The main drawback of these deep learning approaches is the need to generate large amounts of annotated training data, which, for 6D pose estimation tasks, is prohibitively expensive to obtain. As a solution, it has been shown that high-fidelity synthetic data of models of physical objects can be used to train pose networks that perform well in locating their real counterparts [10], [11].

In the surgical context, synthetic data generation is a more challenging endeavor as it is important to generate samples that portray sensible instrument motions and realistic tissue-instrument interaction. Currently available simulation environments such as Vision Blender [12] or BlenderProc [13] can be utilized to render annotated images of surgical instruments in surgical backgrounds; however, they offer limited capabilities on how the objects can be moved or interact with each other in the scene. Furthermore, they do not offer good support to work with articulated robotic instruments.

In previous work, an open-source platform for surgical suturing was introduced to address some of the limitations of surgical robotics simulation platforms [14]. In particular, this work, built with the Asynchronous Multi-Body Framework (AMBF) [15], introduced improved robot control algorithms and teleoperation capabilities, and provided access to ground-truth imaging data. Although it enabled the collection of realistic suturing motions via teleoperation, it still lacked the capabilities to automatically generate the large-scale datasets needed for neural network training. Furthermore, the scene provided a simplified phantom, not resembling any real physical phantom, which complicated the task of physically reproducing the virtual environment.

To address these limitations, we have developed an automated data generation pipeline on top of [14] to produce large-scale and diverse datasets from pre-recorded trajectories generated with teleoperation. Moreover, we scanned and added a commercially available training suturing pad to our simulation environment to improve realism and to ensure that the virtual scene could be physically reproduced. The goal of these improvements was to facilitate the creation of the large-scale datasets needed for deep learning-based algorithms.

In this paper, we showcase an application of our pipeline by generating data to train a state-of-the-art 6D pose estima-

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. Email: jbarrag3@jhu.edu, jzhan247@jhu.edu
[2]Department of Robotics Engineering, Worcester Polytechnic Institute, Worcester, MA 01608, USA. Email: hzhou6@wpi.edu

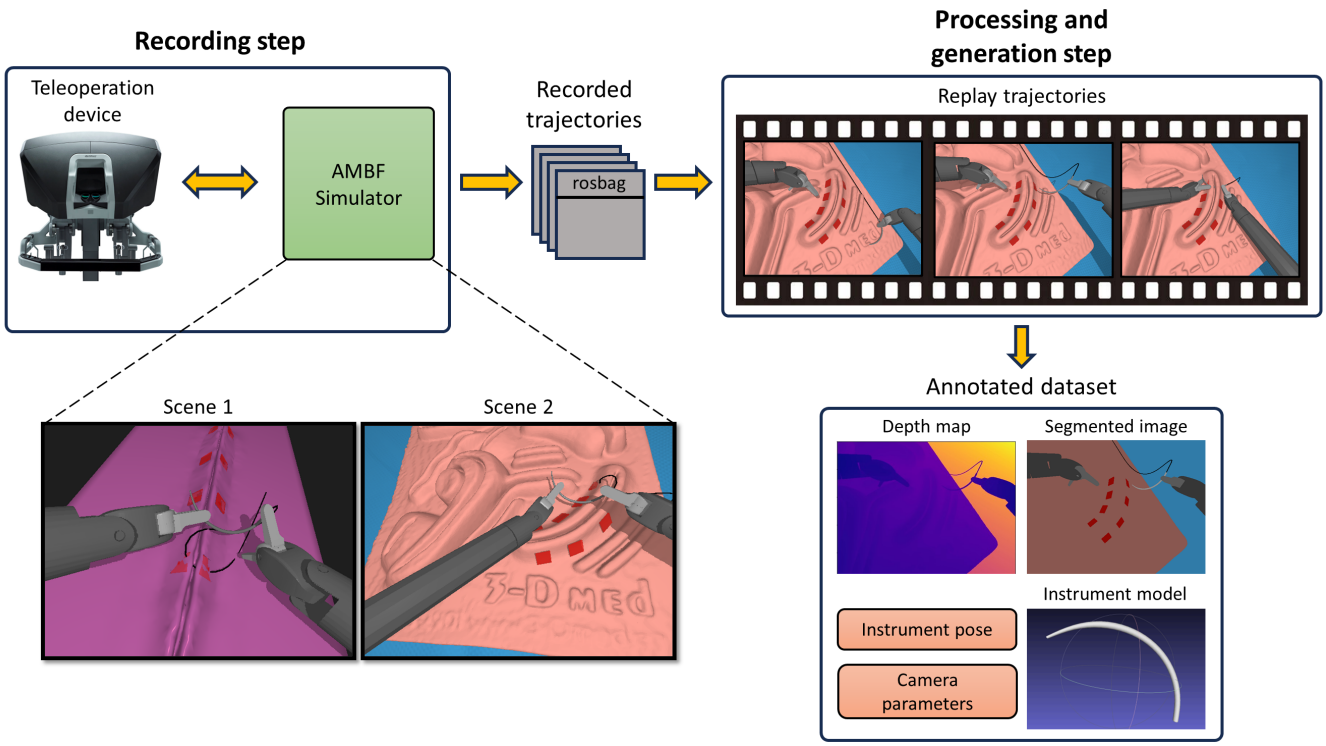

Fig. 1: Proposed data collection pipeline. Realistic data generation with our proposed system requires two steps. First, trajectories of the robotic manipulators are collected using a teleoperation device. Second, trajectories are replayed automatically multiple times from different camera viewpoints to generate diverse set of images. While replaying, our pipeline stores depth and segmentation maps and the ground truth pose of all the objects with respect to the camera.

tion network to predict the pose of a needle while performing suturing maneuvers. Regarding the network architecture, the GDR-Network [16] was chosen as it is one of the first fully differentiable networks for the task of pose estimation from monocular RGB. We envision that our work will significantly benefit the community of surgical robotics researchers as we provide a standardized platform for generating, evaluating, and deploying novel vision algorithms. Lastly, we highlight that thanks to the modular nature of the base simulation engine, our data generation pipeline can be used for a wide range of tasks and objects in surgical robotics setups.

In summary, this work presents the following contributions:

1) An automated data generation pipeline for 6D pose estimation of surgical instruments.
2) A realistic simulation environment for surgical suturing based on a commercially available suturing pad model.
3) A dataset of 7.5k images with 6D pose annotations for a simulated surgical needle.
4) Evaluations on a state-of-the-art 6D pose estimation neural network on the task of surgical needle pose estimation.

II. Related Work

A. Offline Simulation Data Generation

Vision Blender [12] is a Blender add-on for generating synthetic computer vision data such as RGB, depth, segmentation, optical flow, and surface normals. Designed as a tool to efficiently generate data for surgical robotic system development, Vision Blender supports converting generated data to Robot Operating System (ROS) [17] messages. Another modular procedural pipeline for generating simulation data based on Blender is BlenderProc [13]. It shares data generation capabilities with Vision Blender, with the added feature of bounding box generation. BlenderProc also supports importing data in URDF format, expanding its capability for modeling complex robotics systems. Compared to using the native Blender Python, which demands a deep understanding of the Blender infrastructure, BlenderProc offers an intuitive Python interface to simplify the scene-building and data-acquisition process.

B. Robot Simulator & Digital Twins

NVIDIA® Isaac Sim is a scalable robotics simulator and synthetic data generator. Powered by the GPU-accelerated physics simulation engine PhysX and physically-based rendering technology Iray, Isaac Sim is capable of simulating physically accurate virtual environments and generating photorealistic data. Isaac Sim offers support for Universal Scene Descriptor (USD) and Unified Robot Description Format (URDF), enabling developers to seamlessly import intricate 3D environment definitions and robot configurations with ease. In addition, Isaac Sim allows developers to establish connections between Isaac Sim and their custom robot ap-

plications via integrated Robot Operating System (ROS1 & ROS2) interfaces. With extensions such as Isaac Gym [18] and Isaac Orbit [19], developers can efficiently test and refine their robotic systems and robot learning algorithms.

Defining multi-body robots using formats like URDF or Standard Description Format (SDF) can lead to ambiguous definitions in cases of densely connected, sparsely connected, or unconnected bodies. To address this constraint, Munawar et al. [15] introduced the Asynchronous Multi-Body Framework (AMBF), an innovative front-end description format for multi-body simulation, aimed at simulating complex closed-loop robots. AMBF leverages Bullet Physics [20] for its physics simulation and CHAI-3D [21] for graphics rendering and haptic volume rendering. Within AMBF, every object features a custom OpenGL shader, facilitating diverse data generation capabilities, including RGB, depth, and segmentation maps.

Building upon AMBF, many research efforts have emerged to advance robot-assisted surgeries. These include using AMBF to design image-guided feedback modalities [22], a novel framework for skull base surgeries featuring high-precision optical tracking and real-time simulation [23], and a causality-driven robot tool segmentation algorithm [24]. In this work, we selected AMBF simulation over other simulation alternatives due to its support of a broad array of input devices, which facilitates the collection of realistic motions of the surgical instruments. In particular, its tight integration with the da Vinci Research Kit (dVRK) [25] allows collecting robotic surgical motions with a similar setup to what is used in surgery.

III. Methodology

The primary motivation of this work was to provide a data generation tool for 6D pose estimation of surgical instruments. In this regard, we adapted an open-source simulation environment to automatically generate sequences of images of robotic-assisted surgical actions with their corresponding ground-truth maps. The proposed data generation pipeline (See figure 1) was developed with the goal of generating programmatically large and diverse datasets. The methodology section is divided as follows. In section III-A, we present the pipeline for automatic data generation. Section III-B shows the improvements in the surgical virtual scene. Section III-C describes the generation of a dataset for the task of needle pose estimation. Section III-D describes the pose estimation deep learning model trained to estimated the needle's pose. Lastly, section III-E, describes the evaluation metrics used for the predictions of the trained neural network.

A. Data generation pipeline

Our proposed data generation pipeline is composed of two stages: a recording step, and a processing and generation step. During the data recording step, a teleoperation device is used to move the virtual robotic manipulators to perform the surgical task. While teleoperating, joint and Cartesian positions of the robotic manipulators, and the poses of other objects in the simulation are stored in a rosbag file[1].

During the processing and generation step, the stored robotic trajectories are replayed multiple times under different camera viewpoints and lighting conditions. While replaying the trajectories, a collection script stores the resulting monocular or stereoscopic RGB images with their corresponding ground-truth information, i.e., depth map, segmented images, camera intrinsic parameters, and pose of objects expressed with respect to the camera coordinate frame.

1) Format for generated data: To store the data, it was decided to use the Benchmark for 6D Object Pose Estimation (BOP) format [10]. This is a standardized format adopted by several benchmark 6D pose estimation datasets such as HOPE [26], YCB [27], and others. Moreover, it is a standard format used for an annual 6D pose competition [10]. In the BOP format, related data are grouped under a *scene_id*. For our pipeline, data from each trajectory replay was stored in a different *scene_id*.

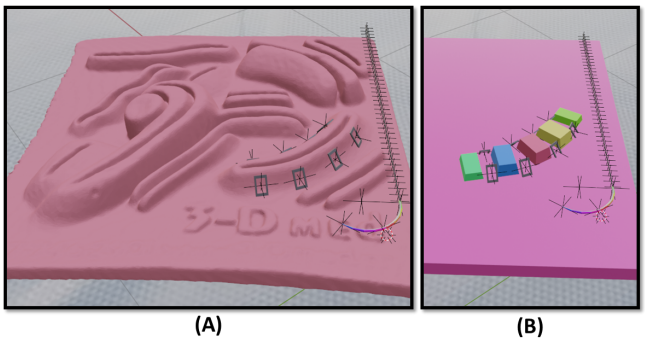

Fig. 2: (a) Visual mesh of the 3-Dmed phantom after preprocessing. (b) Simplified collision mesh composed of multiple convex subcomponents assembled into a single mesh. The collision mesh was only provided for a single ridge of the phantom.

B. Improvements of virtual scene

To improve the realism of our virtual scenes, a commercially available suturing pad (3-Dmed, Franklin, OH, US) was added to the simulation. The suturing pad was initially MRI scanned to obtain a mesh that was preprocessed using 3D Slicer [28] and Meshlab [29]. The MRI scanning was selected over other modalities as it provides higher contrast for soft tissue phantoms [30]. Using the resulting mesh, an AMBF Description File (ADF) is made by utilizing the Blender-AMBF addon plugin [14]. As observed in figure 2, the full-resolution suturing pad is used for visualization, while a simplified mesh made of convex subshapes is used for collision. Small corridors are left on the collision mesh to allow for needle insertions similar to the scene developed

[1]A rosbag is a file format used to store messages, from the Robot Operating System (ROS) middleware. It is ideal for storing trajectories from a robot.

in [14]. Collision meshes are simplified to optimize the simulation's performance.

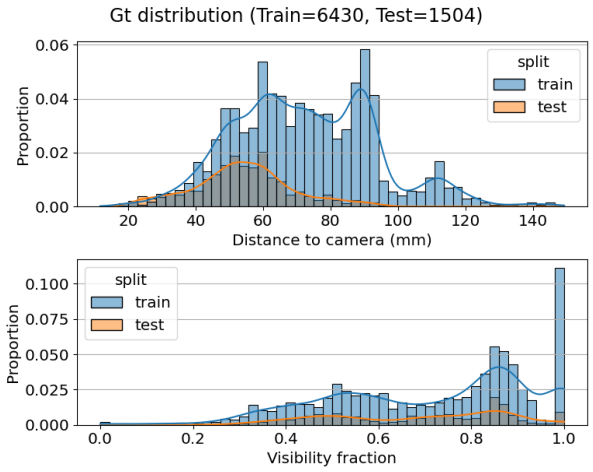

Fig. 3: Ground-truth distribution of the collected needle 6DoF detection dataset.

C. Generated dataset for surgical needles pose detection

Using our improved simulation environment, we collected a dataset for the task of 6D pose estimation of an 18.65 mm surgical needle. First, we collected 6 rosbag recordings of suturing motions using a dVRK robot's surgical console [25]. Two recordings were done in scene 1 and four in scene 2 (See figure 1).

Each of the 6 collected recordings was then replayed on the simulator 20 times, each time from different camera positions and view angles. Data from 4 recordings were used as a training set and 2 for the testing set. Camera positions were specified in the joint space of a virtual endoscopic camera manipulator (ECM) provided by the base simulation environment. To produce unique viewpoints with every replayed recording, a small random offset was added to the selected ECM joints.

After filtering images where the needle was not present, 6430 training and 1500 testing images with a 640x480 resolution were obtained. As observed in figure 3, the resulting dataset is more challenging and realistic than the one presented in [4] as the needles in the images present varying levels of occlusion and distance to the camera. The visibility fraction is calculated with

$$\text{visibility} = \frac{\text{area of visible mask}}{\text{area of projected mask}} \tag{1}$$

where the visible mask is the set of pixels in the RGB image that correspond to the needle and the projected mask is the set of pixels obtained by projecting the needle's CAD model to the image plane using the ground-truth pose.

D. Selected deep learning model for 6D pose estimation

Using the dataset described in section III-C, we trained the state-of-the-art network for 6D pose estimation GDR-Net [16]. This network was selected as it is one of the first fully differentiable pose estimation methods in the literature and the winner of the BOP pose estimation competition of 2022 [10]. This network receives as an input a 2D RGB region, where the object of interest is located, and outputs three intermediate geometric feature maps: the *visible object mask*, a map of 2d-3d dense correspondences, and a surface region attention map. These intermediate maps are concatenated and then given as input to a fully-differentiable Patch-PnP module that regresses the final rotation and translation of the object.

Training of this network can be performed in an end-to-end manner and only requires the RGB image and the object CAD model to generate ground truth for the intermediate geometric maps. As mentioned above, the network requires a region of interest (ROI) where the object is located. To obtain these ROIs during our experiments, the off-the-shelf detector YOLOX [31] was also trained with our dataset for the task of 2D bounding box detection for the needle.

E. Evaluation metrics for the pose estimates

Pose estimations from the neural network were evaluated using three common error metrics: (1) translation error (e_{TE}), (2) rotation error (e_{RE}) [32], and (3) Maximum Symmetry-Aware Surface Distance (e_{MSSD}) [33]. Metric 1 measures the translational error using the Euclidean distance. Metric 2 measures the rotational error using the axis angle representation of rotation matrices. Lastly, metric 3 measures the maximum distance between a vertex of the object model transformed with the ground truth and estimated pose. Given a ground truth pose $\bar{P} = (\bar{R}, \bar{t})$, an estimated pose $\hat{P} = (\hat{R}, \hat{t})$, and a set of vertices V_M belonging to the object model, the metrics e_{TE}, e_{RE} and e_{MSSD} can be calculated with

$$e_{TE} = ||\bar{t} - \hat{t}|| \tag{2}$$

$$e_{RE} = arccos((Tr(\bar{R}\hat{R}^{-1} - 1)/2) \tag{3}$$

$$e_{MSSD} = \min_{\mathbf{S} \in S_M} \max_{\mathbf{x} \in V_M} \|\hat{\mathbf{P}}\mathbf{x} - \bar{\mathbf{P}}\mathbf{S}\mathbf{x}\|_2 \tag{4}$$

where S_M is a set of symmetry transformations for the object whose pose is being estimated.

IV. Experiments and Results

For the evaluation experiments, first, the YOLOX and GDR-Net networks were trained with the generated training dataset. YOLOX was trained for 30 epochs using the Ranger Optimizer [34], a batch size of 16 and a learning rate of 1e-3. GDR-Net was trained with the Ranger Optimizer for 450 epochs, a batch size of 48 images and a learning rate of 8e-4. This training setup was similar to the one used in [16]. At test time, the trained bounding box detector was used to predict a single region of interest for each image. This region of interest was then used as input for the GDR-Net. Only images where at least 30 percent of the needle was visible were used for evaluation. Some sample images from

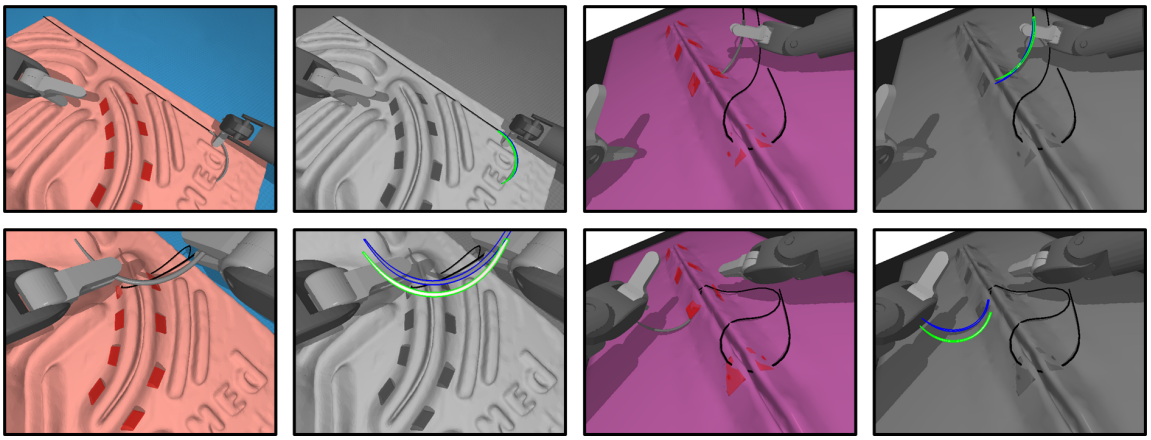

Fig. 4: Test set sample frames and corresponding pose prediction visualizations. Colored images show samples from the test dataset. Masks in the grayscale images are generated by projecting the needle model to the image with the ground-truth (blue mask) and the network's estimated pose (green mask). Higher overlaps between the green and blue masks are indicative of better pose estimates.

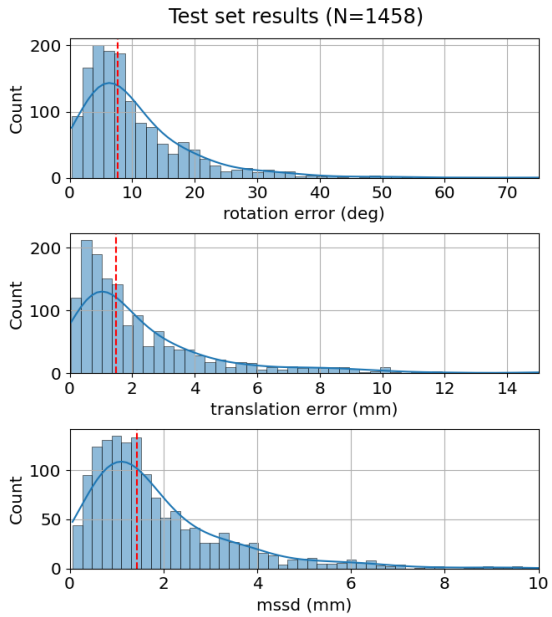

Fig. 5: Error distribution for the best GDRNet model on the test dataset. The x-axis represents the different error metrics, and the y-axis is the number of samples within each bin. The red dotted line indicates the median performance. The x-axes of the histograms were truncated respectively at 70 mm, 15 deg and 10 mm for visualization purposes.

the test set with their corresponding pose predictions can be observed in figure 4.

Final pose errors can be seen in table I. On the evaluated images, GDR-Net obtained a median rotational error of 7.74 degrees and a median translation error of 1.49 mm (less than 20 percent of the needle's diameter). These results are comparable to the pose detection results of non-occluded needles presented in [4] even though needles in our test set present varying levels of occlusion. Lastly, the median MSSD error is 1.43 mm. Pose error distribution in figure 5 indicates that the network can have sporadic predictions with significantly higher errors. High pose errors can be mainly attributed to images where several needle poses cannot be distinguished from each other.

Test set results (N=1458)			
	e_{RE} (deg)	e_{TE} (mm)	e_{MSSD} (mm)
mean	11.85	2.59	2.09
std	17.52	3.41	2.43
median	7.74	1.49	1.43
min	0.33	0.04	0.06
max	170.5	33.01	20.91

TABLE I: GDRNET test set results. Only images where at least 30 percent of the needle was visible were included in the evaluation. The diameter of the detected needle was 18.65 mm.

V. Discussion and Future Work

In this work, we developed a data generation pipeline for 6D estimation tasks of surgical instruments on top of the simulation framework AMBF. The proposed pipeline generates monocular or stereoscopic RGB images, and pose annotations for any rigid or articulated instrument in the scene. Moreover, each generated RGB image is accompanied by its corresponding depth and segmentation maps.

The focus of the work was to enable the automatic generation of large and diverse datasets showing realistic tissue-instrument interaction and sensible trajectories for robotic manipulators. In this regard, we divide our data generation pipeline into two steps: (1) a data recording step where robotic trajectories of a surgical task are recorded, and (2) a processing and generation step where each collected trajectory is replayed multiple times from different camera view angles and lighting conditions.

To showcase the applicability of our pipeline, we generated a dataset of 7.5k images with pose annotations for a surgical needle to evaluate a state-of-the-art pose estimation

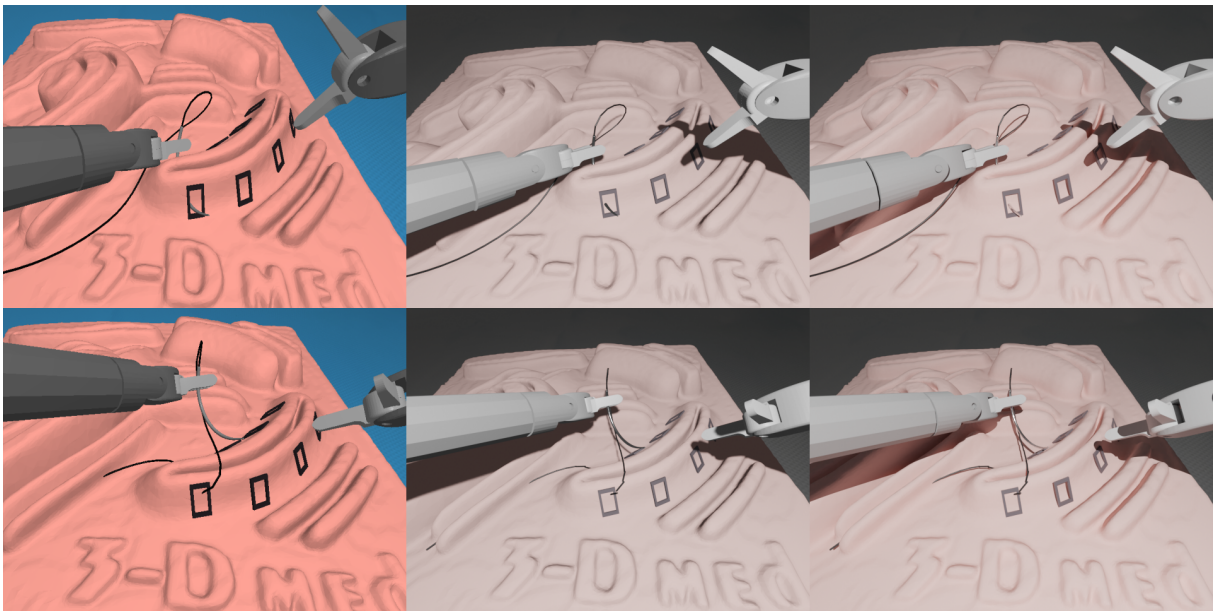

Fig. 6: Rendering quality comparison between AMBF (left), Eevee (center), and Cycles (right). Each row represents the same scene. Shadow quality and metal shininess are superior in Cycles due to more comprehensive and exhaustive ray tracing, while the needle is less glossy and the shadow is unrealistically uniform in Eevee. Nevertheless, both Eevee and Cycles produce significantly higher fidelity rendering than AMBF.

neural network. After training, the network had translation and rotation errors comparable to previous works [4], [5] while being tested on a challenging dataset where the needle could be partially occluded by the instruments and the tissue.

Although the network showed good performance on average, it is important to remember that the model makes predictions solely based on the visual appearance of the object, and therefore cases where multiple poses are indistinguishable from each other will result in high pose errors. Specifically for surgical needles, there are two main scenarios leading to pose ambiguities: (1) images where both the needle's tail and tip are occluded and (2) images where the needle's curvature cannot be observed, i.e., the needle appears as a straight line. As a solution, pose ambiguities could be resolved by using the network's predictions with a model-based tracker that uses additional priors, such as the robot's kinematic motion or the pose of the needle in previous frames.

In future work, we will leverage our data generation pipeline to study different techniques for transferring pose detection models from simulation to reality, a problem that is often referred to in the literature as the "domain gap" [35]. As noted by [10] and [36], rendering realism plays an important role in transferring neural networks from synthetic to real objects. This hints that models trained based on our current synthetic data (generated with the simpler Blinn-Phong shading technique [37]) might suffer from degraded performance when applied to data from the physical surgical platform.

To mitigate this limitation, we implemented a preliminary real-time pipeline to improve the rendering quality by transferring the object pose (including cameras and lights) from the AMBF simulator to Blender. This allows us to utilize the two state-of-the-art rendering engines included in Blender since version 3.0: Eevee (a real-time rasterization-based renderer) and Cycles (a physically based path tracer). As shown in figure 6, shadows and metal shininess rendered using Blender are significantly better than AMBF. Future studies will focus on understanding the effects of different rendering algorithms on the simulation-to-real transfer of neural networks. Additional future improvements on our simulation platform will include more accurate models for the robotic instruments and advanced materials that more accurately reflect surgical tools.

Acknowledgments

This work was supported in part by NSF AccelNet awards OISE-1927354 and OISE-1927275. We thank Irene Kim, Haochen Wei and Nicholas Greene for their invaluable insights on 6D pose estimation models.

Supplementary Information

For more information, visit the project repository at https://github.com/surgical-robotics-ai/realistic-6dof-data-generation

References

[1] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in Surgical Robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, May 2021.

[2] K. L. Schwaner, D. Dall'Alba, P. T. Jensen, P. Fiorini, and T. R. Savarimuthu, "Autonomous Needle Manipulation for Robotic Surgical Suturing Based on Skills Learned from Demonstration," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, Aug. 2021, pp. 235–241.

[3] 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)*. Philadelphia, PA, USA: IEEE, May 2022, pp. 9637–9643.

[4] Y. Jiang, H. Zhou, and G. S. Fischer, "Markerless Suture Needle Tracking From A Robotic Endoscope Based On Deep Learning," in *2023 International Symposium on Medical Robotics (ISMR)*, Apr. 2023, pp. 1–7.

[5] 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)*, Oct. 2022, pp. 5286–5292.

[6] 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, Apr. 2021.

[7] J. A. Barragan, D. Chanci, D. Yu, and J. P. Wachs, "SACHETS: Semi-Autonomous Cognitive Hybrid Emergency Teleoperated Suction," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. Vancouver, BC, Canada: IEEE, Aug. 2021, pp. 1243–1248.

[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)*, May 2021, pp. 4783–4789.

[9] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, p. 381–395, jun 1981. [Online]. Available: https://doi.org/10.1145/358669.358692

[10] M. Sundermeyer, T. Hodaň, Y. Labbé, G. Wang, E. Brachmann, B. Drost, C. Rother, and J. Matas, "BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2023, pp. 2785–2794.

[11] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 23–30.

[12] J. Cartucho, S. Tukra, Y. Li, D. S. Elson, and S. Giannarou, "VisionBlender: A tool to efficiently generate computer vision datasets for robotic surgery," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 0, no. 0, pp. 1–8, Dec. 2020.

[13] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi, "BlenderProc: Reducing the Reality Gap with Photorealistic Rendering."

[14] A. Munawar, J. Y. Wu, G. S. Fischer, R. H. Taylor, and P. Kazanzides, "Open Simulation Environment for Learning and Practice of Robot-Assisted Surgical Suturing," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3843–3850, Apr. 2022.

[15] A. Munawar, Y. Wang, R. Gondokaryono, and G. S. Fischer, "A real-time dynamic simulator and an associated front-end representation format for simulating complex robots and environments," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. [Online]. Available: https://par.nsf.gov/biblio/10207704

[16] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 16 606–16 616.

[17] M. Quigley, "ROS: an open-source robot operating system," in *IEEE International Conference on Robotics and Automation*, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:6324125

[18] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[19] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar *et al.*, "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robotics and Automation Letters*, 2023.

[20] E. Coumans, "Bullet physics simulation," in *ACM SIGGRAPH 2015 Courses*, ser. SIGGRAPH '15. New York, NY, USA: Association for Computing Machinery, 2015. [Online]. Available: https://doi.org/10.1145/2776880.2792704

[21] S. Musić and S. Hirche, "Haptic shared control for human-robot collaboration: A game-theoretical approach," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10 216–10 222, 2020, 21st IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240589632033514X

[22] H. Ishida, J. A. Barragan, A. Munawar, Z. Li, A. Ding, P. Kazanzides, D. Trakimas, F. X. Creighton, and R. H. Taylor, "Improving surgical situational awareness with signed distance field: A pilot study in virtual reality," 2023.

[23] H. Shu, R. Liang, Z. Li, A. Goodridge, X. Zhang, H. Ding, N. Nagururu, M. Sahu, F. X. Creighton, R. H. Taylor, and et al., "Twin-s: A digital twin for skull base surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 6, p. 1077–1084, 2023.

[24] H. Ding, J. Zhang, P. Kazanzides, J. Y. Wu, and M. Unberath, "CaRTS: Causality-driven robot tool segmentation from vision and kinematics data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 387–398. [Online]. Available: https://doi.org/10.1007/978-3-031-16449-1_37

[25] 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)*, May 2014, pp. 6434–6439.

[26] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-DoF Pose Estimation of Household Objects for Robotic Manipulation: An Accessible Dataset and Benchmark," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 13 081–13 088.

[27] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," May 2018.

[28] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, and et al., "3D Slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, p. 1323–1341, 2012.

[29] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: an Open-Source Mesh Processing Tool," in *Eurographics Italian Chapter Conference*, V. Scarano, R. D. Chiara, and U. Erra, Eds. The Eurographics Association, 2008.

[30] M. P. Hiorns, "Imaging of the urinary tract: the role of CT and MRI," *Pediatric nephrology*, vol. 26, no. 1, pp. 59–68, 2011.

[31] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," *ArXiv*, vol. abs/2107.08430, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236088010

[32] T. Hodan, J. Matas, and S. Obdrzálek, "On evaluation of 6d object pose estimation," in *ECCV Workshops*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:14980684

[33] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger, "Introducing MVTEC Itodd — a dataset for 3D object recognition in industry," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[34] L. Wright, "Ranger - a synergistic optimizer." https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019.

[35] D. Schraml, "Physically based synthetic image generation for machine learning: A review of pertinent literature," in *Photonics and Education in Measurement Science 2019*, vol. 11144. SPIE, Sep. 2019, pp. 108–120.

[36] S. A. Heredia Perez, M. Marques Marinho, K. Harada, and M. Mitsuishi, "The effects of different levels of realism on the training of CNNs with only synthetic images for the semantic segmentation of robotic instruments in a head phantom," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 8, pp. 1257–1265, Aug. 2020.

[37] J. F. Blinn, "Models of light reflection for computer synthesized pictures," *SIGGRAPH Comput. Graph.*, vol. 11, no. 2, p. 192–198, jul 1977. [Online]. Available: https://doi.org/10.1145/965141.563893