REAL-TIME SPATIO-TEMPORAL RECONSTRUCTION OF DYNAMIC ENDOSCOPIC SCENES WITH 4D GAUSSIAN SPLATTING

Fengze Li[1,†], Jishuai He[†], Jieming Ma[1,2,⋆], Zhijing Wu[3]

University of Liverpool, Liverpool, UK[1],
Xi'an Jiaotong-Liverpool University, Suzhou, China[2],
University of Cambridge, Cambridge, UK[3]

ABSTRACT

Dynamic scene reconstruction is essential in robotic minimally invasive surgery, providing crucial spatial information that enhances surgical precision and outcomes. However, existing methods struggle to address the complex, temporally dynamic nature of endoscopic scenes. This paper presents ST-Endo4DGS, a novel framework that models the spatio-temporal volume of dynamic endoscopic scenes using unbiased 4D Gaussian Splatting (4DGS) primitives, parameterized by anisotropic ellipses with flexible 4D rotations. This approach enables precise representation of deformable tissue dynamics, capturing intricate spatial and temporal correlations in real time. Additionally, we extend spherindrical harmonics to represent time-evolving appearance, achieving realistic adaptations to lighting and view changes. A new endoscopic normal alignment constraint (ENAC) further enhances geometric fidelity by aligning rendered normals with depth-derived geometry. Extensive evaluations show that ST-Endo4DGS outperforms existing methods in both visual quality and real-time performance, establishing a new state-of-the-art in dynamic scene reconstruction for endoscopic surgery.

Index Terms— 3D Reconstruction, Gaussian Splatting, Robotic Surgery, Endoscopic Image

1. INTRODUCTION

Accurately modeling dynamic scenes from 2D images and rendering lifelike novel views in real-time are critical capabilities in surgical robotics, particularly for endoscopic surgery. Dynamic scene reconstruction significantly enhances the accuracy of robotic-assisted minimally invasive procedures by accurately restoring the spatial and physical properties of human tissues [1]. This, in turn, contributes to better patient outcomes by supporting higher precision in surgical interventions [2, 3].

Advanced techniques such as Neural Radiance Fields (NeRF) [4] and 3D Gaussian Splatting (3DGS) [5] have recently shown remarkable potential in rendering highly realistic static scenes. EndoNeRF was the first to apply NeRF in endoscopic scenes, modeling tissue deformation and canonical density to enhance dynamic scene reconstruction accuracy [1]. EndoSurf advanced this by introducing signed distance functions to impose self-consistency constraints on neural fields, improving geometric fidelity, though at the cost of extended training times [6]. SurgicalGS achieves computational efficiency through explicit Gaussian primitives and splatting-based rendering, yet lacking temporal resolution for complex surgical dynamics [7]. However, adapting these methodologies to dynamic endoscopic scene reconstruction presents substantial challenges.

The continuous motion and temporal dynamics in surgical settings complicate the adaptation of conventional methods to endoscopic applications. As endoscopic procedures capture dynamic scenes through monocular video streams [8], it is challenging to implement traditional approaches that train static representations per frame and merge them into a dynamic model. Endo-4DGS [9] tackled real-time dynamics using a 4D Gaussian deformation field with pseudo-depth data from Depth-Anything [10], enabling real-time reconstruction without ground-truth depth. LGS further optimized 4DGS for surgical settings by incorporating deformation-aware pruning and feature compression, reducing memory and computational costs while maintaining quality [11]. Additionally, EndoGaussian introduced holistic Gaussian initialization to enhance tracking stability and computational efficiency [12]. EndoGS integrates depth-guided supervision, spatio-temporal weighting, and surface-aligned regularization to improve geometric fidelity and robustness against occlusions [13]. Despite these advancements, accurately capturing the spatio-temporal structure of dynamic scenes and managing scale and deformation modeling remain challenging, as many methods still experience local overfitting and limited alignment with true surface geometry.

To address these issues, this study redefines the task by approximating an endoscopic scene's underlying spatio-temporal 4D volume through a set of 4D Gaussian functions, offering a flexible and unified approach to representing com-
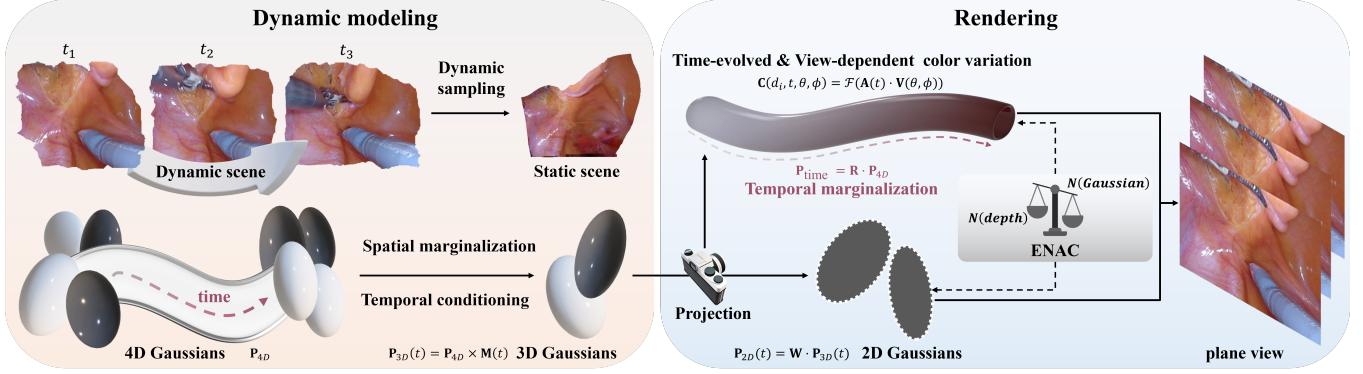
---

† Equally Contributed.

Fig. 1. Diagram of the ST-Endo4DGS framework, demonstrating the transition from 4D Gaussians to 2D projections, effectively capturing dynamic endoscopic scenes. Temporal and spatial correlations are preserved through marginalization, with ENAC ensuring alignment of rendered and depth-derived geometry for enhanced fidelity.

plex endoscopic dynamics [14]. By enabling 4D rotations, these Gaussians adapt to the 4D manifold, effectively capturing the intrinsic motion in endoscopic scenes. Furthermore, we extend endoscopic spherindrical harmonics, a generalization of 4D spherical harmonics [14, 15], to simulate the time-evolved appearance of dynamic endoscopic scenes, allowing for realistic lighting and appearance changes over time. Moreover, inspired by unbiased depth rendering techniques [16] and normal consistency [17], which are effective in preserving surface detail and reducing local overfitting in complex scenarios, we introduce an endoscopic normal alignment constraint (ENAC) to enhance alignment with true surface geometry, thereby improving the geometric fidelity of dynamic endoscopic scenes.

Therefore, the primary contributions are: **(i)** We propose ST-Endo4DGS, a novel framework that utilizes unbiased 4DGS for precise spatio-temporal modeling of dynamic endoscopic scenes, effectively capturing the complex spatial and temporal dynamics of deformable tissues. **(ii)** We extend endoscopic spherindrical harmonics to model time-evolved appearance, capturing complex lighting and appearance changes in endoscopic scenes with improved temporal fidelity and interpretability. **(iii)** The endoscopic normal alignment constraint (ENAC) is proposed to enhance alignment between rendered and depth-derived normals, improving the accuracy of monocular depth estimation. **(iv)** Extensive validation demonstrates that ST-Endo4DGS outperforms existing methods in visual quality and real-time performance.

2. ST-ENDO4DGS

We introduce ST-Endo4DGS, a framework designed for dynamic endoscopic scene modeling, illustrated in Fig. 1. ST-Endo4DGS utilizes unbiased 4DGS for precise spatio-temporal modeling, extends spherindrical harmonics to capture time-evolving appearance under dynamic lighting, and

employs the ENAC constraint to align rendered normals with depth-derived geometry, enhancing depth estimation accuracy. Therefore, this section is organized into four subsections. Subsection 2.1 reviews 3DGS, establishing its role as a foundational approach for dynamic scene reconstruction. Subsection 2.2 presents our 4DGS-based method for comprehensive spatio-temporal modeling. Subsection 2.3 details the extension of spherindrical harmonics for dynamic appearance adaptation, while Subsection 2.4 discusses ENAC's role in improving geometric fidelity for monocular depth estimation.

2.1. 3DGS as a baseline

3DGS models static 3D scenes using anisotropic Gaussians, enabling real-time, high-fidelity view synthesis through a GPU-optimized rasterizer. This technique has become a foundational baseline for endoscopic scene reconstruction and serves as a key inspiration for our approach. In 3DGS, the scene is represented by a cloud of 3D Gaussians, each affecting a spatial point $x \in \mathbb{R}^3$ via an unnormalized Gaussian function:

$$p(x|\mu, \Sigma) = \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right), \quad (1)$$

where $\mu \in \mathbb{R}^3$ is the mean vector, and $\Sigma = RS^2R^T$ is the anisotropic covariance matrix, with $S = \text{diag}(s_x, s_y, s_z)$ and R denoting rotation via a quaternion. Each Gaussian includes view-dependent color encoded with spherical harmonics coefficients and an opacity α, both optimized through rendering loss. To refine geometry and quality, 3DGS applies densification and pruning on the Gaussians during optimization.

2.2. Spatio-temporal modeling with 4DGS

To capture the complex spatio-temporal dynamics in endoscopic scenes, we extend 3DGS into a 4D framework

by integrating the temporal dimension, which enables ST-Endo4DGS to model both spatial structure and temporal evolution, capturing the dynamic nature of tissue deformation and motion. Temporal conditioning is applied to transform a 4D Gaussian into a time-conditioned 3D Gaussian, effectively reducing the 4D representation to a 3D spatial model that varies over time:

$$P_{3D}(t) = P_{4D} \times M(t), \tag{2}$$

where $P_{3D}(t)$ is the time-conditioned 3D Gaussian, P_{4D} denotes the complete 4D Gaussian distribution, and $M(t)$ encodes temporal dependencies. Then, ST-Endo4DGS projects the time-conditioned 3D Gaussian to a 2D Gaussian in the image plane by marginalizing out the depth component:

$$P_{2D}(t) = W \cdot P_{3D}(t), \tag{3}$$

where W is the projection matrix that maps the 3D Gaussian to the 2D plane, aligning with the view-based rendering process. This step captures the spatial layout while retaining temporal variation. To model time-evolving, view-dependent appearance, we define the color variation $C(d_i, t, \theta, \phi)$ as a function of both time t and viewing direction d_i:

$$C(d_i, t, \theta, \phi) = \mathcal{F}\big(A(t) \cdot V(\theta, \phi)\big), \tag{4}$$

where $A(t)$ is a time-dependent coefficient matrix, $V(\theta, \phi)$ represents the viewing direction, and $\mathcal{F}$ captures the interaction between temporal changes and view-dependent lighting effects. This formulation allows ST-Endo4DGS to dynamically adapt appearance based on both time and viewpoint. Temporal marginalization further refines the temporal aspect of the Gaussian model by consolidating the 4D representation into a time-marginalized distribution:

$$P_{time} = R \cdot P_{4D}, \tag{5}$$

where R is a transformation matrix responsible for temporal integration.

By combining these transformations, ST-Endo4DGS achieves a unified spatio-temporal model that accommodates both static and dynamic properties in endoscopic scenes. The final rendering equation, integrating all components, is expressed as:

$$\mathcal{I} = \sum_{i=1}^{n} \alpha_i \cdot \big(W \cdot (P_{4D} \times M(t))\big) \cdot \big(R \cdot P_{4D}\big) \cdot \mathcal{F}\big(A(t) \cdot V(\theta, \phi)\big), \tag{6}$$

where each term incorporates spatial conditioning, temporal dependency, and view-based color adaptation, enabling high-fidelity spatio-temporal rendering of dynamic endoscopic scenes.

2.3. Time-evolved appearance with endoscopic spherindrical harmonics

In ST-Endo4DGS, capturing the time-evolved and view-dependent appearance of dynamic endoscopic scenes requires a 4D extension of spherical harmonics to incorporate both spatial and temporal dimensions within a unified model. We introduce endoscopic spherindrical harmonics, specifically designed for this purpose, to efficiently encode temporal evolution and angular dependencies within a single representation, minimizing redundancy while ensuring temporal consistency. Thus, we redefine the endoscopic spherindrical harmonics basis function as:

$$Z_{nl}^m(t, \theta, \phi) = Y_l^m(\theta, \phi) \cdot \sin\left(\omega_n t + \phi_n\right), \tag{7}$$

where $Y_l^m(\theta, \phi)$ captures spatial orientation, and $\sin(\omega_n t + \phi_n)$ introduces temporal variation. Here, $\omega_n = \frac{2\pi n}{T}$ is the angular frequency corresponding to the temporal component with period T, and ϕ_n is a phase shift parameter aligning appearance changes with endoscopic dynamics. This construction establishes an orthonormal basis over the spatio-temporal domain, enabling ST-Endo4DGS to represent time-dependent, view-consistent appearance in a compact and temporally coherent form.

2.4. Enhancing geometric fidelity with ENAC

The ENAC enhances geometric fidelity by aligning normals derived from rendered depth maps, $N(\text{depth})$, with those from the Gaussian representation, $N(\text{Gaussian})$. This alignment ensures a precise depiction of surface geometry in dynamic endoscopic scenes. Thus, ENAC introduces an L_1 regularization term:

$$L_{\text{ENAC}} = |N(\text{depth}) - N(\text{Gaussian})|_1. \tag{8}$$

This term acts as a supplementary constraint, refining surface accuracy while supporting the main rendering objective. The rendering model in ST-Endo4DGS, as detailed in Equation 6, integrates both spatial and temporal aspects of the 4D Gaussian representation. The overall optimization objective, balancing rendering fidelity and geometric consistency, is defined as:

$$L_{\text{total}} = L_{\text{render}} + \lambda_{\text{ENAC}} \cdot L_{\text{ENAC}}, \tag{9}$$

where L_{render} is the main rendering loss, L_{ENAC} ensures normal alignment, and λ_{ENAC} controls its weight in the optimization. Thus, by seamlessly integrating ENAC within ST-Endo4DGS, our approach achieves enhanced alignment with true surface normals, resulting in finer detail and improved coherence across complex endoscopic scenes.

3. EXPERIMENTS

We evaluate our model's performance on the EndoNeRF dataset [1], which includes two prostatectomy samples captured with stereo cameras, providing depth maps and presenting challenges such as tool occlusions and non-rigid tissue deformations. We follow a 7:1 train-validation split and assess 3D scene reconstruction quality using PSNR, SSIM,
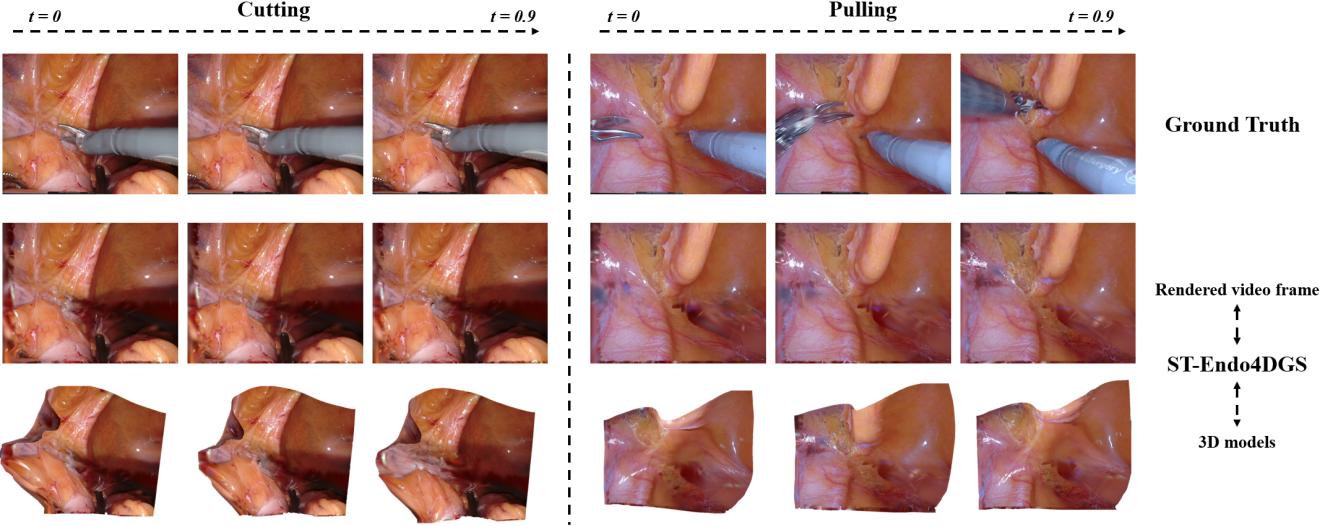
Fig. 2. Comparative results on the EndoNeRF dataset showcasing the performance of ST-Endo4DGS under Cutting and Pulling scenarios. The Rendered video frames and 3D models generated by ST-Endo4DGS closely match the ground truth.

Table 1. Comparison and ablation experiments on the EndoNeRF dataset [1], where our models and the best results are in bold.

Models	EndoNeRF-Cutting			EndoNeRF-Pulling			FPS ↑
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
EndoNeRF [1]	35.84	0.942	0.057	35.43	0.939	0.064	0.2
EndoSurf [6]	34.89	0.952	0.107	34.91	0.955	0.120	0.04
Endo-4DGS [9]	36.00	0.950	0.039	37.71	0.960	0.041	100
EndoGS [13]	37.16	0.953	0.045	36.19	0.941	0.041	70
ST-Endo4DGS without ENAC	39.17	0.950	0.039	37.78	0.960	0.041	**123**
ST-Endo4DGS	**39.29**	**0.973**	**0.016**	**38.28**	**0.966**	**0.024**	**123**

and LPIPS metrics. All experiments are conducted on an NVIDIA RTX4090 GPU using PyTorch. Consistent with the Endo-4DGS [9] baseline, we adopt a learning rate of 1.6×10^{-3}.

In our experimental results (shown in Table. 1 and Fig. 2), we compare the performance of ST-Endo4DGS against recent models, including EndoNeRF [1], EndoSurf [6], Endo-4DGS [9], and EndoGS [13], focusing on PSNR, SSIM, LPIPS, and FPS metrics. Our ST-Endo4DGS model demonstrates significant improvements over the counterparts, achieving state-of-the-art performance. Specifically, ST-Endo4DGS achieves up to 6.5% higher PSNR in the Cutting and Pulling scenarios, with values reaching 39.29 and 38.28, respectively. In terms of perceptual quality, ST-Endo4DGS also records lower LPIPS scores, reflecting a 41.5% improvement over Endo-4DGS. Furthermore, ST-Endo4DGS operates at 123 FPS, markedly faster than Endo-4DGS's 100 FPS, representing a 23% increase in speed. This enhanced efficiency, combined with superior visual fidelity, underscores the advantages of our approach, making ST-Endo4DGS more

suitable for real-time applications in dynamic and complex endoscopic environments. Our model not only provides improved image quality but also ensures practical applicability through its computational efficiency.

4. CONCLUSION

We present ST-Endo4DGS, a pioneering framework for dynamic endoscopic scene reconstruction that achieves real-time, high-fidelity novel view synthesis by approximating the spatio-temporal 4D volume through 4D Gaussian primitives. By leveraging unbiased 4D Gaussian Splatting, extended spherindrical harmonics, and the ENAC constraint, ST-Endo4DGS effectively captures intricate tissue deformations and maintains accurate surface alignment. Extensive evaluations on the EndoNeRF dataset confirm the superiority of our approach, with ST-Endo4DGS outperforming existing methods in reconstruction quality, achieving state-of-the-art results. This work establishes a new benchmark for real-time reconstruction of dynamic endoscopic scenes.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 431–441.

[2] Valentina Vitiello, Su-Lin Lee, Thomas P Cundy, and Guang-Zhong Yang, "Emerging robotic platforms for minimally invasive surgery," *IEEE reviews in biomedical engineering*, vol. 6, pp. 111–126, 2012.

[3] Jenna Seetohul, Mahmood Shafiee, and Konstantinos Sirlantzis, "Augmented reality (ar) for surgical robotic and autonomous systems: State of the art, challenges, and solutions," *Sensors*, vol. 23, no. 13, pp. 6202, 2023.

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023.

[6] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge, "Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 13–23.

[7] Jialei Chen, Xin Zhang, Mobarakol Islam, Francisco Vasconcelos, Danail Stoyanov, Daniel S Elson, and Baoru Huang, "Surgicalgs: Dynamic 3d gaussian splatting for accurate robotic-assisted surgical scene reconstruction," *arXiv preprint arXiv:2410.09292*, 2024.

[8] David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[9] Yiming Huang, Beilei Cui, Long Bai, Ziqi Guo, Mengya Xu, Mobarakol Islam, and Hongliang Ren, "Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 197–207.

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10371–10381.

[11] Hengyu Liu, Yifan Liu, Chenxin Li, Wuyang Li, and Yixuan Yuan, "Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 660–670.

[12] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan, "Endogaussian: Gaussian splatting for deformable surgical scene reconstruction," *arXiv preprint arXiv:2401.12561*, 2024.

[13] Lingting Zhu, Zhao Wang, Zhenchao Jin, Guying Lin, and Lequan Yu, "Endogs: Deformable endoscopic tissues reconstruction with gaussian splatting," *arXiv preprint arXiv:2401.11535*, 2024.

[14] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang, "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting," in *International Conference on Learning Representations (ICLR)*, 2024.

[15] Robert T Seeley, "Spherical harmonics," *The American Mathematical Monthly*, vol. 73, no. 4P2, pp. 115–121, 1966.

[16] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang, "Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction," *arXiv preprint arXiv:2406.06521*, 2024.

[17] Lue Fan, Yuxue Yang, Minxing Li, Hongsheng Li, and Zhaoxiang Zhang, "Trim 3d gaussian splatting for accurate geometry representation," *arXiv preprint arXiv:2406.07499*, 2024.