3-D Reconstruction from consecutive endoscopic images using Gaussian Splatting

Hung-Le Minh[1], Duy-Van Truong[1], Huy-Xuan Manh[4], Viet-Hang Dao[2,3], Phuc-Binh Nguyen[2], Thanh-Tung Nguyen[2], and Hai-Vu[1]

[1] Hanoi University of Science and Technology (HUST), Hanoi, VietNam
[2] Institute of Gastroenterology and Hepatology (IGH), HaNoi, VietNam
[3] Hanoi Medical University (HMU), HaNoi, VietNam
[4] Koh Young Technology Vietnam, VietNam

Abstract. Recent advancements in 3D reconstruction helped endoscopy doctors analyze the patients' gastrointestinal surfaces and abnormality detections. In this work, we expand this development further with a reconstruction method based on both classic techniques like structure from motion and recent advanced techniques like neural radiation fields and Gaussian splatting with new Gaussian encoding-decoding modules. In addition, an unique dataset was collected with some videos from daily endoscopy examinations. This development helped us achieve better reconstruction results and lower training time compared to existing methods.

Keywords: Gaussian Splatting · 3D Recontruction · EndoScopy · Radiance Fields · Gastrointestinal Imaging

1 Introduction

Gastrointestinal diseases, including lesions in the esophagus, stomach, and duodenum appear quite popularly in digestive examinations. Several studies have shown that common lesions include gastritis (62.7%), gastric and duodenal ulcers (6.3%), reflux esophagus (41.3%), esophageal candidiasis (1.9%), and polyps (1.8%). Doctors usually use visual inspection with tools like tube camera or capsule camera to see the internal gastrointestinal surface, detecting abnormal like lesion and tumors. In these examinations, size is an important factor as it can be used to determine the stage of the tumors, thus give valuable information for doctors to provide treatments. However, the information in endoscopy data is usually presented as 2D RGB images that does not contribute much 3D information. Therefore, a computed-aided diagnostic or a CAD tool which is able to reconstruct 3D information from the video pipeline or an image sequences is needed.

3-D reconstruction from the consecutive image sequence is a conventional task in the computer vision field. Snavely et al.'s "Photo Tourism: Exploring Photo Collections in 3D" [11] introduced a ground-breaking approach to 3D

reconstruction from photo collections of multiple viewpoints, laying the foundation for Structure-from-Motion (SfM). Since then, SfM has been used widely in different domains, and improved, notably with work of Schonberger et al. [10] to improve the traditional SfM (Shape-from-motion) pipeline. The authors address challenges in feature matching, camera model estimation, and point triangulation, proposing several improvements that advance the state-of-the-art in SfM. The revised SfM pipeline demonstrates superior performance, particularly in large and complex datasets, and has become the foundational reference in 3D reconstruction. With the advancement of deep learning and big data-based techniques, 3D reconstruction in endoscopy is promised to overcome challenges which made SfM and 3D computer vision methods alike produce sub-optimal results like requiring multiple viewpoints, light reflectance, narrow operating channels and tissue deformation.

Some studies about 3D reconstruction in other domains have yielded promising results. For example, the study of utilizing Gaussian Splatting [4], but the use of 3D reconstruction in endoscopy still faces several challenges, such as the inability to provide multiple angles and the deformation of damaged surfaces over time as doctors perform endoscopy. Additionally, due to the presence of numerous water bubbles in the gastrointestinal tract, the damaged areas may be overly illuminated, presenting a significant challenge, along with the lack of computational power and a large dataset for 3D in endoscopy. Therefore, in this study, we propose a solution that requires low computational resources, and can be integrated with endoscopy systems to reconstruct lesion surfaces efficiently. The main contribution is a new initialization method to eliminate input noise before training and to encode temporal changes of the damaged surfaces method. Based on HexPlane [2], it can reduce training time and resources. The experimental results show that the training time and memory required to train and run the deep learning model is much lighter than the based-line 3D reconstruction methods.

The remaining of this paper is organized as follows: Section 2 briefly reviews recent works on 3-D reconstructions from the endoscopic images sequence. Section 3 describes the proposed method. Section 4 shows the 3-D reconstruction results from consecutive images and evaluates in term of both quantitative and qualitative indexes. Finally, the conclusions and future works are given in Section 5.

2 Related Work

Recent advancements in depth estimation for monocular endoscopy have leveraged deep learning to overcome the challenges of accurately reconstructing 3D structures from 2D images. The paper "Deep learning-based depth estimation from a synthetic endoscopy image training set" [7] of Nicholas J. Durr and his colleges explores the use of synthetic datasets to train deep networks, allowing for robust depth estimation even in complex endoscopic environments. In additional, facing the challenges lack of in-depth information that makes training

difficult, thus Xingtong Liu et al.'s "Dense Depth Estimation in Monocular Endoscopy With Self-Supervised Learning Methods" [6] introduces self-supervised approaches that eliminate the need for ground truth depth data, making the process more adaptable to real-world scenarios. Complementing these efforts, "A geometry-aware deep network for depth estimation in monocular endoscopy" [14] was introduced by Yongming Yang et al. integrates geometric constraints into the deep learning model, enhancing the accuracy of depth predictions by leveraging the inherent structure of the endoscopic scenes. One of the most successful studies on depth prediction presented by Beilei Cui et al is "EndoDAC: Efficient Adapting Foundation Model for Self-Supervised Depth Estimation from Any Endoscopic Camera." [3] This method addresses challenges in self-supervised depth estimation from endoscopic images by incorporating Dynamic Vector-Based Low-Rank Adaptation (DV-LoRA) and Convolutional Neck blocks. EndoDAC autonomously estimates camera intrinsics, enabling it to work with various surgical video datasets without explicit camera information. The results show that EndoDAC outperforms existing methods in accuracy and efficiency, with fewer training epochs and reduced computational demands.

A different approach than the studies mentioned above to be able to reconstruct 3D scenes. The study [8] by Mildenhall et al. (2020) introduced a groundbreaking method where Neural Radiance Fields (NeRF) use neural networks to model the radiance field of a scene. This approach learns a continuous function to predict the density and color of points in 3D space, enabling the generation of new 2D images from different viewpoints. NeRF significantly enhances image quality and detail compared to previous methods by leveraging information from multiple views to construct a more accurate 3D representation of the scene (Mildenhall et al., 2020). Building on this, the works in [1] by Barron et al. (2021) extends NeRF's capabilities by introducing a multiscale representation to handle scenes with varying levels of resolution. Mip-NeRF improves the quality of rendered images by applying different layers of resolution during the learning process, enhancing detail and reducing artifacts in high-resolution areas. In [9], Pumarola et al. (2021) introduces D-NeRF, an extension of the Neural Radiance Fields (NeRF) framework designed to handle dynamic scenes. Unlike traditional NeRF, which is limited to static scenes, D-NeRF incorporates temporal information to model and reconstruct 3D scenes with moving objects or changing environments. This method enhances NeRF's capability to accurately represent and synthesize images of dynamic scenes by integrating dynamic changes into the 3D model. With the results that the above research brings as well as new approaches, there are a number of studies that have applied them in the medical field. The paper [12] by Yuehao Wang et al. (2022) introduces an adaptation of the NeRF framework specifically for endoscopic imaging. EndoNeRF addresses challenges like distortion, low resolution, and uneven lighting by modifying NeRF to better handle the unique characteristics of endoscopic data. It enables the accurate modeling and reconstruction of complex anatomical structures by predicting radiance and density at specific points in the scene. This allows for the creation of detailed 3D models from 2D endoscopic images,

significantly improving reconstruction quality and offering potential benefits for medical applications such as diagnostics and surgical planning.

In recent years, the field of computer graphics has seen significant advancements in rendering techniques, particularly in the area of radiance field rendering. As researchers continue to push the boundaries of visual quality and computational efficiency, a new method has emerged that combines both speed and high fidelity. The paper [4] by Bernhard Kerbl et al. presents a novel approach for real-time radiance field rendering using a 3D Gaussian scene representation. This method achieves visual quality comparable to or surpassing previous state-of-the-art methods while requiring significantly shorter optimization times and delivering superior rendering speed. The approach leverages three key innovations: representing the scene with 3D Gaussians to avoid unnecessary computations in empty spaces, continuously optimizing the properties of the Gaussians including their anisotropic covariance, and developing a fast, visibility-aware rendering algorithm that supports real-time rendering at $\geq$ 30 fps at 1920x1080 resolution. The experimental results demonstrate that this method achieves state-of-the-art visual quality on various established datasets and provides the first real-time rendering solution for complex scenes with large depth complexity.

3 Proposed Method

The application of Structure from Motion (SfM) to generate input models for the Gaussian method has demonstrated significant improvements in both processing time and 3D reconstruction efficiency, as outlined in [4]. This approach enables the creation of three-dimensional structures from sequences of two-dimensional images, thereby enhancing the accuracy and effectiveness of the reconstruction process. However, when applied in the context of endoscopy, particularly in gastrointestinal endoscopy, several substantial challenges have emerged.

The environment within the gastrointestinal tract has unique characteristics, such as low light conditions and high levels of noise caused by unstable movements and interactions with soft tissues. Additionally, during endoscopic procedures, complex surgical instruments frequently enter the field of view, obstructing visibility and degrading the quality of the captured images. These factors increase the difficulty of accurately reconstructing 3D scenes from endoscopic images, leading to potential inaccuracies or gaps in the generated 3D models.

Recognizing these challenges, we have proposed several solutions to improve the reliability and efficiency of the 3D reconstruction process in endoscopic conditions. These solutions include optimizing image processing algorithms to minimize noise and enhance image recognition capabilities in low-light environments. Furthermore, we have focused on developing new approaches to better handle data processing when surgical instruments are present in the field of view. These advancements not only enhance the quality of the 3D models but also make the reconstruction process faster and more accurate, meeting the stringent demands of modern endoscopy.

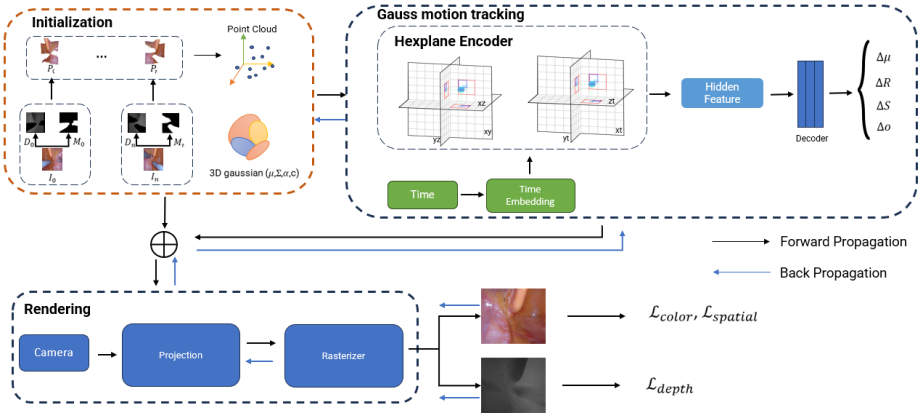

Fig. 1: The proposed method consists of three main modules. The initialization module helps to use mask image and depth image to convert 2D image to point cloud and remove noise. The motion encoding module helps to track the movement of tissues over time. The image rendering module helps to generate 2D depth image and 2D RGB color image after removing noise.

3.1 Initialization

The input data from SfM has several key limitations:

- **Sparse initialization points**: SfM often produces sparse initialization points, especially in areas with significant depth variation or rapid movement, leading to inaccurate models and requiring additional optimization to improve accuracy.
- **Training and rendering performance**: The sparsity of the initialization points can negatively impact the performance of model training and image rendering, particularly in complex scenes with significant depth variation.

To address the identified limitations, this study proposes a new initialization method designed to enhance the accuracy and efficiency of the 3D reconstruction process. Specifically, the system takes as input a sequence of images $\{I_i\}_{i=1}^{L}$ captured from a single camera, where L denotes the temporal length or number of frames in the sequence. In addition, based on the requirements set by the doctors to evaluate whether the input image contains noisy areas or not, we will then use the tool to create mask images $\{M_i\}_{i=1}^{L}$ to cover up the corresponding noisy areas. Instead of relying on traditional techniques such as SfM, we employ an advanced deep learning model [13] to predict depth maps $\{D_i\}_{i=1}^{L}$ for each image in the sequence.

This depth prediction process generates a depth map for each frame, representing the distance from points in the scene to the camera. With the obtained depth maps and the original input images, we combine them to reconstruct new images in the camera's space. This approach enables more accurate 3D space

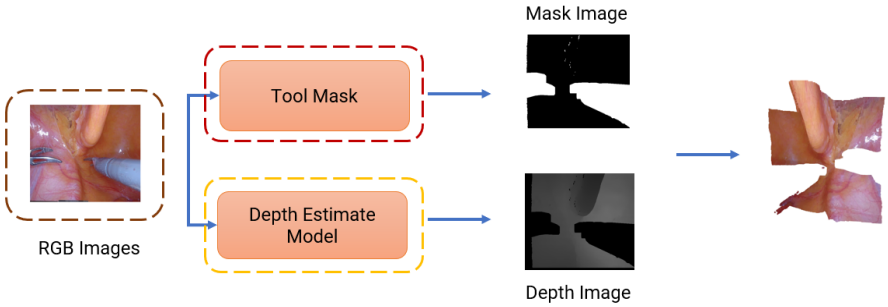

Fig. 2: The process of combining depth images and mask images to generate depth information and remove noise.

reconstruction, especially in complex environments such as endoscopy, where depth and shape variations are critical factors.

Based on the images obtained above, we propose a formula to initialize the point clouds.

$$P_i = K^{-1} T_i D_i (I_i \odot M_i) \tag{1}$$

where K, T_i, D_i, I_i, M_i are Extrinsic matrix, Intrinsic matrix, depth map, the image, the mask respectively

Since the images are captured from a single camera, they may be limited and some areas might be obstructed by medical equipment, bubbles, etc. Thus, this study will integrate all point clouds to create a comprehensive initialization. Additionally, we want to create a common space that contains all the variations of the point clouds over time. So that when we optimize them, we can learn the most common features between the variations of the point clouds over time. This helps us to optimize better when combined with time embedding.

$$P = \{P_1, P_2, P_3, \ldots, P_T\} \tag{2}$$

3.2 Gaussian motion Tracking

During surgery, tissues may move or change shape, requiring an accurate model to track and represent these changes. To address this issue, a deformation field D(μ, t) is needed to monitor the displacement of the attribute ΔG for each Gaussian distribution at time t. The deformed Gaussian distributions can be calculated as:

$$G_t = G_0 + \Delta G \tag{3}$$

However, there are several challenges in practical implementation when using large neural networks to approximate the deformation field D(μ, t). First, the use of large neural networks can lead to slow inference speeds, making the

computational process inefficient. Second, this model may not achieve optimal performance during training and inference, leading to subpar results in modeling surface dynamics.

Based on research [9], this study divides the deformation field into two lighter modules: D = F ∘ E. The E module is a disentangled voxel encoder that encodes the 4D inputs (the center of each Gaussian (x, y, z) and time t) into temporal latent features. The F module is a Gaussian transformation decoder that uses these latent features to compute Gaussian transformations. This approach improves inference speed and optimization while maintaining the accuracy of the model. The disentangled voxel encoder E is designed to effectively encode the 4D inputs.

Gaussian encoding module This module is inspired by the research [2], which uses a multi-resolution HexPlane structure to represent the 4D voxel encoding. This means that the voxel encoding is split into six planes with corresponding vectors. The module E(µ, t) encodes the center of each Gaussian µand the time t into temporal latent features, making the modeling of the temporal transformation of Gaussian distributions more efficient and accurate. After calculating the feature values of each plane and combining them, a latent feature is obtained. This process results in a complete and accurate latent feature from the 4D input coordinates. The combination formula is detailed below:

$$E = E_{XY} \otimes E_{ZT} \otimes v_1 + E_{XY} \otimes E_{ZT} \otimes v_2 + E_{XY} \otimes E_{ZT} \otimes v_3 \tag{4}$$

where $\otimes$ represents matrix multiplication, and $\mathbf{E}_{AB} \in \mathbb{R}^{A \times B}$ is a feature plane that has been learned from the training data. This matrix contains learned feature values and is used to represent important information of the data in the feature space $\mathbb{R}^{A \times B}$. This plane helps reduce the complexity of the data and focus on the most significant factors. $\mathbf{v}_i \in \mathbb{R}^D$ denotes the feature vector along the i-th axis. This vector contains feature values for a specific dimension in the input space. By using feature vectors along the axes, it is possible to analyze and represent data along different dimensions, allowing the model to capture the structure and characteristics of the data comprehensively.

Gaussian decoding module This module plays a crucial role in adjusting and reconstructing the Gaussians over time. Designed to handle the variations of the Gaussian parameters from latent features, this decoder uses small multilayer perceptrons (MLPs) to predict changes in position, rotation, scale, and opacity of each Gaussian. Specifically, this module is applied to predict changes in position ($\Delta\boldsymbol{\mu}$), changes in rotation ($\Delta\mathbf{R}$), changes in scale ($\Delta\mathbf{S}$), and changes in opacity ($\Delta\mathbf{o}$). Each MLP aims to optimize the reconstruction of the Gaussians after they have been transformed, ensuring that the model can accurately reflect the dynamic changes in the surface over time. Therefore, the transformation of the Gaussian at time t is given by the following formula:

$$G_t = G_t + \Delta G = (\boldsymbol{\mu} + \Delta\boldsymbol{\mu}, \mathbf{R} + \Delta\mathbf{R}, \mathbf{S} + \Delta\mathbf{S}, \mathbf{o} + \Delta\mathbf{o}, \mathbf{SH}) \tag{5}$$

3.3 Optimization

Color Loss In the process of scene reconstruction, ensuring that the colors of the reconstructed scene match the actual colors is crucial. The color loss function helps minimize the difference between the predicted colors and the actual colors, ensuring that the reconstructed image is realistic and accurate.

$$\mathcal{L}_{\text{color}} = \sum_{x \in \zeta} \left\| M(x) \left(\hat{C}(x) - C(x) \right) \right\|_1 \tag{6}$$

where M, $\hat{C}$, C and ζ are binary tool masks, predicted colors, real colors and 2D coordinate space.

Depth Loss The depth loss function helps the model learn to accurately represent the distances between objects in the scene. This is essential for generating images with greater depth and accuracy. Additionally, this loss function utilizes soft constraints (sort constraints), which enable the model to focus on aligning spatial structures (such as shapes and relative distances between objects) rather than the absolute depth values.

$$\mathcal{L}_{\text{depth}} = 1 - \frac{\text{Cov}(M \odot \hat{D}, M \odot D)}{\sqrt{\text{Var}(M \odot \hat{D}) \, \text{Var}(M \odot D)}} \tag{7}$$

where M, $\hat{D}$, D and are binary tool masks, predicted depths, real depths.

Spatio-temporal Loss To ensure the spatial and temporal smoothness of the reconstructed result, total variation (TV) functions are applied. This helps to avoid black/white regions in areas obscured by surgical instruments.

$$\mathcal{L}_{\text{spatial}} = \text{TV}(\hat{C}) + \text{TV}(\hat{D}^{-1}) \tag{8}$$

Total Loss With the 3 loss functions mentioned above. They will be combined and create a synthetic loss function to ensure the quality of the image produced with the objectives each loss function brings.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{color}} + \lambda_2 \mathcal{L}_{\text{depth}} + \lambda_3 \mathcal{L}_{\text{spatial}} \tag{9}$$

Where, the weights λ_1, λ_2, λ_3 play an important role in balancing the different loss functions. These weights help to adjust the influence of each loss function on the overall optimization of the model. But to get stable results, the color loss function will be focused on the most and the remaining two loss functions we test the settings to get the best results 4.

4 Experimental Results

4.1 Dataset and Evaluation Metrics

We performed experiments using two public datasets, including EndoNeRF [12] and another dataset. ENDONERF features two instances of in-vivo prostate surgery data recorded from a stereo camera at a single viewpoint, showcasing complex scenes with non-rigid deformations and instrument occlusions. In accordance with prior research [27], we divided the frame data from each scene into training and testing sets with a 7:1 ratio. Our method was assessed using several metrics such as PSNR, SSIM, and LPIPS. Additionally, we documented the training duration, inference speed (FPS, frames per second), and GPU memory consumption during training.

In addition, we collected three endoscopic, single viewpoint videos from IGH. On average, videos have a total of 430 frames, but we only select several frames according to the criteria related to lighting and angles to serve the reconstruction of the surface. The data was also divided into 8 parts, with 7 parts for training and 1 part for testing.

Name Dataset	Total Frame
ENDONERF-cutting	155
ENDONERF-pulling	63
IGH dataset	430

Table 1: Overview of the datasets

4.2 Implementation details

All our experiments are performed on a computer with an Intel core i7 12700k configuration and with a single RTX 3080 graphics card. Using the Adam [5] optimization function. The initial learning rate is initialized to 1.6×10^{-3}. We will perform the point cloud thickening without updating among the models for about 1000 iterations And the training process is performed for 6000 iterations

4.3 Results

We compared our proposed method with previous studies on surface structure reconstruction, including EndoNeRF and EndoSurf. The results are shown in Table 2 and Table 3 where we can see that the performance of EndoNeRF and EndoSurf achieved excellent results for temporally varying data but required substantial computational resources and lengthy training times. Our results show that our method achieved average result is 37.4895 (PSNR) with only 10 minutes of training on the EndoNeRF dataset, significantly faster than previous studies, and yielded improved outcomes. We have illustrated several scenes in EndoNeRF

dataset which are reconstructed using the proposed method in Fig. 4 and Fig. 3. In both cases, they are reconstructed successfully even with the obstruction of some objects such as the endoscope tube and the cutting-device appeared in the endoscopy examination. In Fig. 5, we examine the proposed method to a practical image sequence. The 2-D original image sequence shows a stomach cancer region. The reconstructed images help examining doctor observe the abnormal region from different view-point (as given in 3-D visualization at following link http://surl.li/wedpmm) as well as clearly observe the shape of this abnormality. These results suggest that the proposed method has the potential to be applied in preoperative lesion assessment.

Method	EndoNeRF [12]			EndoSurf [15]			**Ours**		
Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ENDONERF-cutting	34.186	0.932	0.151	34.981	0.953	0.106	37.849	0.963	0.089
ENDONERF-pulling	34.212	0.938	0.161	35.004	0.956	0.120	37.13	0.930	0.004
Average	34.199	0.935	0.156	34.4925	0.9545	0.113	37.4895	0.9465	0.0465

Table 2: Quantitative metrics of appearance (PSNR/SSIM/LPIPS)

Method	EndoNeRF [12]		EndoSurf [15]		**Ours**	
Metric	GPU ↓	Time Training ↓	GPU ↓	Time Training ↓	GPU ↓	Time Training ↓
ENDONERF-cutting	19GB	7 hours	19GB	9 hours	2GB	10 mins
ENDONERF-pulling	19GB	7 hours	19GB	9 hours	2GB	10 mins

Table 3: Comparison of GPU usage and training time across different methods.

λ_1	λ_2	λ_3	PSNR ↑
0.9	0.01	0.09	37.27
0.9	0.02	0.08	37.23
0.9	0.03	0.07	36.94
0.9	0.04	0.06	37.22
0.8	0.01	0.19	36.24
0.8	0.02	0.18	36.07
0.8	0.03	0.17	36.56
0.8	0.04	0.16	26.49

Table 4: Weights for Loss Functions

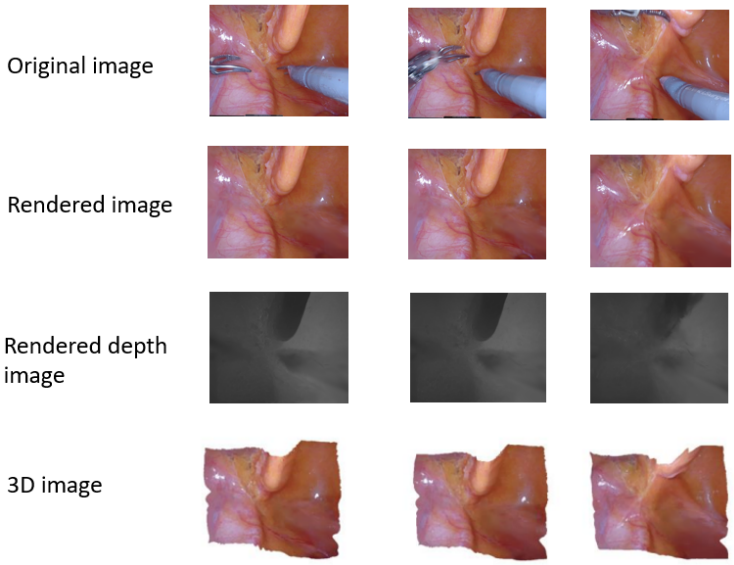

Fig. 3: Rendered images, depth images and 3D images are reconstructed from the original images from the dataset ENDONERF-pulling

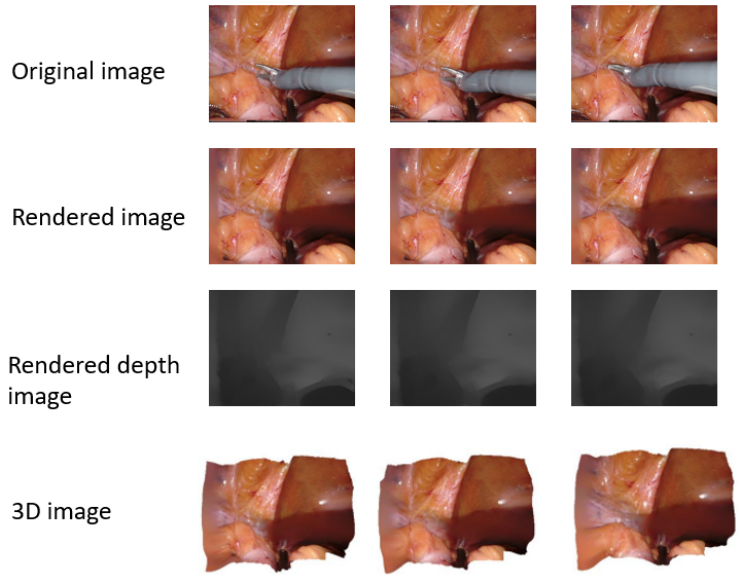

Fig. 4: Rendered images, depth images and 3D images are reconstructed from the original images from the dataset ENDONERF-cutting

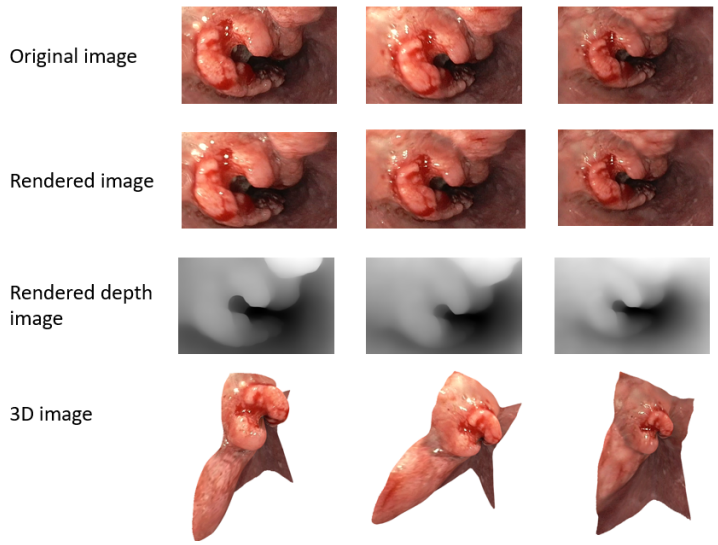

Fig. 5: Rendered images, depth images and 3D images are reconstructed from the original images from the IGH dataset (Please refer the reconstruction results in different views at following link http://surl.li/wedpmm)

5 Conclusion

In this paper, we present a real-time, high-quality framework for reconstructing dynamic surgical scenes. By leveraging Endo-Gaussian Initialization and Spatio-Temporal Gaussian Tracking, we effectively address the challenges of Gaussian initialization and tissue deformation. Extensive experiments demonstrate that our EndoGaussian achieves state-of-the-art reconstruction quality while significantly improving rendering speed. We believe that emerging Gaussian Splatting-based reconstruction techniques can open new pathways for better understanding robotic surgical scenes and support various clinical tasks, particularly in intra-operative applications.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5855–5864 (October 2021) 3
2. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 130–141 (June 2023) 2, 7
3. Cui, B., Islam, M., Bai, L., Wang, A.C., Ren, H.: Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera.

ArXiv **abs/2405.08672** (2024), https://api.semanticscholar.org/CorpusID:269761398 3
4. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering (2023), https://arxiv.org/abs/2308.04079 2, 4
5. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
6. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Transactions on Medical Imaging **39**(5), 1438–1447 (2020). https://doi.org/10.1109/TMI.2019.2950936 3
7. Mahmood, F., Durr, N.J.: Deep learning-based depth estimation from a synthetic endoscopy image training set. In: Angelini, E.D., Landman, B.A. (eds.) Medical Imaging 2018: Image Processing. vol. 10574, p. 1057421. International Society for Optics and Photonics, SPIE (2018). https://doi.org/10.1117/12.2293785, https://doi.org/10.1117/12.2293785 2
8. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (dec 2021). https://doi.org/10.1145/3503250, https://doi.org/10.1145/3503250 3
9. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10318–10327 (June 2021) 3, 7
10. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) 2
11. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (jul 2006). https://doi.org/10.1145/1141911.1141964, https://doi.org/10.1145/1141911.1141964 1
12. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues inrobotic surgery. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 431–441. Springer Nature Switzerland, Cham (2022) 3, 9, 10
13. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10371–10381 (June 2024) 5
14. Yang, Y., Shao, S., Yang, T., Wang, P., Yang, Z., Wu, C., Liu, H.: A geometry-aware deep network for depth estimation in monocular endoscopy. Engineering Applications of Artificial Intelligence **122**, 105989 (2023). https://doi.org/https://doi.org/10.1016/j.engappai.2023.105989, https://www.sciencedirect.com/science/article/pii/S0952197623001732 3
15. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos (2023), https://arxiv.org/abs/2307.11307 10