

# Evaluation of deep learning-based surface reconstruction and tracking of surgical scenes for minimally invasive intervention

Nati Nawawithan<sup>a,b</sup>, James Yu<sup>a,c</sup>, Kelden Pruitt<sup>a,b</sup>, Uy Seng<sup>a</sup>, Baowei Fei<sup>a,b,c\*</sup>

<sup>a</sup> Center for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX

<sup>b</sup> Department of Bioengineering, University of Texas at Dallas, Richardson, TX

<sup>c</sup> Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

\* Corresponding author: [bfei@utdallas.edu](mailto:bfei@utdallas.edu), Website: <https://fei-lab.org>

## ABSTRACT

Minimally invasive procedures enable clinicians to perform interventions with reduced tissue damage and expedited recovery times. However, laparoscopic imaging inherently provides a limited field of view, compelling surgeons to mentally integrate preoperative imaging with intraoperative scenes, which is prone to localization errors. Augmented reality (AR) offers a solution by overlaying preoperative 3D organ models onto live laparoscopic video; however, precise 3D reconstruction and localization of the surgical scene is critical for accurate overlays. In this study, we evaluated three deep learning-based surface reconstruction methods, DROID-SLAM, MAST3R-SLAM, and the visual geometry grounded transformer (VGGT), for reconstructing tissue surfaces from laparoscopic video. We evaluated these algorithms with a dataset comprised of laparoscopic RGB frames of *ex-vivo* porcine tissue, ground-truth camera trajectories obtained via an optical tracking system and known camera intrinsic parameters. Dense point cloud and depth map predications were then qualitatively compared between models and estimated camera trajectory was compared quantitatively using absolute trajectory error (ATE) and relative pose error (RPE). Preliminary results indicate that all models produced plausible tissue surface reconstructions, with estimated depth maps and camera trajectories demonstrating close alignment with the ground truth. These findings suggest that deep learning-based 3D reconstruction methods can effectively support image-guided interventions without relying on ground-truth camera poses or depth maps and may be integrated with AR to support minimally invasive surgical workflows.

**Keywords:** Surface reconstruction, minimally invasive intervention, laparoscopic imaging, deep learning

## 1. INTRODUCTION

Minimally invasive techniques have transformed surgical practice by reducing postoperative recovery time, scarring, and intraoperative complications.<sup>1</sup> These procedures are typically guided by preoperative imaging modalities such as computed tomography (CT) or magnetic resonance imaging (MRI), or by intraoperative ultrasound, but they still depend heavily on the operator's ability to mentally correlate imaging data with patient anatomy. Real-time 3D surface reconstruction offers a solution by providing augmented reality (AR) overlays of subsurface targets or anatomical models derived from preoperative 3D images, thereby improving localization accuracy and reducing error rates.<sup>2</sup> Our previous work has demonstrated the feasibility of AR frameworks in minimally invasive surgery<sup>3</sup> and biopsy<sup>4</sup> procedures.

Recent advances in dense visual simultaneous localization and mapping (SLAM) have increasingly shifted from purely geometric pipelines toward learned end-to-end frameworks that integrate deep feature representations with global optimization. For example, our group previously implemented dense surface reconstruction from monocular laparoscopic video using a learning-based visual SLAM.<sup>5</sup> DROID-SLAM<sup>6</sup>, MAST3R-SLAM<sup>7</sup>, and the visual geometry grounded transformer (VGGT)<sup>8</sup> exemplify three distinct but convergent paradigms for jointly estimating camera motion and scene structure from monocular image sequences. DROID-SLAM combines recurrent neural networks with differentiable bundle adjustment, predicting dense pixel correspondences that are iteratively refined through a ConvGRU-based update operator and optimized using Gauss Newton-based dense bundle adjustment over a frame graph. Its tightly coupled frontend-backend architecture enables multi-view consistent depth estimation and global pose refinement, maintaining robustness in low-texture and dynamic environments through confidence-weighted geometric optimization.

In contrast, MAST3R-SLAM incorporates learned 3D reconstruction priors in the form of dense point maps, using directional ray-based matching and iterative projective alignment to track frames and fuse geometry into canonical keyframe maps. Global consistency is enforced via pose graph optimization that minimizes ray-based errors, offering

improved resilience to depth ambiguity and noisy predictions. VGGT adopts a more fully end-to-end strategy, leveraging a transformer-based backbone pretrained with large-scale geometric prior to directly regress camera intrinsics, extrinsics, depth, and 3D structure from arbitrary image sets. Broadly, these methods illustrate a spectrum from tightly integrated learned-geometric optimization (DROID-SLAM) to prior-guided geometric fusion (MASt3R-SLAM), then to foundation-model-style direct 3D regression (VGGT), reflecting the evolving balance between explicit geometric reasoning and learned representation in modern SLAM systems.

In this study, we investigate deep learning-based 3D reconstruction methods for surgical scenes, including SLAM-based and transformer-based frameworks, with the objective of identifying robust approaches suitable for image-guided interventions.

## 2. MATERIALS AND METHODS

### 2.1 Laparoscopic frame acquisition procedure

The experimental setup is similar to our previous work,<sup>9</sup> which included a laparoscope (WAIR100A, Olympus), and an optical tracking system. Porcine abdominal tissue was used for dataset acquisition. The laparoscope was equipped with an RGB camera (E3ISPM, ToupTek) and coupled to a halogen light source (OSL2 High-Intensity Fiber-Coupled Illuminator, Thorlabs, Newton, NJ) via optical fiber to provide consistent illumination of the tissue surface. Laparoscopic poses were captured at 120 frames per second (fps) using retroreflective markers in conjunction with an infrared tracking system. The laparoscope intrinsic parameters were calculated with a checkerboard pattern using Open3D library. During acquisition, the operator manually maneuvered the laparoscope around the tissue specimen to ensure comprehensive coverage of the scene. Continuous laparoscopic video was recorded for approximately 2-4 minutes per trial.

### 2.2 Surface reconstruction framework

The overview of the tissue surface evaluation framework can be seen in Figure 1. Each frame of the recorded laparoscopic feeds was extracted and synchronized with the positional data from optical tracking system. All the frames were sampled down with 2 fps to decrease data redundancy and free up graphics processing unit (GPU) memory during 3D reconstruction process. 3D reconstruction models, including DROID-SLAM, MASt3R-SLAM, and VGGT were selected for this study. Each algorithm was executed in a Docker container on the high-performance computer equipped with six NVIDIA RTX A6000 GPUs.

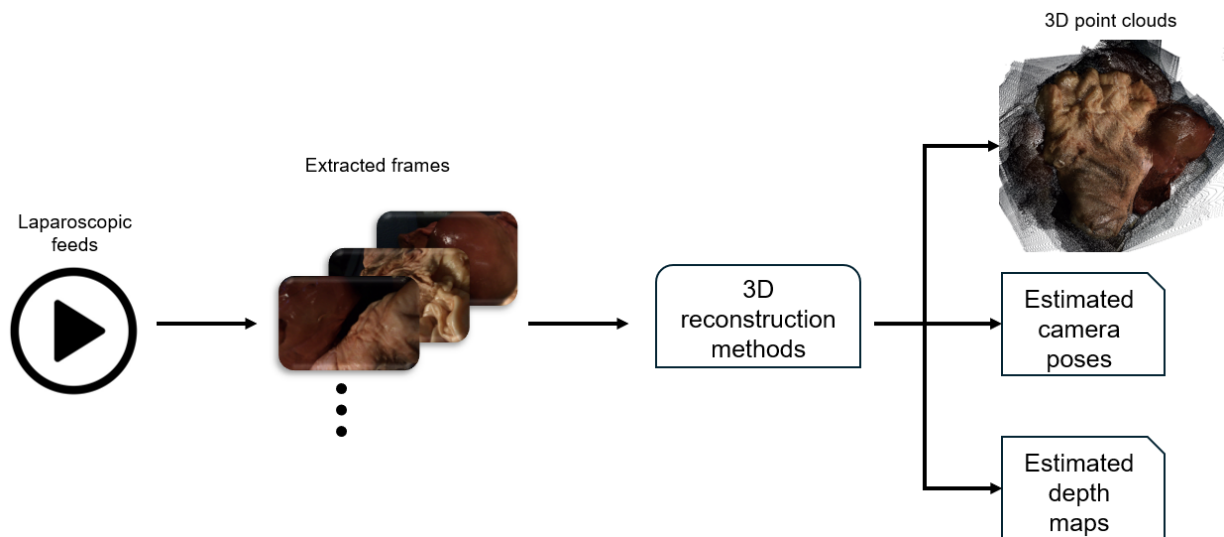


Figure 1. Diagram of the 3D reconstruction framework with our dataset.

The sampled frames and camera intrinsic parameters were given to DROID-SLAM. Nevertheless, only the sample frames were provided to MASt3R-SLAM and VGGT. The outputs of those models are 3D point clouds, estimated camera poses,

and estimated depth maps. For evaluation, 3D point clouds and depth maps were evaluated qualitatively. The estimated camera poses, and the ground truth ones were compared both qualitatively and quantitatively.

### 2.3 Evaluation metrics

To evaluate the camera pose estimation performance of the learning-based models, we employ the absolute trajectory error (ATE) and relative pose error (RPE) metrics.<sup>5,10</sup> The ATE assesses the global consistency of a trajectory by comparing the absolute positions of the estimated poses against the ground truth following alignment. ATE serves as a standard metric for SLAM systems as it captures the cumulative error over the entire trajectory, thereby quantifying the global consistency of the resulting map.

RPE quantifies local accuracy and drift by evaluating motion consistency over trajectory sub-segments. Unlike metrics designed for globally consistent SLAM systems, RPE is particularly effective for assessing odometry systems, which are inherently susceptible to drift. To compute this metric, the trajectory is partitioned into sub-sequences of fixed lengths, *i.e.*, 10 meters. For each segment, the estimated relative transformation is compared against the corresponding ground truth transformation, and the resulting errors are averaged across all possible sub-sequences of the given length. All error calculations in this study were performed using the *evo* package.<sup>11</sup>

## 3. RESULTS

### 3.1 Qualitative results

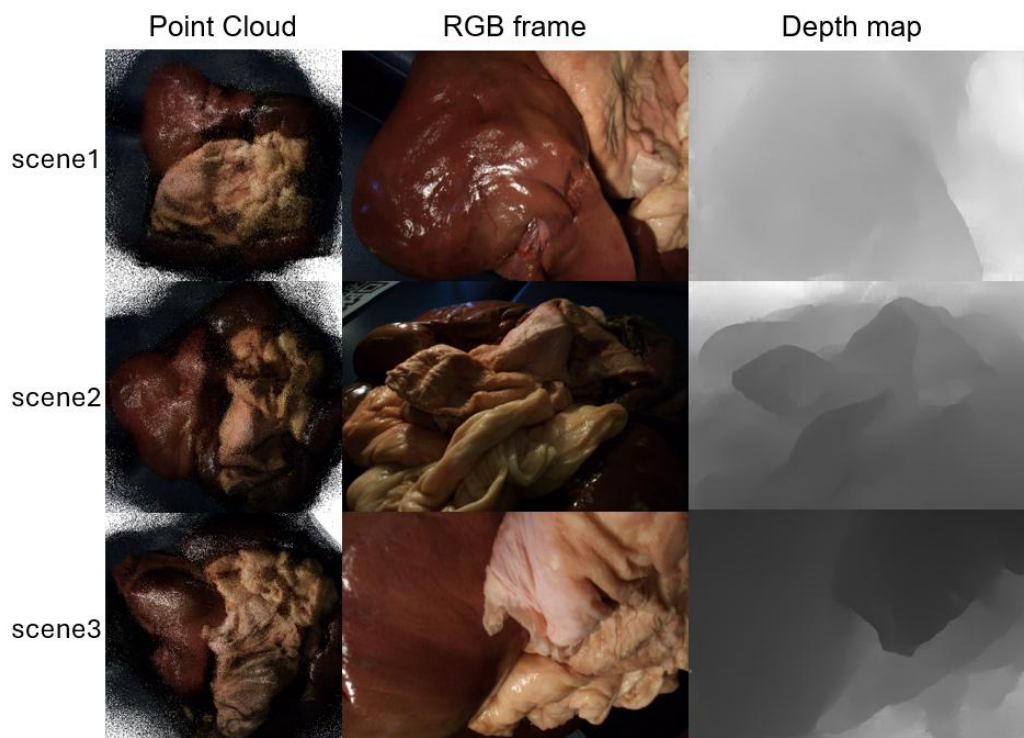


Figure 2. 3D Point clouds of tissue surface, and depth maps with their corresponding RGB frames from DROID-SLAM.

Figures 2-4 illustrate the 3D point clouds and depth maps generated by each model. The point clouds were visualized using the Open3D library, while the depth maps are presented as grayscale images, where lighter regions correspond to greater distances from the camera. To evaluate the robustness of the frameworks, three distinct animal tissue scenes were utilized. In terms of input requirements, MAST3R-SLAM and VGGT performed surface reconstruction using solely RGB frames, whereas DROID-SLAM required both RGB frames and camera intrinsic parameters. Qualitatively, all algorithms provided accurate shape and color representations across the scenes; however, VGGT yielded lower-quality surface reconstructions in scenes 1 and 2. While the generated depth maps were generally consistent with the corresponding RGB frames, the

point clouds and depth maps produced by DROID-SLAM and MAST3R-SLAM exhibited significantly less background noise than those from VGGT.

Figure 5 compares the ground truth camera trajectories (gray) with the estimated trajectories (blue) generated by each algorithm. The results indicate that the trajectories estimated by both DROID-SLAM and MAST3R-SLAM align closely with the ground truth. In contrast, the trajectories estimated by VGGT exhibit greater deviation from the reference poses.

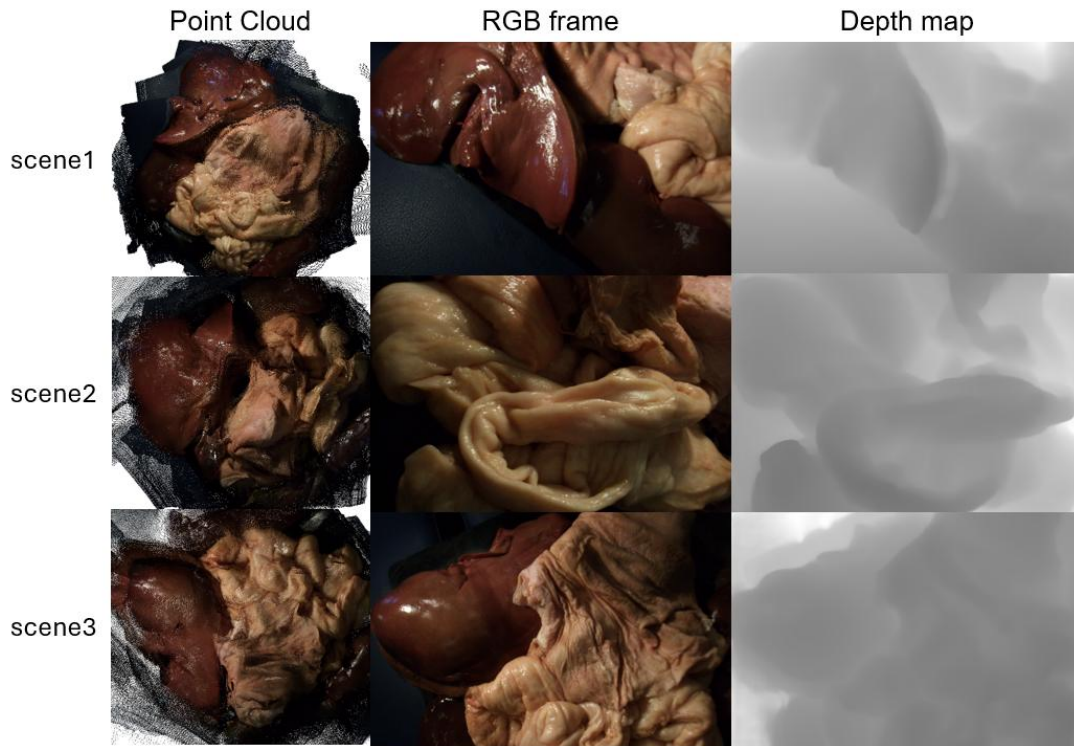


Figure 3. 3D Point clouds of tissue surface, and depth maps with their corresponding RGB frames from MAST3R-SLAM.

### 3.2 Quantitative results

The average ATE and translational RPE between the ground truth and estimated camera trajectories were evaluated. Each estimated camera trajectory was scaled and aligned with the ground truth one before evaluation to make sure the results are consistent despite using different 3D reconstruction methods. The average ATE and translational RPE of each animal tissue scene on different 3D reconstruction methods are listed in Table 1. DROID-SLAM and MAST3R-SLAM provide robust estimated trajectories with 3-4 mm of ATE. Nevertheless, MAST3R-SLAM and VGGT have higher translational RPE than those from DROID-SLAM.

Table 1. Quantitative results of the mean absolute trajectory error (ATE) and translational relative pose error (RPE) between the ground truth camera poses and the estimated camera poses from DROID-SLAM, MAST3R-SLAM, and VGGT methods.

Methods	Scene 1		Scene2		Scene3	
	ATE (mm)	Trans. RPE (mm)	ATE (mm)	Trans. RPE (mm)	ATE (mm)	Trans. RPE (mm)
DROID-SLAM	<b>4.38 ± 1.66</b>	<b>1.99 ± 1.06</b>	<b>4.38 ± 2.11</b>	<b>1.67 ± 1.05</b>	<b>3.47 ± 1.70</b>	<b>1.39 ± 1.03</b>
MASt3R-SLAM	4.63 ± 1.90	12.41 ± 6.85	4.51 ± 2.26	9.09 ± 5.52	3.92 ± 1.61	7.44 ± 5.82
VGGT	8.72 ± 4.20	16.10 ± 8.30	10.95 ± 4.62	9.24 ± 5.28	10.30 ± 4.23	7.40 ± 5.38

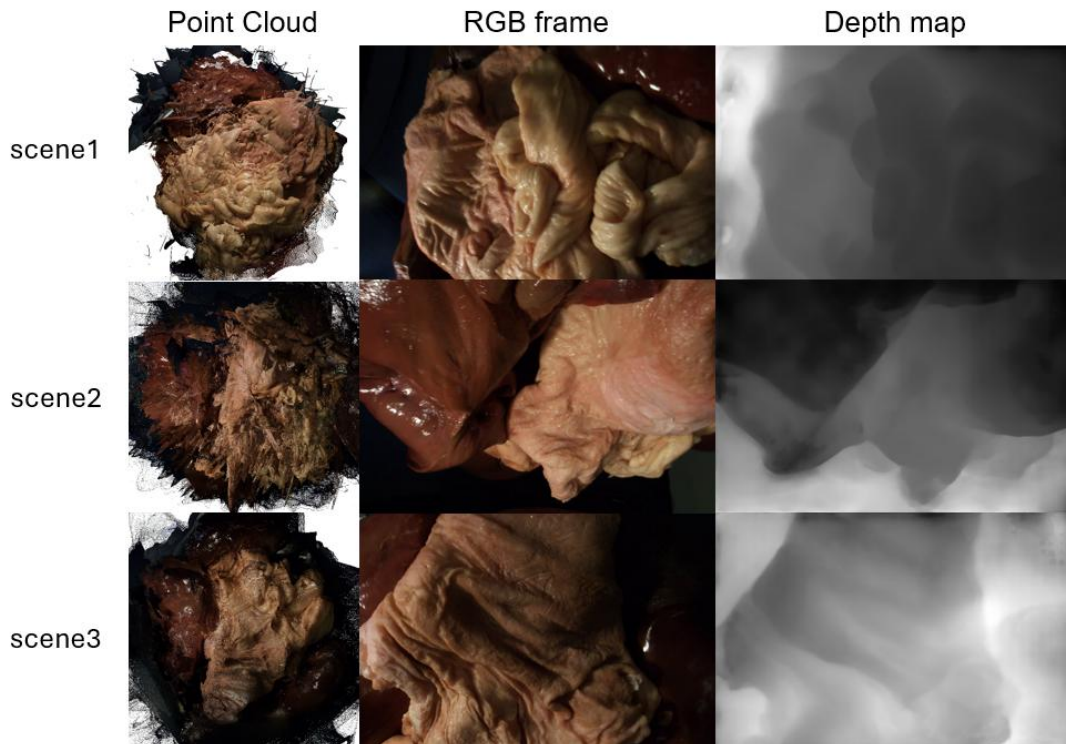


Figure 4. 3D Point clouds of tissue surface, and depth maps with their corresponding RGB frames from VGGT.

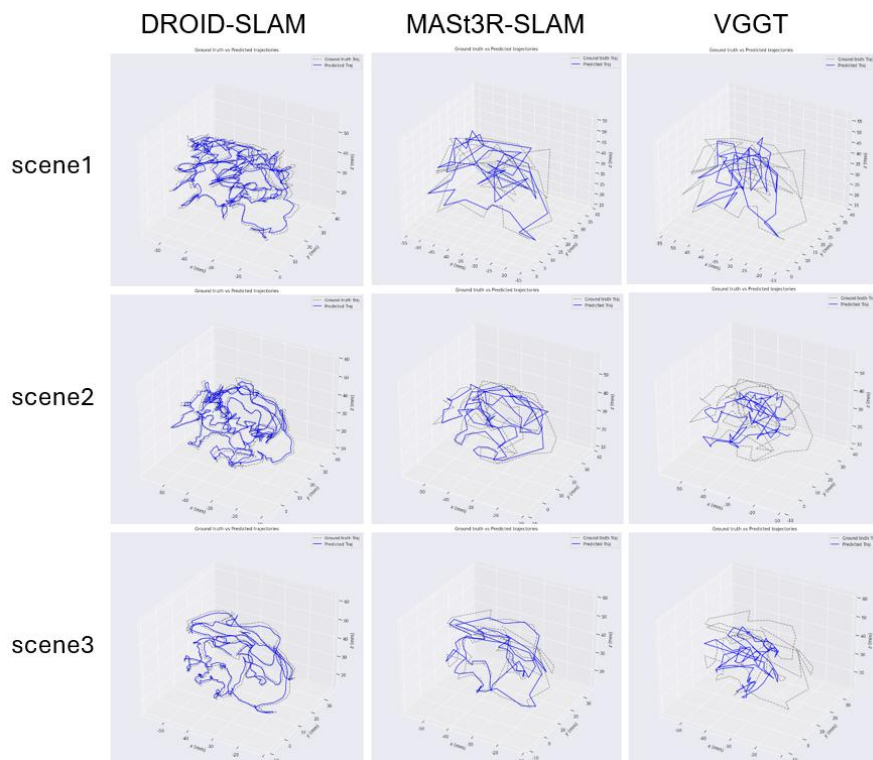


Figure 5. Ground truth (gray) and estimated (blue) camera pose trajectories from each method in different scenes.

## 4. DISCUSSION AND CONCLUSION

In this study, we evaluated three tissue surface reconstruction methods that can be used for augmented reality-guided medical interventions. However, the majority of learning-based 3D reconstruction methods have been evaluated primarily on public datasets designed for autonomous driving, which differ significantly from medical imaging scenarios. We assessed three reconstruction frameworks, DROID-SLAM, MAST3R-SLAM, and VGGT, using three distinct tissue scenes comprising multi-view 2D frames and ground truth camera trajectories acquired via an optical tracking system. We utilized absolute trajectory error and relative pose error as quantitative metrics to evaluate the performance of each model.

The results demonstrate that the point clouds generated by all methods effectively preserved the original shape and color of the tissue surfaces. Furthermore, the estimated depth maps for all scenes were consistent with their corresponding RGB frames. In terms of camera tracking, trajectories estimated by DROID-SLAM and MAST3R-SLAM exhibited strong alignment with the ground truth, whereas those from VGGT showed greater deviation. The average ATE for both DROID-SLAM and MAST3R-SLAM was comparable and significantly lower than that of VGGT. Conversely, the translational RPE for MAST3R-SLAM and VGGT was higher than that of DROID-SLAM. This discrepancy suggests that while these reconstruction methods are subject to high-frequency local noise (jitter) in frame-to-frame estimation, as captured by the higher translational RPE, these errors are predominantly zero-mean and random rather than systematic. Consequently, local errors tend to cancel out over longer sequences, preventing significant accumulation of drift. The low ATE values confirm that the systems maintain robust global structural consistency despite local instability.

Accurate reconstruction of tissue surfaces holds significant potential for advancing image-guided interventions. For instance, Sun *et al.*<sup>12</sup> demonstrated a novel depth estimation network and reconstruction framework capable of enhancing surgical scene perception and providing precise surgical site information. Such 3D spatial information could facilitate robust AR navigation and the automation of surgical tasks in robot-assisted minimally invasive surgery procedures. Comprehensive studies are necessary to further evaluate and test these techniques in clinical settings.

## ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA288379 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award Number RP240289 and RP240542. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## REFERENCES

- [1] R. Wei, B. Li, H. Mo *et al.*, "Stereo Dense Scene Reconstruction and Accurate Localization for Learning-Based Navigation of Laparoscope in Minimally Invasive Surgery," *IEEE Trans Biomed Eng.*, 70(2), 488-500 (2023). <https://www.doi.org/10.1109/TBME.2022.3195027>.
- [2] L. Boretto, E. Pelanis, A. Regensburger, and O. J. Elle, "Laparoscopic Feature-Less 3D Reconstruction Using Neural Radiance Fields and Optical Tracking," *Advances in Digital Health and Medical Bioengineering*, 601-609 (2024). [https://www.doi.org/10.1007/978-3-031-62520-6\\_67](https://www.doi.org/10.1007/978-3-031-62520-6_67).
- [3] N. Nawawithan, J. Young, P. Bettati *et al.*, "An augmented reality and high-speed optical tracking system for laparoscopic surgery," *Proc. SPIE 12928, Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling*, 129280E, (29 March 2024); <https://www.doi.org/10.1117/12.3008448>.
- [4] P. Bettati, J. Young, A. Rathgeb *et al.*, "An augmented reality-guided biopsy system using a high-speed motion tracking and real-time registration platform," *Proc. SPIE 12928, Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling*, 129281G, (29 March 2024); <https://www.doi.org/10.1117/12.3008573>.
- [5] J. Yu, K. Pruitt, N. Nawawithan *et al.*, "Dense surface reconstruction using a learning-based monocular vSLAM model for laparoscopic surgery," *Proc. SPIE 12928, Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling*, 129280J, (29 March 2024); <https://www.doi.org/10.1117/12.3008768>.
- [6] Z. Teed, and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, (2021). <https://www.doi.org/10.48550/arXiv.2108.10869>.

- [7] R. Murai, E. Dexheimer, and A. J. Davison, "MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors," The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025).  
<https://www.doi.org/10.1109/CVPR52734.2025.01556>.
- [8] J. Wang, M. Chen, N. Karaev *et al.*, "VGGT: Visual Geometry Grounded Transformer," The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), (2025).  
<https://www.doi.org/10.1109/CVPR52734.2025.00499>.
- [9] N. Nawawithan, J. Yu, K. Pruitt *et al.*, "Novel view synthesis using neural radiance fields for laparoscopic surgery navigation," Proc. SPIE 13408, Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling, 134081U, (7 April 2025); <https://www.doi.org/10.1117/12.3048817>.
- [10] Z. Zhang, and D. Scaramuzza, "A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), (2018).  
<https://www.doi.org/10.1109/IROS.2018.8593941>.
- [11] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM.," (2017);  
<https://github.com/MichaelGrupp/evo>
- [12] X. Sun, F. Wang, Z. Ma, and H. Su, "Dynamic surface reconstruction in robot-assisted minimally invasive surgery based on neural radiance fields," Int J Comput Assist Radiol Surg, 19(3), 519-530 (2024).  
<https://www.doi.org/10.1007/s11548-023-03016-8>.