3D Scene Reconstruction for IMA Division Surgery

Anand Srinivasan
anand.srinivasan@yale.edu

Advisor: Daniel Rakita
daniel.rakita@yale.edu

A Senior Thesis as a partial fulfillment of requirements for the Bachelor of Science in Computer Science

Department of Computer Science
Yale University
Dec 12, 2024

Acknowledgements

I want to express my gratitude to my advisor Dr. Daniel Rakita and my two PhD student mentors Xiatao Sun and TJ Vitchutripop for their support and guidance throughout this project. Special thanks the Yale Department of Computer Science for providing a fantastic environment for academic growth.

Contents

3D Scene Reconstruction for IMA Division Surgery

Anand Srinivasan

Abstract

Many colorectal surgeries require division of the inferior mesenteric artery (IMA), yet standardized metrics for this common procedure do not exist. 3D scene reconstruction is a necessary step towards the generation of such metrics. Various challenges exist for 3D scene reconstruction in the endoscopic domain. Ambiguous lighting conditions, dynamic tissue and tool movement, and a lack of camera intrinsic and extrinsic information all present significant obstacles for robust scene reconstruction. Recently, Gaussian splatting has gained popularity for its ability to efficiently represent complex 3D scenes. For endoscopic 3D scene reconstruction of IMA division surgery, two Gaussian splatting-based algorithms, Endo-4DGS and InstantSplat, are tested with the aim of reconstructing a small clip of video surrounding a single IMA division. Both models vary slightly in regard to expected input and use cases. Endo-4DGS requires sparse 2D images, initial 3D point cloud predictions, and camera extrinsic estimations for initialization. It is built to handle dynamic scenes. InstantSplat requires sparse 2D images as input and is designed to reconstruct static scenes. For Endo-4DGS, depth maps are constructed using depth estimation models to generate initial 3D point cloud estimates. An extrinsic estimation pipeline consisting of SIFT feature matching ahead of the Kabsch algorithm is used to determine relative camera poses. Results show reasonable success from both models at characterizing 3D scenes from the narrow set of viewpoints that are well-represented in the training dataset. Characterizing scenes from lesser-represented viewpoints remains challenging. The findings indicate that in their current form, 3D scene reconstructions generated using the approaches tested are not sufficiently robust for optimal surgical metric analysis. Future work will aim to test alternative approaches that allow for the incorporation of a larger set of viewpoints as input.

1 Introduction

Inferior mesenteric artery (IMA) division is a common component of colorectal surgery to help remove lymph nodes in patients with colorectal cancer [1, 2, 3]. Surgeons operate laparoscopically from a robotic surgical system, using a Da Vinci stapler – a state-of-the-art stapling device equipped with real-time feedback and fine-tuned control – to make a precise incision in the IMA. These procedures require adequate separation of the IMA from proximal vessels and careful positioning of the Da Vinci stapler device to ensure safe division and prevent potential anastomotic leaks. Studies show an anastomotic leakage incidence of up to 6-10% in surgical patients [4, 5]. Despite the significant precision requirement of IMA division and the high rate of leakage post-surgical resection, standard surgical metrics for IMA division do not exist. Such metrics would be useful for training residents to help guide appropriate stapler positioning and proper sequestering of the IMA before division, as opposed to relying on feel and experience for these important determinations, which remains the standard of care to date. Moreover, metrics would be helpful to standardize care across colorectal surgeons and assess surgical performance in order to maintain a high standard of care.

As a step towards creating robust surgical metrics for IMA division, 3D scene reconstruction of surgical video data is needed to allow for accurate angular and volumetric calculations within a standard 3D reference frame. Current state-of-the-art endoscopic 3D scene reconstruction approaches rely on Gaussian splatting, a recently popularized method which represents points in 3D space as Gaussian distributions, for scene representation. Gaussian splatting has been found to achieve robust novel view synthesis, indicating its capacity for 3D scene representation [6].

In this work, two Gaussian splatting-based reconstruction approaches are tested: Endo-4DGS and InstantSplat. Endo-4DGS is a dynamic endoscopic reconstruction approach, which leverages MLPs to infer temporal information [7]. InstantSplat is an SfM-free static reconstruction approach that infers extrinsic camera pose information through pixel alignment [8]. While InstantSplat works directly with a set of sparse 2D images as input, the Endo-4DGS model requires initial point cloud and extrinsic camera pose estimates for initialization. For this purpose, the MoGe general purpose 3D geometry estimation model is used to generate depth maps and corresponding 3D point cloud estimates for each image [9]. A camera extrinsic estimation pipeline is created, using SIFT key point matching ahead of the Kabsch for determining optimal relative transformation matrices between sequential image pairs [10, 11].

For an assessment of 3D scene reconstruction performance, both models are tested on a sample video clip from a single IMA division surgery. Frames are selected from a short, 10-second section of video that includes the physical division of the IMA by the Da Vinci stapler. The Da Vinci stapler is held steady by the surgeon for the duration of the incision,

allowing it to function as a spatial anchor during pose estimation.

Qualitative reconstruction results show that both models achieve reasonable success at characterizing 3D scenes from the narrow set of viewpoints that are well represented in the training dataset. Characterizing scenes from lesser-represented viewpoints remains challenging. The findings indicate that in their current form, 3D scene reconstructions generated using the Guassian splatting techniques tested are not sufficiently robust for optimal surgical metric analysis.

To achieve more robust reconstruction results, methods which can incorporate a more expansive portion of the IMA division video should be tested. Expanding the applicable portion of surgical video will require either a more robust method for extrinsic camera pose estimation in endoscopic scenes that does not rely on a spatial anchor or the adoption of a dynamic reconstruction model which is able to infer extrinsic camera poses.

2 Methodology

2.1 Dataset

The IMA division dataset consists of four endoscopic surgical videos of colorectal surgeries that include an IMA division. For the analysis of the efficacy of reconstruction approaches tested, a short 10-second clip from a single video was chosen to be reconstructed. This clip shows the physical division of the IMA, therefore representing the most important moment for metric analysis in IMA division surgery. The clip includes limited dynamic tool movement, making it more suitable for both reconstruction approaches.

2.2 Initialization Overview

Endo-4DGS requires initial point cloud and extrinsic camera pose estimates for initialization. Although end-to-end scene reconstruction tools like COLMAP exist for camera pose and point cloud estimation from a set of sparse images, these tools struggle to analyze endoscopic scenes due to the dynamic nature of deformable tissue and the inconsistent lighting conditions throughout endoscopic datasets. In place of standard SfM algorithms, depth maps are used in conjunction with camera intrinsic estimations to construct 3D point clouds for each image in a dataset. A variety of depth estimation models exist for estimating depth from monocular image input, and three are tested in this work. For extrinsic pose estimation, matched 2D key points are deduced between sequential image pairs using a scale-invariant feature transform (SIFT). The Kabsch algorithm is subsequently applied to compute relative transformations between sequential images using these matched key points.

2.3 Depth Estimation

Three depth estimation models are tested on the IMA dataset: Depth Anything [12], EndoDAC [7], and MoGe [9]. Depth Anything is a general-purpose depth estimation model trained on over 63 million images. EndoDAC is a self-supervised depth estimation framework meant to adapt foundation models to analyze endoscopic scenes. Finally, MoGe is a general-purpose 3D geometry estimation model which estimates depth, camera intrinsics, and 3D point clouds from a monocular image input. Since ground truth depth information is not available for the IMA division videos, all models are used as zero-shot prediction models. Moreover, due to the lack of ground truth comparison values, all prediction models are solely assessed qualitatively at this stage in the pipeline.

2.4 Point Cloud Estimation

Depth maps are used in conjunction with an estimation of the monocular camera intrinsic matrix to estimate 3D point clouds within the camera's local reference frame. The 3D projection of a pixel (u, v) is calculated as follows:

$$p_x = d(u - c_x)/f_x$$
$$p_y = d(v - c_y)/f_y$$
$$p_z = d$$

where d is the depth value of pixel (u, v) determined by the depth map, (c_x, c_y) is the camera's principal point, (f_x, f_y) are the camera's horizontal and vertical focal lengths, and p= (p_x, p_y, p_z) is the 3D projection of pixel (u, v). The endoscopic camera's intrinsics are estimated by the MoGe model, and this estimate is used for the 3D point projection of all depth maps.

2.5 Extrinsic Estimation

In order to estimate the relative transformation matrix T between a pair of consecutive images I1 and I2, a known set of matched key points representing analogous spatial points between the two distanced views is needed. Scale-invariant feature transform (SIFT) is a popular algorithm for determining these key points [10]. The application of SIFT to endoscopic images, however, is made difficult by poor lighting conditions, homogenous tissue appearance, and the deformable nature of tissues themselves, leading to a high proportion of outlier key point matches. To address this challenge, the DaVinci stapler is selected as an anchor on which to locate key points. First, Meta's Segment Anything model [13] is used to generate tool masks for the DaVinci stapler in each frame. SIFT is subsequently employed to locate key points between the masked regions on the 2D images.

Although selecting the DaVinci stapler as an anchor allows for the detection of robust key points, this choice also imposes a set of restrictions on the portion of video data permissible to use in the reconstruction. Because using the DaVinci stapler as an anchor assumes that its position is static in global 3D space, any clip selected for reconstruction must not include motion of the stapler itself, as any change in positioning of the stapler on screen is attributed to extrinsic camera pose changes. The portion of video immediately surrounding the division of the IMA remains valid for reconstruction, since the surgeon must hold the stapler steady when performing the resection.

After determining a set of matched key points, the Kabsch algorithm is employed to determine the relative transformation from a source image to a target image based on the set of matched 3D projections across the two images. Consider a set of source points $P \in R^{Nx3}$ and the set of matched target points $Q \in R^{Nx3}$ in a subsequent image. Our objective is to find a transformation matrix $T \in SE(3)$ such that the mean square distance between target and transformed points, calculated as:

$$\sum_{i}^{N} \| Q_i - T(P_i) \|^2$$

is minimized, where $T(P_i)$ represents the $SE(3)$ matrix transformation of points P_i. Following the Kabsch algorithm, we can center each set of points by computing their respective centroids $\overline{P}$ and $\overline{Q}$ and subtracting these centroids from their respective point distributions to find centered point distributions P^* and Q^*. The covariance matrix $H = P^{*T}Q^*$ can be decomposed using singular value decomposition: $H = U\Sigma V^T$. The optimal rotation matrix R can then be computed as $R = VU^T$ and the optimal translation vector $\mathbf{t}$ can be computed as $\mathbf{t} = \overline{Q} - R\overline{P}$. Finally, the transformation matrix T can be composed as $\begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix}$.

For a dataset $\mathbf{D} = \{I_1, I_2, \ldots, I_N\}$ of N images, relative transformations are estimated between each pair of consecutive frames, with the prior frame acting as the source image and the later frame acting as the target, in order to generate a set of N–1 transformations $\mathbf{T} = \{T_{1ß2}, T_{2ß3}, \ldots, T_{(N-1)ßN})\}$. The reference frame of the first image I_1 is taken to be the global frame. The camera pose of all subsequent images is calculated in reference to this global frame. For an image I_k, the extrinsic pose P_k is determined as:

$$P_k = T_{1\to k} = T_{(k-1)\to k} \cdot T_{(k-2)\to(k-1)} \cdot \ldots \cdot T_{1\to 2}$$

2.6 Gaussian Splatting

For endoscopic scene reconstruction, two Gaussian splatting-based algorithms are tested: Endo-4DGS and InstantSplat. Camera pose estimations derived from the above extrinsic estimation pipeline and MoGe-generated point clouds are used for Endo-4DGS initialization. Both algorithms are tested on the single 10-second clip chosen for reconstruction.

3 Results

3.1 Depth Estimation

Zero-shot depth estimation results for all three models are shown in figure 1. Although ground truth depth information is not available for the IMA dataset, qualitatively, MoGe seems to achieve the best depth predictions. Notably, the MoGe and Depth Anything models seem to better distinguish tool edge geometry in comparison to the EndoDAC model.

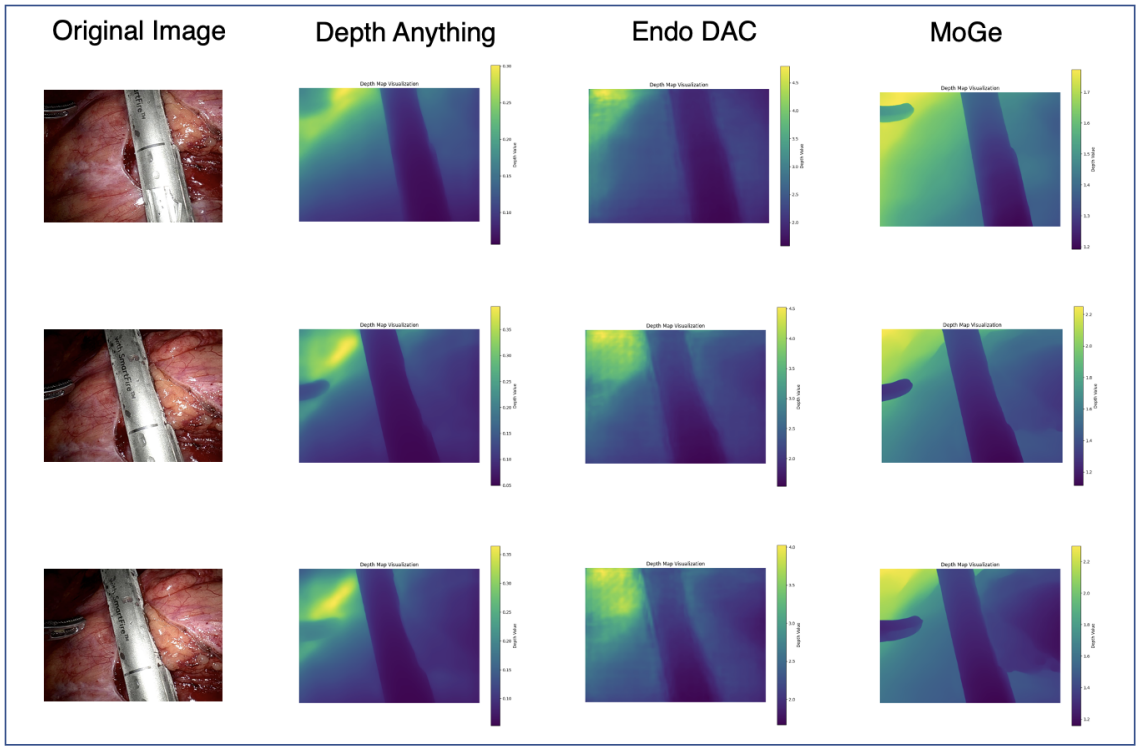

Fig. 1: Zero-shot depth map predictions for Depth Anything, EndoDAC, and MoGe models on three representative images.

3.2 Point Cloud Estimation

3D point cloud projections for images from depth maps generated by all depth models tested are depicted in figure 2. All projections use camera intrinsics estimated by the

MoGe model. Supplemental videos 1-3 show visualizations of Depth Anything, Endo-DAC, and MoGe depth-based point clouds from multiple viewpoints. Qualitatively, the MoGe-generated point cloud again appears to best estimate the scene's geometry.

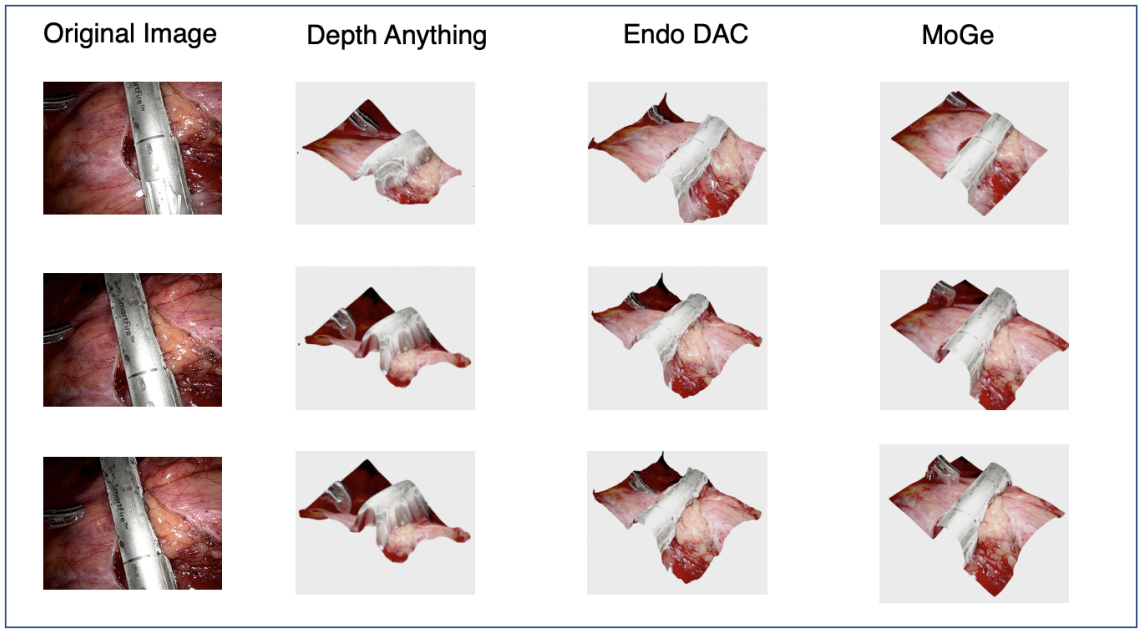

Fig. 2: 3D point cloud projections using depth maps generated by Depth Anything, EndoDAC, and MoGe models on three representative images. Camera intrinsics estimated by MoGe model for all projections. MoGe estimated relative focal lengths of $(f_x, f_y) = (1.00, 1.35)$ and a relative principal point of $(c_x, c_y) = (0.5, 0.5)$.

3.3 Extrinsic Estimation

SIFT results with and without using the DaVinci stapler as an anchor are depicted in figure 3. Coarse- and fine-grained time steps between sequential images (1 seconds and 0.1 seconds, respectively) are tested. On average across a 15 second dataset (150 fine-grained frames or 15 coarse grained frames), smaller, anchored transformations result in smaller total squared distances between transformed matched key points in comparison to larger, unanchored transformations. Supplemental video 4 depicts an anchored transformation between two 3D point clouds estimated from frames that were a second apart. The transformation matrix applied is calculated as the chain of 10 fine-grained transformation matrices.

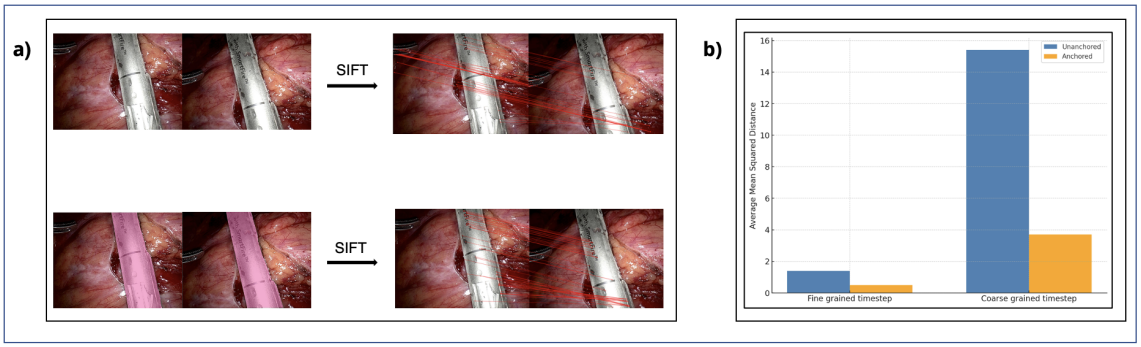

Fig. 3: **(a)** Visualization of SIFT point matches across a representative sequential pair of images using unanchored (top) and anchored (bottom) approaches. **(b)** Average total squared distance between matched, transformed key points for sequential image pairs for coarse and fine-grained approaches with and without the use of the DaVinci stapler as an anchor.

3.4 Gaussian Splatting

Gaussian splatting results for the Endo-4DGS and InstantSplat models are shown in figure 4. Supplementary videos 5-6 show a larger collection of synthesized views from the Endo-4DGS and InstantSplat scene reconstruction models. Both models achieve reasonable results for viewpoints that are well-represented in the training set, but are unable to effectively represent the scenes from angles that are less represented in the training set.

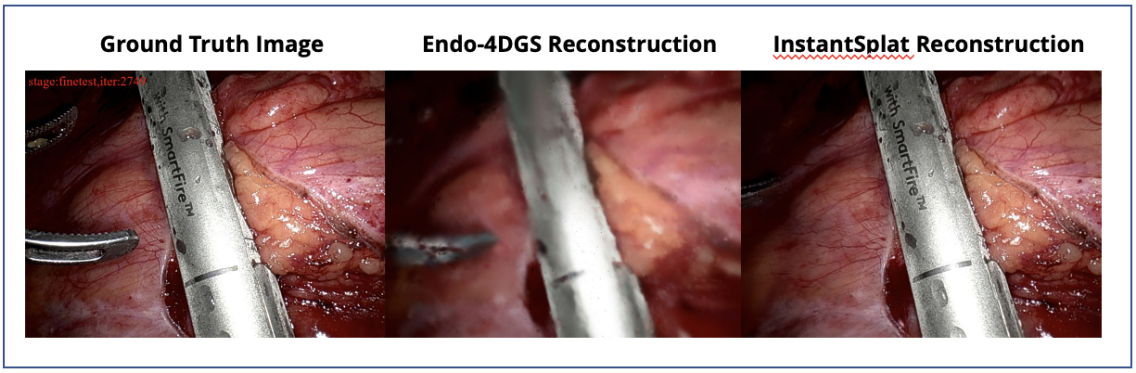

Fig. 4: Qualitative Endo-4DGS and InstanstSplat reconstruction results for an example view that was well represented in the training dataset.

4 Related Work

Traditional 3D scene reconstruction approaches leverage structure from motion (SfM) approaches ahead of Mutli-View Stereo (MVS) algorithms for camera pose estimation and dense 3D point cloud reconstruction. End-to-end tools like COLMAP automate this pipeline [14]. Although effective for controlled, static scenes, traditional scene reconstruction approaches struggle with endoscopic datasets due to poor lighting conditions and dynamic elements like deformable tissues and moving surgical tools.

Recently, Neural Radiance Fields (NeRF) emerged as a new method for novel view synthesis from a sparse set of input images using MLPs for continuous volumetric scene representations [15]. EndoNeRF presents a method to leverage dynamic neural radiance field for deformable endoscopic tissue reconstruction tasks. This approach assumes a single, fixed viewpoint setting, limiting its applicability to many real-world endoscopic datasets [16].

3D Gaussian Splatting uses 3D Gaussians to represent scenes, avoiding empty space computations. It also introduces a fast visibility-aware rendering algorithm, effectively reducing training times for 3D reconstruction while maintaining the quality achieved by NeRF reconstructions [6]. Endo-4DGS uses 3D Gaussian Splatting for 3D scene representation in combination with lightweight MLPs for capturing temporal dynamics in endoscopic settings [7].

The above NeRF and Gaussian splatting-based scene reconstruction methods require accurate camera pose information for multi view scenes. In endoscopic scenarios where SfM approaches for pose estimation can be unreliable, this restriction can preclude the applicability of such methods. If inaccurate pose information is used as input, reconstruction quality can deteriorate.

Aside from SfM approaches, other algorithms exist for relative pose estimation from a set of 2D image input. The scale-invariant feature transform (SIFT) is a common algorithm used for detecting analogous keypoints between two distinct views of the same scene [10]. The Kabsch algorithm is a method for finding the optimal rotation and translation from one set of 3D point onto another [11]. Placing SIFT ahead of the Kabsch algorithm therefore allows for the estimation of relative poses between distinct image pairs. Still, this approach is limited in its need for a static anchor within each set of image pairs in order to ensure accurate extrinsic estimation.

In order to address the difficulty of obtaining camera pose information for certain types of scenes, InstantSplat introduces an SfM-free self-supervised framework to generate 3D reconstructions from sparse 2D input without the need for camera pose input [8]. EndoGSlam integrates a Gaussian scene representation with a simultaneous localization and mapping (SLAM) approach to construct a map of the 3D scene while determining the location of the camera within this map for camera pose estimation [17]. Although

these methods circumvent the issue of camera pose estimation, they are designed for reconstructing static scenes, making them partially ill-suited for dynamic endoscopic scene reconstruction.

5 Discussion and Future Work

Both Gaussian splatting approaches tested have significant limitations. Endo-4DGS requires extrinsic information as input. In order to obtain this extrinsic information reliably, the DaVinci stapler was selected as a spatial anchor. As a result, the portion of video applicable to the reconstruction was restricted to that in which the stapler was static. Although the portion of video immediately surrounding the division met these criteria, these frames included a limited diversity of 2D viewpoints, making it challenging to reconstruct the scene from a large range of angles. Similarly, InstantSplat requires static scene input, similarly limiting the applicable portion of video. Because both approaches introduce strict restrictions that limit the valid range of input views, thereby rendering them incapable of representing the 3D scene from a wide variety of angles, the application of these approaches for accurate volumetric calculations is currently infeasible.

Future work should aim to accurately model surgical IMA division scenes across a greater diversity of viewpoints. In order to expand the portion of video data applicable to 3D scene reconstruction, alternative extrinsic-free approaches (like InstantSplat) that also incorporate a method of modeling temporal scene dynamics (like Endo-4DGS) should be tested. Alternatively, different pose estimation approaches could be tested ahead of dynamic reconstruction approaches like Endo-4DGS. If robust endoscopic pose estimation could be accomplished without the requirement of a spatial anchor, a larger range of views would become available to the model as input. This greater diversity of input would likely allow for a more accurate 3D scene reconstruction.

Furthermore, although the current 3D scene reconstruction may not be immediately applicable for accurate volumetric calculations, the estimated 3D point clouds generated from depth maps may allow for robust 3D angle estimations between relevant clinical segmentations.

Bibliography

[1] Giulio Mari et al. ""High or low Inferior Mesenteric Artery ligation in Laparoscopic low Anterior Resection: study protocol for a randomized controlled trial" (HIGHLOW trial)". In: *Trials* 16 (Jan. 27, 2015), p. 21. ISSN: 1745-6215. DOI: 10.1186/s13063-014-0537-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4311448/ (visited on 09/15/2024).

[2] Xiaolan You et al. "High versus low ligation of inferior mesenteric artery during laparoscopic radical resection of rectal cancer: A retrospective cohort study". In: *Medicine* 99.12 (Mar. 2020), e19437. ISSN: 1536-5964. DOI: 10.1097/MD.0000000000019437.

[3] Antonio Brillantino et al. "Inferior Mesenteric Artery Ligation Level in Rectal Cancer Surgery beyond Conventions: A Review". In: *Cancers* 16.1 (Jan. 2024). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 72. ISSN: 2072-6694. DOI: 10.3390/cancers16010072. URL: https://www.mdpi.com/2072-6694/16/1/72 (visited on 09/15/2024).

[4] Melissa N. N. Arron et al. "Mesenteric occlusive disease of the inferior mesenteric artery is associated with anastomotic leak after left-sided colon and rectal cancer surgery: a retrospective cohort study". In: *International Journal of Colorectal Disease* 37.3 (2022), pp. 631–638. ISSN: 0179-1958. DOI: 10.1007/s00384-021-04089-0. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8885551/ (visited on 09/15/2024).

[5] S. Fujii et al. "Randomized clinical trial of high versus low inferior mesenteric artery ligation during anterior resection for rectal cancer". In: *BJS Open* 2.4 (June 8, 2018), pp. 195–202. ISSN: 2474-9842. DOI: 10.1002/bjs5.71. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6069351/ (visited on 09/15/2024).

[6] Bernhard Kerbl et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *ACM Transactions on Graphics* 42.4 (Aug. 2023), pp. 1–14. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/3592433. URL: https://dl.acm.org/doi/10.1145/3592433 (visited on 09/15/2024).

[7] Yiming Huang et al. *Endo-4DGS: Endoscopic Monocular Scene Reconstruction with 4D Gaussian Splatting*. Apr. 2, 2024. DOI: 10.48550/arXiv.2401.16416. arXiv: 2401.16416[cs]. URL: http://arxiv.org/abs/2401.16416 (visited on 12/12/2024).

[8] Zhiwen Fan et al. *InstantSplat: Sparse-view SfM-free Gaussian Splatting in Seconds*. Aug. 20, 2024. DOI: 10.48550/arXiv.2403.20309. arXiv: 2403.20309[cs]. URL: http://arxiv.org/abs/2403.20309 (visited on 12/12/2024).

[9] Ruicheng Wang et al. *MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision*. Nov. 27, 2024. DOI: 10.48550/arXiv.2410.19115. arXiv: 2410.19115[cs]. URL: http://arxiv.org/abs/2410.19115 (visited on 12/12/2024).

[10] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94. URL: http://link.springer.com/10.1023/B:VISI.0000029664.99615.94 (visited on 12/12/2024).

[11] Jim Lawrence, Javier Bernal, and Christoph Witzgall. "A Purely Algebraic Justification of the Kabsch-Umeyama Algorithm". In: *Journal of Research of the National Institute of Standards and Technology* 124 (Oct. 9, 2019), p. 124028. ISSN: 2165-7254. DOI: 10.6028/jres.124.028. arXiv: 1902.03138[math]. URL: http://arxiv.org/abs/1902.03138 (visited on 12/12/2024).

[12] Lihe Yang et al. *Depth Anything V2*. arXiv.org. June 13, 2024. URL: https://arxiv.org/abs/2406.09414v1 (visited on 09/15/2024).

[13] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: ().

[14] Johannes L. Schonberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pp. 4104–4113. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.445. URL: http://ieeexplore.ieee.org/document/7780814/ (visited on 12/12/2024).

[15] Ben Mildenhall et al. *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*. Aug. 3, 2020. DOI: 10.48550/arXiv.2003.08934. arXiv: 2003.08934[cs]. URL: http://arxiv.org/abs/2003.08934 (visited on 12/12/2024).

[16] Yuehao Wang et al. *Neural Rendering for Stereo 3D Reconstruction of Deformable Tissues in Robotic Surgery*. June 30, 2022. DOI: 10.48550/arXiv.2206.15255. arXiv: 2206.15255[cs]. URL: http://arxiv.org/abs/2206.15255 (visited on 12/12/2024).

[17] Kailing Wang et al. *EndoGSLAM: Real-Time Dense Reconstruction and Tracking in Endoscopic Surgeries using Gaussian Splatting*. Mar. 22, 2024. DOI: 10.48550/arXiv.2403.15124. arXiv: 2403.15124[cs]. URL: http://arxiv.org/abs/2403.15124 (visited on 12/12/2024).

A Appendix

All supplemental videos and results can be accessed here.