

Novel view synthesis using neural radiance fields for laparoscopic surgery navigation

Nati Nawawithan^{a,b}, James Yu^{a,c}, Kelden Pruitt^{a,b}, Baowei Fei^{a,b,c*}

^aCenter for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX

^bDepartment of Bioengineering, University of Texas at Dallas, Richardson, TX

^cDepartment of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

* Corresponding author: bfei@utdallas.edu, Website: <https://fei-lab.org>

ABSTRACT

Laparoscopic surgery is a widely used minimally invasive technique to treat a variety of pathologies within the abdominopelvic region. Compared to conventional open surgery, laparoscopic surgery reduces iatrogenic trauma and hospital stay length. Nevertheless, surgeons have a limited field of view during operation because surgical instruments are inserted through the patient's body via small incisions. To improve surgeon's perception, surgical navigation techniques are implemented to improve visualization. A number of surgical navigation systems being developed are integrated with computer algorithms to enhance their efficiency. In this work, we generate novel views, depths, and camera poses of corresponding laparoscopic images to augment our dataset, which can then be used to train and evaluate learning-based computer vision algorithms such as simultaneous localization and mapping (SLAM) and visual odometry. Our existing dataset was generated from porcine tissue images using a laparoscope and an optical tracking system to capture camera poses. We performed tissue surface reconstruction of the dataset via the NeRF-SLAM algorithm. Camera poses were transformed to acquire novel views of RGB laparoscopic images and their corresponding depth images. The results show that the simulated RGB images have an average peak signal-to-noise ratios (PSNR) between 29.17 - 29.79 dB, structural similarity index measure (SSIM) value between 0.6061 - 0.6174, and learned perceptual image patch similarity (LPIPS) value between 0.568 - 0.592 compared to the corresponding input images. This work provides a foundation for the generation of synthetic datasets, which could be useful for learning-based autonomous surgical navigation systems.

Keywords: Laparoscopic surgery, image-guided intervention, neural radiance fields

1. INTRODUCTION

Modern surgical procedures have benefitted greatly from the advent of minimally invasive techniques.¹ These techniques have led to improved patient outcomes, such as reduced post-operative stay and decreased intra-operative blood loss.² The field has also advanced with the introduction of surgical robotics, which allows greater precision and surgical control.³

Computer vision (CV) algorithms rely on well-labeled data often consisting of RGB images, camera position, and depth information. Data beyond the RGB image itself has traditionally been gathered using a complex suite of sensors including depth sensors, such as LiDAR, and optical tracking systems. That said, while stereo vision is becoming more popular and access to a suite of these other tools is increasing, many systems utilize monocular endoscopes and don't include some of these advanced tools, as equipment for these data gathering techniques can be expensive and invasive when considering surgical applications.⁴ Therefore, monocular depth estimation (MDE) is required in these applications but remains a challenging problem in robotic applications and augmented/virtual reality (AR/VR).

Yang et al.⁵ introduced a foundation model to predict high-quality monocular depth maps, which is helpful in the data acquisition process. These deep learning approaches require extensive datasets to conduct training, and manually labeled datasets are scarce. Instead, a photo-realistic synthetic dataset can be used to train and evaluate learning-based CV algorithms as camera poses and depth maps are precisely labelled. For example, Teed et al.⁶ trained a neural network (DROID-SLAM) to handle simultaneous localization and mapping (SLAM) problems using a synthetic dataset called TartanAir.⁷ The performance of DROID-SLAM showed that the predicted camera poses on real-world datasets were robust and accurate after training with the synthetic dataset. Our work seeks to address some of these monocular endoscopes' drawbacks to collect surgical data by exploring the use of neural radiance fields (NeRFs) to generate novel views from arbitrary camera poses. This information can improve the accuracy of surface reconstruction, which has the potential to

improve downstream tasks such as AR applications in medical procedures.⁸⁻¹³ For example, AR assistance can improve the targeting accuracy compared to conventional biopsy of soft tissue lesions,¹⁴ and prostate interventions¹⁵ including laparoscopic surgery.¹⁶ Yu et al.⁹ demonstrated that DROID-SLAM's performance in monocular camera pose estimation can translate well to laparoscopic surgery.

The aim of our work is to demonstrate the feasibility of novel view synthesis to create augmented datasets for future development and fine-tuning of CV algorithms. Our work focuses on the ability of neural radiance fields (NeRFs) to generate synthetic minimally invasive surgery data and improving the quality of 3D tissue reconstruction by MDE. Our work highlights the potential of simulated NeRF datasets to further CV capabilities in surgical settings and paves the way for advanced applications in abdominal surgery.

2. MATERIALS AND METHODS

2.1 Tissue imaging procedure

Our experimental setup consists of porcine abdominal tissues (liver, spleen, kidney, stomach, and intestine), a laparoscope (WAIR100A, Olympus), and an optical tracking system as illustrated in Figure 1. The laparoscope was integrated with an RGB camera (E3ISPM, Touptek) and a halogen light source (OSL2 High-Intensity Fiber-Coupled Illuminator, Thorlabs, Newton, NJ) to illuminate the tissue surface. Laparoscopic poses were acquired at 120 fps using retroreflective markers and infrared tracking. The operator moved the laparoscope about the animal tissue until the whole scene was sufficiently scanned by video. The entire laparoscopic video feed was recorded for ~2-4 minutes.

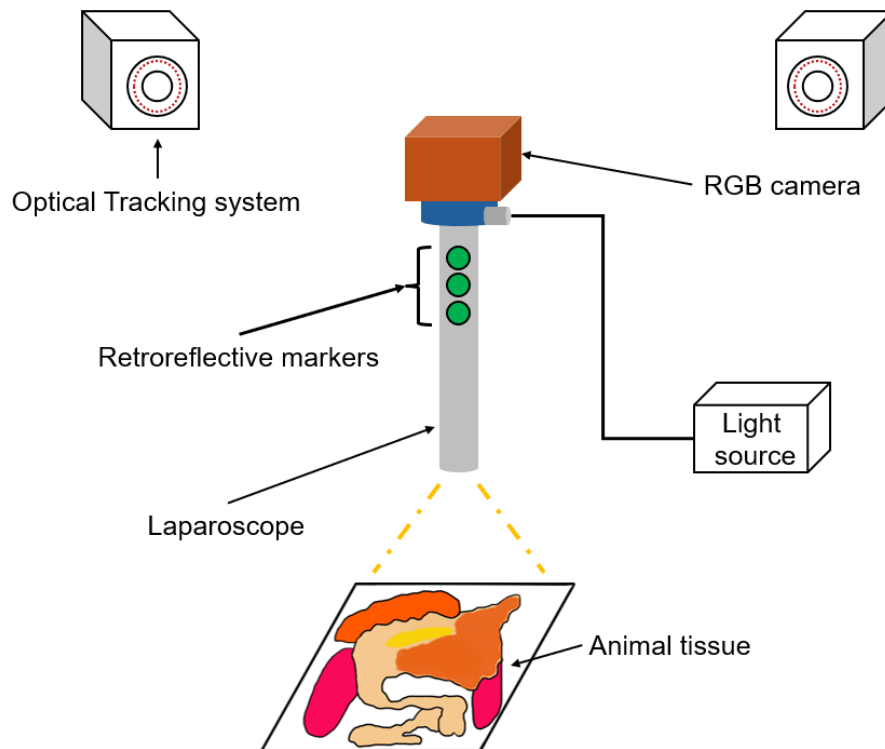


Figure 1. Experimental setup for laparoscopic imaging and optical tracking.

2.2 Synthetic dataset generation

The overview of the pipeline used for dataset synthesis is shown in Figure 2. The laparoscopic video feed was first broken into individual frames to be fed into the network. The number of frames were sampled down to 1415 frames. Each frame has a resolution of 1920×1080. The camera poses and corresponding key frames of the image sequence were estimated

from the DROID-SLAM algorithm. The computation process was performed on our high-performance computer (HPC) using 2 RTX A6000 graphics processing units (GPUs). Then, ground truth camera poses from the optical tracking system were utilized for validation and alignment of the estimated poses. Depth maps for each frame were created from the Depth Anything model.⁵ Calibration board imaging was performed to calculate the camera intrinsic parameters using OpenCV. RGB images, depth maps, camera poses, and camera intrinsic parameters were then provided as inputs to NeRF-SLAM¹⁷ for training a neural implicit representation of the tissue surface. We tested different depth supervision weights from zero to one when training the neural implicit representations. The NeRF was then saved for novel view synthesis to expand the original dataset.

To generate RGB and depth images from the tissue surface, the trained model was inferred programmatically.¹⁸ RGB and depth images from novel views were rendered along the converted camera trajectory from world coordinate system to NeRF's coordinate system. The generated RGB and depth images were then evaluated qualitatively and quantitatively. Peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) were used as quantitative metrics to evaluate photometric accuracy. The higher value of PSNR leads to the better quality of an image. If the SSIM value is closer to 1, the comparing images are more similar. For LPIPS, the lower value means the distance between image patches is smaller or the images are more identical.

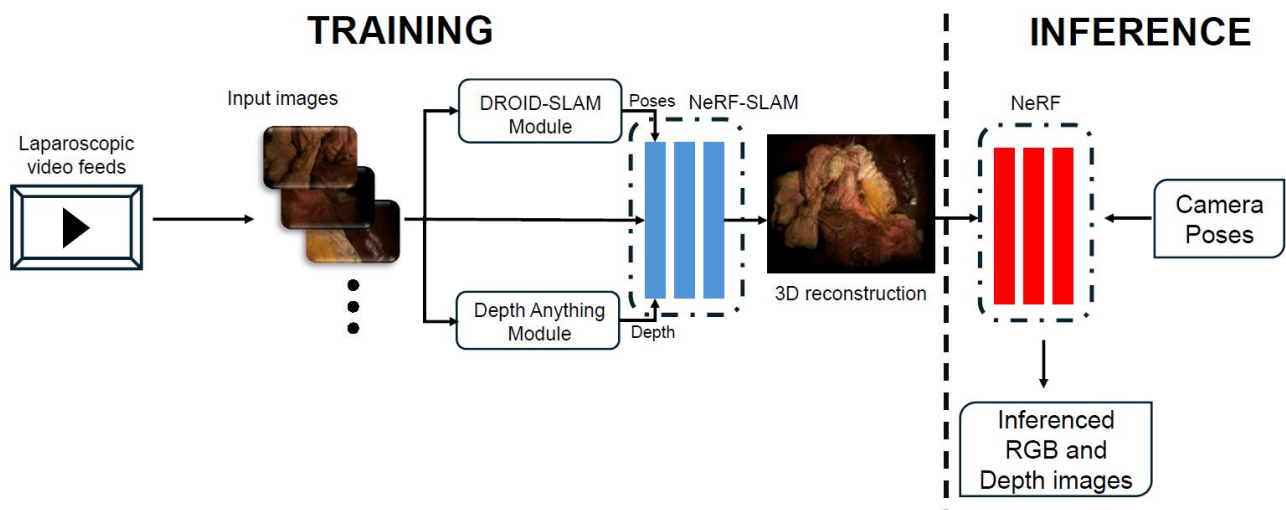


Figure 2. Pipeline for monocular depth estimation, NeRF training, and synthesis of novel views and depths.

3. RESULTS

3.1 Qualitative results

Examples of 3D reconstruction, simulated RGB and depth images from the same frame at different depth supervision levels are depicted in Figure 3. 3D reconstruction of the animal tissue with a depth supervision weight of 1 provided more details than other depth supervision levels. In addition, depth maps from the trained neural implicit representations without depth supervision were noisier than the others with some degree of depth supervision.

3.2 Quantitative results

The results of each of the metrics of interest from different depth supervision levels are reported in Table 1. The simulated RGB images were rendered at the identical camera pose of the corresponding extracted video frames for evaluation. An average PSNR was measured between the input RGB images and the corresponding simulated RGB images at a recent step of the trained weight. PSNR can be improved with depth supervision during training. The trained weight with depth supervision weight of 1 achieved the highest PSNR in this study.

Table 1. Quantitative results of the simulated RGB images (PSNR in dB, SSIM and LPIPS in range of 0-1).

Depth supervision	PSNR	SSIM	LPIPS
0%	29.17	0.6095	0.577
25%	29.77	0.6174	0.568
50%	29.73	0.6083	0.582
75%	29.42	0.6061	0.592
100%	29.79	0.6081	0.583

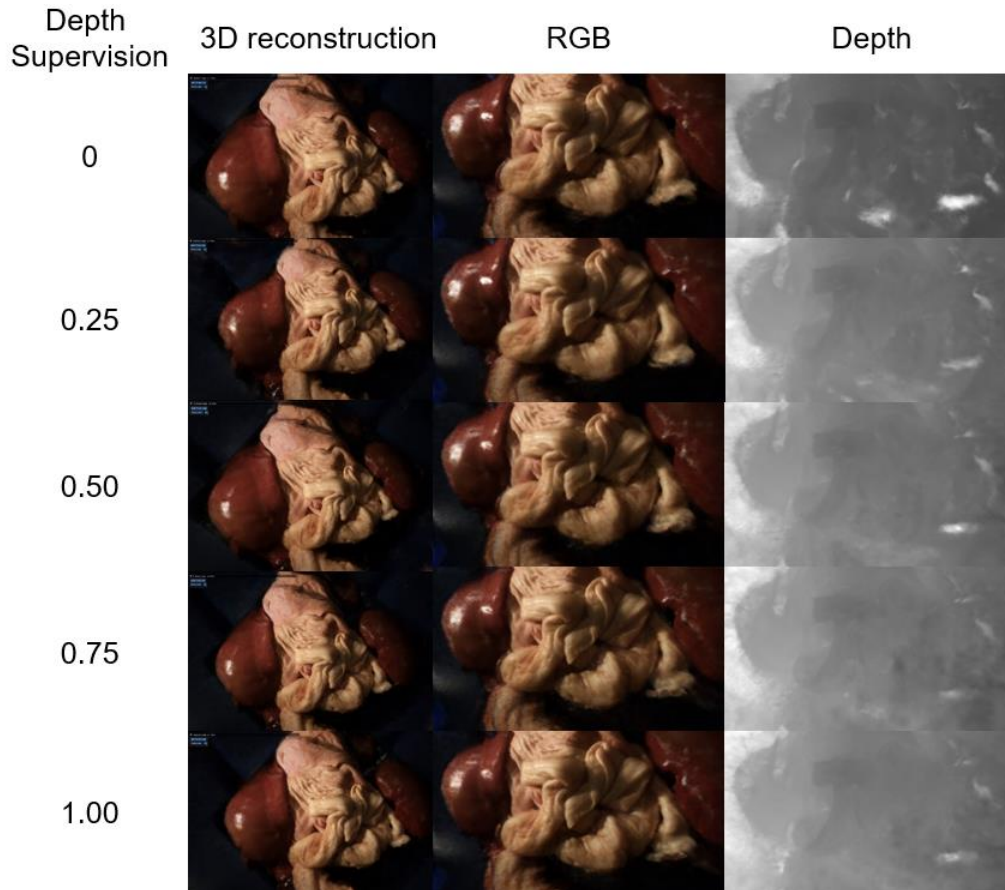


Figure 3. 3D reconstruction, simulated RGB and depth images under different depth supervision levels.

4. DISCUSSION AND CONCLUSION

Our laparoscopic dataset simulation framework was used to augment an existing dataset by rendering novel views of RGB and depth images from an implicit tissue surface representation. This framework has the potential to expand datasets with richer information to be used for training and evaluation for learning-based SLAM algorithms. Moreover, the information from the learning-based CV algorithms can facilitate autonomous surgical navigation systems, providing an accurate landscape for safe navigation. We demonstrated the ability to generate neural implicit representations from monocular RGB frames and inferred them to render novel views and depths across sampled trajectories, which can be used for downstream applications. Our qualitative and quantitative results show that applying depth supervision during neural implicit representation training can help to improve the photometric accuracy of RGB images. To our knowledge, this is

the first usage of MDE for improving tissue surface reconstruction results. Further studies are required to investigate the proposed improvement to dense reconstructions of SLAM algorithms and to explore surgical navigation applications using these improved surface reconstructions.

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 R01CA204254 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] I. Alkatout, U. Mechler, L. Mettler *et al.*, “The Development of Laparoscopy-A Historical Overview,” *Front Surg*, 8, 799442 (2021).
- [2] B. Madhok, K. Nanayakkara, and K. Mahawar, “Safety considerations in laparoscopic surgery: A narrative review,” *World J Gastrointest Endosc*, 14(1), 1-16 (2022).
- [3] C. Schneider, M. Allam, D. Stoyanov *et al.*, “Performance of image guided navigation in laparoscopic liver surgery - A systematic review,” *Surg Oncol*, 38, 101637 (2021).
- [4] I. Rivas-Blanco, C. J. Perez-Del-Pulgar, I. Garcia-Morales, and V. F. Munoz, “A Review on Deep Learning in Minimally Invasive Surgery,” *IEEE Access*, 9, 48658-48678 (2021).
- [5] L. Yang, B. Kang, Z. Huang *et al.*, “Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10371-10381 (2024).
- [6] Z. Teed, and J. Deng, “DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras,” *Proceedings of the 35th International Conference on Neural Information Processing Systems*, (2021).
- [7] W. Wang, D. Zhu, X. Wang *et al.*, “TartanAir: A Dataset to Push the Limits of Visual SLAM,” *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2020).
- [8] M. Pfefferle, S. Shahub, M. Shahedi *et al.*, “Renal biopsy under augmented reality guidance,” *Proc SPIE Int Soc Opt Eng*, 11315, (2020).
- [9] J. Yu, K. Pruitt, N. Nawawithan *et al.*, “Dense surface reconstruction using a learning-based monocular vSLAM model for laparoscopic surgery,” *Proc SPIE Int Soc Opt Eng*, 12928, (2024).
- [10] P. Bettati, and B. Fei, “An Augmented Reality System with Advanced User Interfaces for Image-Guided Intervention Applications,” *Proc SPIE Int Soc Opt Eng*, 12466, (2023).
- [11] R. Fraser, P. Bettati, J. Young *et al.*, “A Fast and Interactive Augmented Reality System for PET/CT-guided Intervention of Neuroblastoma,” *Proc SPIE Int Soc Opt Eng*, 12928, (2024).
- [12] J. Huang, M. Halicek, M. Shahedi, and B. Fei, “Augmented reality visualization of hyperspectral imaging classifications for image-guided brain tumor phantom resection,” *Proc SPIE Int Soc Opt Eng*, 11315, (2020).
- [13] P. Bettati, J. D. Dormer, J. Young *et al.*, “Virtual Reality Assisted Cardiac Catheterization,” *Proc SPIE Int Soc Opt Eng*, 11598, (2021).
- [14] P. Bettati, M. Chalian, J. Huang *et al.*, “Augmented Reality-Assisted Biopsy of Soft Tissue Lesions,” *Proc SPIE Int Soc Opt Eng*, 11315, (2020).
- [15] 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 Int Soc Opt Eng*, 12928, (2024).
- [16] N. Nawawithan, J. Young, P. Bettati *et al.*, “An augmented reality and high-speed optical tracking system for laparoscopic surgery,” *Proc SPIE Int Soc Opt Eng*, 12928, (2024).
- [17] A. Rosinol, J. J. Leonard, and L. Carlone, “NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields,” *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2023).
- [18] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Transactions on Graphics*, 41(4), 1-15 (2022).