SurgicalGaussian: Deformable 3D Gaussians for High-Fidelity Surgical Scene Reconstruction

Weixing Xie[1,2], Junfeng Yao[1,2,3](✉), Xianpeng Cao[1], Qiqin Lin[1]
Zerui Tang[1,2], Xiao Dong[4] (✉), and Xiaohu Guo[5]

[1] Center for Digital Media Computing, School of Film, School of Informatics, Xiamen University, Xiamen, China
yao0010@xmu.edu.cn
[2] National Institute for Data Science in Health and Medicine, Xiamen University
[3] Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism
[4] Guangdong Provincial Key Laboratory IRADS and Department of Computer Science, BNU-HKBU United International College, Zhuhai, China
[5] Department of Computer Science, The University of Texas at Dallas, Dallas, USA

Abstract. Dynamic reconstruction of deformable tissues in endoscopic video is a key technology for robot-assisted surgery. Recent reconstruction methods based on neural radiance fields (NeRFs) have achieved remarkable results in the reconstruction of surgical scenes. However, based on implicit representation, NeRFs struggle to capture the intricate details of objects in the scene and cannot achieve real-time rendering. In addition, restricted single view perception and occluded instruments also propose special challenges in surgical scene reconstruction. To address these issues, we develop SurgicalGaussian, a deformable 3D Gaussian Splatting method to model dynamic surgical scenes. Our approach models the spatio-temporal features of soft tissues at each time stamp via a forward-mapping deformation MLP and regularization to constrain local 3D Gaussians to comply with consistent movement. With the depth initialization strategy and tool mask-guided training, our method can remove surgical instruments and reconstruct high-fidelity surgical scenes. Through experiments on various surgical videos, our network outperforms existing method on many aspects, including rendering quality, rendering speed and GPU usage. The project page can be found at https://surgicalgaussian.github.io.

Keywords: 3D Reconstruction · Gaussian Splatting · Minimally Invasive Surgery.

1 Introduction

In robotic-assisted minimally invasive surgery, reconstructing the surgical scene from endoscopic videos is a critical and challenging task. The reconstruction of surgical scenes can not only help doctors operate instruments more accurately, but also is the basis for a series of downstream clinical applications, such as surgical environment simulation [4,19], robotic surgery automation [15], and medical

teaching [21]. However, the sparse viewpoints, limited movement space, topologically changing tissues and instrument occlusion in endoscopic surgery pose key challenges to dynamic reconstruction. Despite much progress recent years, reconstruction of surgical scenes by existing methods still lose intricate details.

Previous work proposed explicit discrete representations of surgical scenes, such as point clouds [11,23,30] and surfels [14,17]. These methods usually compensate tissue deformation by sparse warp fields, limiting their ability to handle drastic motions and color alteration due to topology changes. With the development of neural radiance fields (NeRF) [6,18], continuous representations of dynamic scene demonstrated superiority in generating high-quality appearance and geometry. Specifically, EndoNeRF [25] improved dynamic NeRF framework [20] to reconstruct stationary monocular endoscopic scenes with depth supervision. EndoSurf [29] focused on surface reconstruction by employing SDF fields [1] and radiance fields to model surface dynamics and appearance. In addition, LerPlane [27] draws on the design of feature planes [5] to efficiently encode space-temporal features of sampling points, significantly reducing the workload of dynamic tissue modeling. However, implicit representation of NeRFs require dense sampling on millions of rays to adapt the implicit function to the surgical scene. This consumes huge computational resources, especially in surgical scenarios with complex motion and high resolution, even with accelerated NeRF versions [10,27]. It is difficult for current work to achieve high-quality reconstruction and real-time rendering of surgical scenes at the same time.

Recently, 3D Gaussian Splatting (3DGS) [8] has emerged as a viable alternative 3D representation to NeRF as it yields realistic rendering while being significantly faster to train than NeRFs. Specifically, 3DGS represents the scene as anisotropic 3D Gaussians and adopts differentiable rasterization pipeline to render images. In this paper, we propose a deformable 3DGS framework tailored for endoscopic videos to reconstruct dynamic surgical scene and remove occluded instruments. To model motion fields, some methods [13,26,27] employ planar structures for feature encoding efficiency. Although decomposing 3D scene into feature planes can speed up and improve the reconstruction quality, these low-rank planar components are not the best choice for encoding complex motion fields. Dynamic scenes possess a higher rank compared to static scenes, and explicit point-based rendering further elevates the rank of the scene [2,28]. In our method, given 3D Gaussians in canonical space, we utilize a deformation network to decouple the motion and geometry of surgical scene, predicting flexible Gaussian motion in observation space.

Compared with existing methods for stereo 3D reconstruction in robotic surgery, our contributions are summarized as follows: 1) We propose an deformable 3D Gaussians framework (SurgicalGaussian) for high-fidelity surgical scene reconstruction in endoscopic video; 2) we propose an efficient Gaussian initialization strategy (GIDM) to use geometry prior based on depth and mask to reduce the motion-appearance ambiguity in single viewpoint; 3) we address the color prediction of occluded areas and the noise of Gaussian deformation fields by using color and deformation regularization respectively; 4) our method demon-

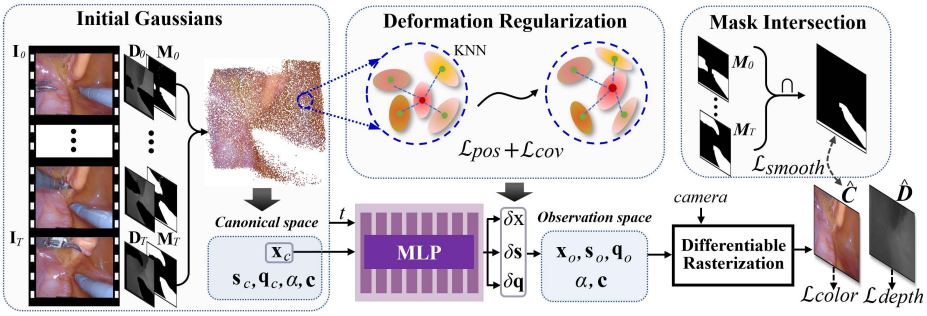

Fig. 1. Framework of the proposed SurgicalGaussian.

strates compelling reconstruct quality with real-time rendering speed, preserving high-frequency details of tissues while removing surgical tools.

2 Method

The architecture of our network SurgicalGaussian is shown in Fig. 1. We take an endoscopic video $V = \{\mathbf{I}_i, \mathbf{D}_i, \mathbf{M}_i : i \in [0, T]\}$ as input, where $\mathbf{I}_i$ is the i-th frame image, $\mathbf{D}_i$ is the depth map and $\mathbf{M}_i$ (1 for tool pixels) is the mask of the surgical tools. The time t of frame i is normalized to $i/T \in [0, 1]$. Given the above inputs, our network builds a deformable 3D Gaussian representation of a surgical scene that can remove surgical instruments and restore deformed soft tissue with high quality.

2.1 Preliminaries

We build dynamic scene representation based on 3DGS [8], which consists of a set of Gaussian primitives $\{\mathcal{G}\}$. The distribution of Gaussian in world space is defined by its center location $\mathbf{x}$ and the covariance matrix $\mathbf{\Sigma}$, denoted as $G(x) = exp(-1/2x^T\mathbf{\Sigma}^{-1}x)$. To ensure positive semi-definite property of Gaussians, the covariance is decomposed as $\mathbf{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$, where $\mathbf{S}$ is a scaling matrix and $\mathbf{R}$ is a rotation matrix. In practice, we store the diagonal vector $\mathbf{s} \in \mathbb{R}^3$ of the scaling matrix and the quaternion vector $\mathbf{q} \in \mathbb{R}^4$ of the rotation matrix. Each Gaussian has opacity α and spherical harmonics coefficients for color. We then denote a 3D Gaussian with its properties as $\mathcal{G} = \{(\mathbf{x}, \mathbf{s}, \mathbf{q}, \alpha, \mathbf{c})\}$.

Given a viewing transformation $\mathbf{V}$ and the Jacobian of the affine approximation of the projective transformation $\mathbf{J}$, we can project 3D Gaussians to 2D image plane for rendering. The 2D covariance matrix in camera coordinates is calculated as follows: $\mathbf{\Sigma}' = \mathbf{JV\Sigma V}^T\mathbf{J}^T$. The estimated color $\hat{C}(\mathbf{r})$ and depth $\hat{D}(\mathbf{r})$ of pixel $\mathbf{r}$ can be rendered by blending Gaussians sorted with their depth:

$$\hat{C}(\mathbf{r}) = \sum_i (\alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j))c_i, \hat{D}(\mathbf{r}) = \sum_i (\alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j))d_i. \quad (1)$$

Here c_i is the color of 3D Gaussian, $\alpha_i' = \alpha_i \mathcal{G}_i'$ is opacity α_i weighted by the probability density of projected 2D Gaussian $\mathcal{G}_i'$ at target pixel location. We denote depth of $\mathcal{G}_i$ as d_i and predict pixel depth in a similar way.

2.2 GIDM initialization strategy

3DGS [8] uses SfM [22] point cloud as the initial position of the 3D Gaussians. Compared with random initialization, this strategy significantly improves the rendering quality of areas not well covered by the training view. However for endoscopic videos, due to limited observation viewpoints, sparse textures of soft tissues and dynamic lighting condition, it is difficult to obtain accurate SfM point clouds, and therefore cannot provide accurate initialization for 3D Gaussians.

In our implementation, we propose an efficient Gaussian initialization scheme using depth maps of the surgical scene. Given the depth $\mathbf{D}$ and mask $\mathbf{M}$ of T frames in surgical video, we project image pixels to world coordinates. Specifically, we first project the tissue pixels on frame 0 into 3D space to obtain the point cloud $\mathbf{P}_0$. This point cloud has large missing due to the removal of surgical tools. However, the tissue occluded by the instrument at frame i may be visible in other frames, so Gaussian points should also be placed in this area for reconstruction of dynamic part. Based on this observation, we check masks in all frames and collect new tissue pixels appear in frame $i+1$ while occluded in previous i frames. We add the collected pixels from other frames to frame 0 to obtain the refined image $\mathbf{I}^*$, depth map $\mathbf{D}^*$ and mask $\mathbf{M}^*$. Based on the projection, we obtain the refined point cloud $\mathbf{P}^*$. There are still holes in $\mathbf{P}^*$ if a small region is occluded the whole time.

$$\mathbf{P}^* = \{\mathbf{D}^* \mathbf{K}_e^{-1} \mathbf{K}_i^{-1} (\mathbf{I}^* \odot (\mathbf{1} - \mathbf{M}^*))\}, \mathbf{M}^* = \bigcap_{i=0}^{T} \mathbf{M}_i. \tag{2}$$

Here, $\mathbf{K}_i$ and $\mathbf{K}_e$ are camera intrinsic matrix and extrinsic matrix respectively. We show point cloud $\mathbf{P}^*$ in Fig. 1, which is used to initialize the position $\mathbf{x}_c$ and color $\mathbf{c}$ of Gaussians $\{\mathcal{G}_c\} = \{(\mathbf{x}_c, \mathbf{s}_c, \mathbf{q}_c, \alpha, \mathbf{c})\}$ in canonical space.

2.3 Deformable 3D Gaussian Representation

We utilize the strong ability of 3D Gaussians on rendering to get high-fidelity reconstruction of surgical scenes. To model dynamic 3D Gaussians that vary over time, we decouple the 3D Gaussians and the deformation field. The Gaussians in canonical space represent geometric prior of the scene, and a deformation network models the changes in position and shape of the Gaussians. This Gaussian point-based representation is flexible in capturing high-rank motion of objects.

The key to our Gaussian deformation modeling is a pure MLP. The network first encodes Gaussian position in the canonical space and time t of current frame and takes them as the input of the MLP network. The deformation network $\mathbf{F}_\Theta$ will learn the offset of each Gaussian's properties in the observation space, such

as position $\delta\mathbf{x}$, scaling $\delta\mathbf{s}$ and rotation $\delta\mathbf{q}$ to encode the motion of the scene. $\mathbf{F}_\Theta$ is a deep MLP with multiple layers and $\gamma(\cdot)$ is the function for Gaussian positional encoding and time encoding using specific frequency [18]. The offset of Gaussian position, scaling and rotation properties are obtained as following:

$$(\delta\mathbf{x}, \delta\mathbf{s}, \delta\mathbf{q}) = \mathbf{F}_\Theta((\gamma(\mathbf{x}_c), \gamma(t)). \tag{3}$$

Then the properties of Gaussians $\{\mathcal{G}_o\} = \{(\mathbf{x}_o, \mathbf{s}_o, \mathbf{q}_o, \alpha, \mathbf{c})\}$ in observation space are obtained:

$$\mathbf{x}_o = \mathbf{x}_c + \delta\mathbf{x}, \mathbf{s}_o = \mathbf{s}_c \cdot \exp(\delta\mathbf{s}), \mathbf{q}_o = \mathbf{q}_c \cdot \delta\mathbf{q}. \tag{4}$$

Note that applying the $\cdot$ operation to quaternion vectors is equivalent to multiplying the corresponding rotation matrices. The deformation network $\mathbf{F}_\Theta$ does not encode α and $\mathbf{c}$ because they are intrinsic properties of Gaussians that do not change with motion.

2.4 Optimization

Under the supervision of reconstruction losses and regularization terms, our network jointly optimizes canonical Gaussians $\mathcal{G}_c$ and parameters of deformation network. Since the surgical instruments in the video are to be removed, we invert the masks to select soft tissues and apply the reconstruction loss to this area. $\hat{\mathbf{C}}_i$ and $\hat{\mathbf{D}}_i$ are rendered image and depth for i-th frame, respectively.

$$\mathcal{L}_{\text{color}} = \left\|(\mathbf{I}_i - \hat{\mathbf{C}}_i)(\mathbf{1} - \mathbf{M}_i)\right\|_1, \mathcal{L}_{\text{depth}} = \left\|(\mathbf{D}_i - \hat{\mathbf{D}}_i)(\mathbf{1} - \mathbf{M}_i)\right\|_1. \tag{5}$$

Deformation regularization. When dealing with a single-view input, the deformation network can be highly under-constrained, leading to noisy deformations from the canonical space to the observation space. To address this issue, we propose a regularization method that ensures nearby Gaussians have similar deformation [16,24]. Specifically, we employ deformation consistency loss to the Gaussian position $\mathbf{x}$ and covariance matrix $\mathbf{\Sigma}$. For a Gaussian $\mathcal{G}_i$, we collect its K (K is set to 5) nearest neighbor Gaussians, calculate Euclidean distance between $\mathcal{G}_i$ and its neighbors in both canonical space and observation space, and constrain the difference after deformation to ensure consistent movement.

$$\mathcal{L}_{pos} = \sum_{i=1}^{N}\sum_{k=1}^{K}\left\|d\left(\mathbf{x}_c^{(i)}, \mathbf{x}_c^{(k)}\right) - d\left(\mathbf{x}_o^{(i)}, \mathbf{x}_o^{(k)}\right)\right\|_1, \tag{6}$$

$$\mathcal{L}_{cov} = \sum_{i=1}^{N}\sum_{k=1}^{K}\left\|d\left(\mathbf{\Sigma}_c^{(i)}, \mathbf{\Sigma}_c^{(k)}\right) - d\left(\mathbf{\Sigma}_o^{(i)}, \mathbf{\Sigma}_o^{(k)}\right)\right\|_1. \tag{7}$$

Occlusion-based color regularization. For the occlusion of surgical instruments, scene representation based on NeRFs can naturally complement the

color of the occluded area benefit from the smoothness of MLP [25]. The representation based on 3DGS is discrete and cannot handle rendering on occluded regions well. We observed that occlusions caused by surgical instruments fall into two types, one visible in other frames and the other invisible in all frames. The mask $\mathbf{M}^*$ discussed in Sec. 2.2 represents the occluded area that has never been observed. Notice that a Gaussian located in this region will hardly be optimized at any time, thus generating holes in rendered images. In order to enable the Gaussians in occluded regions to learn color similar to those of nearby Gaussians, we introduce a total variational loss [5], which helps generate reasonable color in occluded regions. The color regularization loss is defined as:

$$\mathcal{L}_{smooth} = \frac{1}{n}\sum_{p,q}\left(\|\mathbf{C}^{p,q} - \mathbf{C}^{p-1,q}\|_2^2 + \|\mathbf{C}^{p,q} - \mathbf{C}^{p,q-1}\|_2^2\right), \mathbf{C} = \hat{\mathbf{C}}_i \odot \mathbf{M}^*, \quad (8)$$

where p and q are indices of pixel, and n denotes the number of pixels in $\mathbf{C}$. The selection of intersection mask $\mathbf{M}^*$ instead of mask of each frame encourages Gaussians to use the ground truth appearance to recover temporarily occluded areas, thus preventing tissue color in that area from being over smoothed.

Total loss. We combine reconstruction loss and regularization loss to optimize dynamic 3D Gaussian representation. Follow 3DGS [8], we add SSIM loss to ensure structural similarity of rendered image to ground-truth image. We use five parameters to balance the relative importance of different terms. The final optimization objective can be represented as follows:

$$\mathcal{L} = (\mathcal{L}_{color} + \lambda_1\mathcal{L}_{ssim} + \lambda_2\mathcal{L}_{depth}) + (\lambda_3\mathcal{L}_{pos} + \lambda_4\mathcal{L}_{cov} + \lambda_5\mathcal{L}_{smooth}). \quad (9)$$

3 Experiments

3.1 Experimental Settings

Datasets and evaluation metrics. The EndoNeRF dataset[25] contains two prostatectomy cases shot by a binocular camera, with endoscopic images, depth maps, and mask maps of surgical tools. We select two sequences with rapid tissue motion from StereoMIS [7] dataset to evaluate reconstruction ability of existing methods. We estimate depth maps using a pre-trained STTR-light [12] model, and generate tool masks based on SAM-Track [3] model. We adopt three metrics to measure the quality of surgical scene reconstruction, including PSNR, SSIM, and LPIPS. These metrics are calculated by comparing predicted color and ground-truth color excluding surgical instruments denoted by masks. We also report GPU memory usage for training and rendering speed FPS.

Implementation details. The deformation MLP $\mathbf{F}_\Theta$ has 8 hidden layers and a width of 256. We randomly select one frame for each training iteration, and train all scenes with $40k$ iterations. The loss weights in Eq.(9) are empirically set as $\lambda_1 = 0.2$, $\lambda_2 = 0.001$, $\lambda_3 = 1$, $\lambda_4 = 200$ and $\lambda_5 = 0.02$. Adam [9] optimizer is adopted for training, the initial learning rate of the deformation network is set to 1.5×10^{-5}, and the Gaussian optimization parameter settings follow 3DGS [8]. We conduct all experiments on a computer with an RTX 3090 graphics card.

Table 1. Quantitative comparisons of our method with EndoNeRF [25], EndoSurf [29], LerPlane [27] and EndoGaussian [13]. We evaluate reconstruction quality on PSNR↑, SSIM↑, LPIPS↓ together with rendering speed FPS↑ and GPU Usage↓.

Dataset	Methods	"pulling"			"cutting"			FPS	GPU
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS		
EndoNeRF	EndoNeRF	34.217	0.938	0.160	34.186	0.932	0.151	0.04	15 GB
	EndoSurf	35.004	0.956	0.120	34.981	0.953	0.106	0.05	17 GB
	LerPlane	36.241	0.950	0.102	35.580	0.955	0.101	1.02	20 GB
	EndoGaussian	37.308	0.958	0.070	**38.287**	**0.962**	**0.058**	**190**	**2 GB**
	Ours	**38.783**	**0.970**	**0.049**	37.505	0.961	0.062	80	4 GB
StereoMIS		"intestine"			"liver"				
	EndoNeRF	28.694	0.783	0.279	27.738	0.712	0.345	0.06	13 GB
	EndoSurf	29.660	0.853	0.204	28.941	0.820	0.248	0.08	14 GB
	LerPlane	29.441	0.822	0.206	28.852	0.793	0.254	1.45	19 GB
	EndoGaussian	29.024	0.805	0.213	26.174	0.728	0.295	**200**	**2 GB**
	Ours	**31.496**	**0.890**	**0.145**	**31.668**	**0.893**	**0.135**	140	3 GB

Table 2. Ablation study on EndoNeRF [25] dataset.

Model	"pulling"			"cutting"		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o GIDM initialization	37.771	0.962	0.091	36.428	0.944	0.089
w/o $\mathcal{L}_{pos}$,$\mathcal{L}_{cov}$	37.944	0.963	0.094	36.633	0.951	0.090
w/o $\mathcal{L}_{smooth}$	38.609	0.968	0.070	37.081	0.957	0.065
Full model	**38.783**	**0.970**	**0.049**	**37.505**	**0.961**	**0.062**

3.2 Comparisons & Ablations

Comparison experiments. We conduct comparison experiments of SurgicalGaussian and other reconstruction methods on two datasets. Dynamic NeRFs, such as EndoNeRF [25], LerPlane [27] and EndoSurf [29], are based on an inverse mapping from observation space to canonical space when model the motion field, which is unable to achieve high-quality decoupling of canonical space and deformation field [28]. In addition, EndoNeRF uses MLP to encode both motion and appearance of the scene, fails to capture high-frequency details. LerPlane fixes the position of the sampling points on the grid nodes for encoding, and the resulting neural features cannot adapt well to high-rank complex signals. EndoSurf employs smoothness constraint on reconstructed dynamic surfaces, thus compromising rendering quality. We also compare our network with a similar 3DGS-based reconstruction method EndoGaussian [13], which strives for efficient training speed and real-time rendering. However, EndoGaussian faces the same problem as LerPlane, using feature plane, the low-rank tensors, to encode Gaussian deformation fields, resulting in impaired reconstruction of objects moving rapidly. In Table 1 and Fig. 2, we give quantitative evaluations and reconstructed point clouds of these methods. Our method can capture the intricate details of objects and produce highly realistic rendering results.

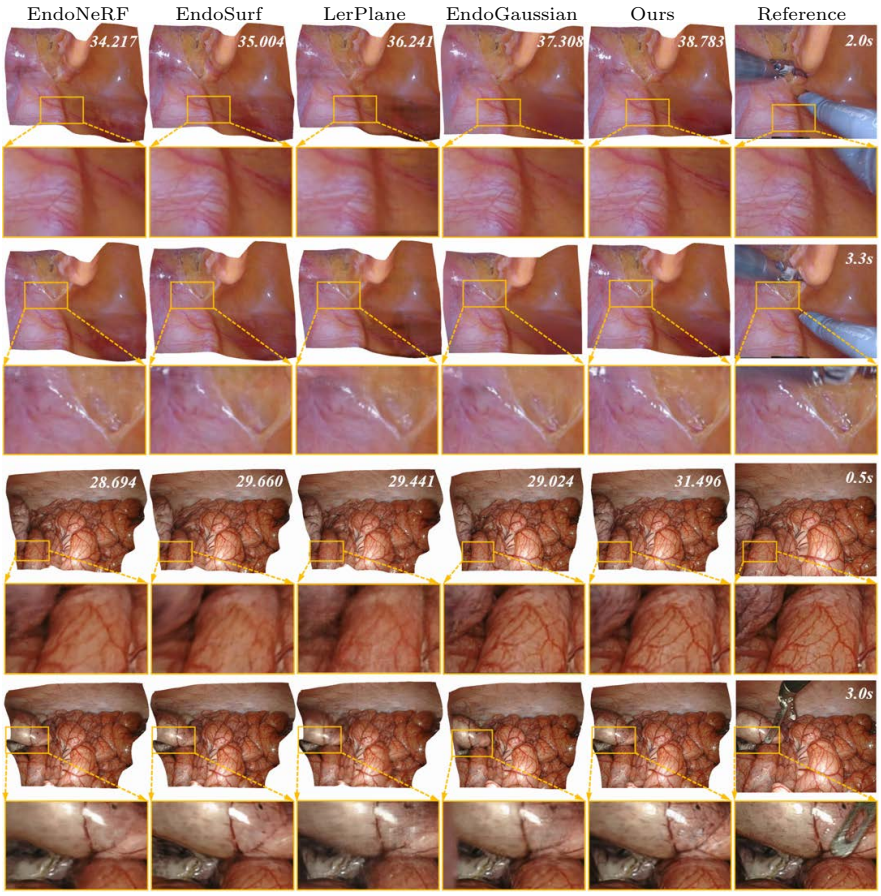

Fig. 2. Comparison of reconstruction results between our SurgicalGaussian and EndoNeRF [25], EndoSurf [29], LerPlane [27] and EndoGaussian [13].

Rendering efficiency. Our method achieves real-time rendering. Although the introduction of deformation MLP may increase the rendering overhead, we are still able to render in real-time owing to the extremely efficient CUDA implementation of 3D Gaussian splatting and our compact MLP structure. The average rendering speed on RTX3090 is $\geq$ 80FPS and GPU usage is $\leq$ 4GB.

Ablations. In Table 2 we conduct ablation analysis on the effectiveness of different modules of our network. We compare reconstruction quality based on random and GIDM initialization. GIDM helps restore the scene geometry thus yielding better quality. The $\mathcal{L}_{pos}$ and $\mathcal{L}_{cov}$ losses smooth tissue surface deformations by reducing overly noisy Gaussian motion fields. Although $\mathcal{L}_{smooth}$ loss slightly improves the quality of the scene, it plays an important role in restoring the color of occluded areas by surgical tools.

4 Conclusion

We present a novel deformable 3D Gaussian Splatting framework, SurgicalGaussian, specifically designed for high-fidelity reconstruction of dynamic surgical scenes. The 3D Gaussian-based representation in canonical space captures intricate textures of the tissue, while the forward-mapping deformation field enhances its ability to model complex motions. The ablation study demonstrates the effectiveness of the proposed modules in our network, such as depth initialization, deformation regularization and color smoothness. We conduct extensive experiments, and our method significantly outperforms the SOTA methods in reconstruction quality. Our method provides a new idea for the reconstruction of surgical scenes and will further promote the development of robot-assisted surgery and intelligent medical care.

Acknowledgments. This work is supported in part by the Natural Science Foundation of China (No.62072388), the public technology service platform project of Xiamen City(No.3502Z20231043), and the Fujian Sunshine Charity Foundation, and in part by the Guangdong Higher Education Upgrading Plan of "Rushing to the Top, Making Up Shortcomings, and Strengthening Special Features" (No.2023KQNCX091), and the Guangdong Provincial Key Laboratory IRADS (No.2022B1212010006, R0400001-22).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Batlle, V.M., Montiel, J.M., Fua, P., Tardós, J.D.: Lightneus: Neural surface reconstruction in endoscopy using illumination decline. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14229, pp. 502–512. Springer, Cham (2023), https://doi.org/10.1007/978-3-031-43999-5_48
2. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. p. 333–350. Springer-Verlag, Berlin, Heidelberg (2022), https://doi.org/10.1007/978-3-031-19824-3_20
3. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023)
4. Chong, N., Si, Y., Zhao, W., Zhang, Q., Yin, B., Zhao, Y.: Virtual reality application for laparoscope in clinical surgery based on siamese network and census transformation. In: MICAD. pp. 59–70. Springer (2021)
5. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: CVPR. pp. 12479–12488 (2023)
6. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. NeurIPS **35**, 33768–33780 (2022)

7. Hayoz, M., Hahne, C., Gallardo, M., et al.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery **18**, 1185–1192 (2023)
8. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
9. Kinga, D., Adam, J.B., et al.: A method for stochastic optimization. In: ICLR. San Diego, California; (2015)
10. Li, R., Gao, H., Tancik, M., Kanazawa, A.: Nerfacc: Efficient sampling accelerates nerfs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18537–18546 (2023)
11. Li, Y., Richter, F., Lu, J., Funk, E.K., Orosco, R.K., Zhu, J., Yip, M.C.: Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics. IEEE Robotics and Automation Letters **5**(2), 2294–2301 (2020)
12. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6197–6206 (2021)
13. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2305.04966 (2024)
14. Long, Y., et al.: E-dssr: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 415–425. Springer, Cham (2021), https://doi.org/10.1007/978-3-030-87202-1_40
15. Lu, J., Jayakumari, A., Richter, F., Li, Y., Yip, M.C.: Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 4783–4789. IEEE (2021)
16. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
17. Ma, R., Wang, R., Pizer, S., Rosenman, J., McGill, S.K., Frahm, J.M.: Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 573–582. Springer, Cham (2019), https://doi.org/10.1007/978-3-030-32254-0_64
18. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
19. Montana-Brown, N., Saeed, S.U., et al.: Saramis: Simulation assets for robotic assisted and minimally invasive surgery. In: NeurIPS. vol. 36 (2024)
20. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR. pp. 10318–10327 (2021)
21. Schmidt, A., Mohareri, O., DiMaio, S., Yip, M.C., Salcudean, S.E.: Tracking and mapping in medical computer vision: A review. Medical Image Analysis pp. 103–131 (2024)
22. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016)
23. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robotics and Automation Letters **3**(1), 155–162 (2017)
24. Tagliabue, E., Piccinelli, M., Dall'Alba, D., Verde, J., Pfeiffer, M., Marin, R., Speidel, S., Fiorini, P., Cotin, S.: Intra-operative update of boundary conditions for patient-specific surgical simulation. In: de Bruijne, M., et al. (eds.)

MICCAI 2021. LNCS, vol. 12904, pp. 373–382. Springer, Cham (2021), https://doi.org/10.1007/978-3-030-87202-1_36
25. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: Wang, L., et al. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 431–441. Springer, Cham (2022), https://doi.org/10.1007/978-3-031-16449-1_41
26. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Xinggang, W.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR (2024)
27. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14228, pp. 46–56. Springer, Cham (2023), https://doi.org/10.1007/978-3-031-43996-4_5
28. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: CVPR (2023)
29. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14228, pp. 13–23. Springer, Cham (2023), https://doi.org/10.1007/978-3-031-43996-4_2
30. Zhou, H., Jayender, J.: Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 331–340. Springer, Cham (2021), https://doi.org/10.1007/978-3-030-87202-1_32