This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Endo-GSMT: Endoscopic Monocular Scene Reconstruction with Dynamic Gaussian Splatting and Motion Tracking

Hao Gou[1] ⋆, Changmiao Wang[4] ⋆, Jiahao Yang[5], Yaoqun Liu[6], Fucang Jia[2], Deqiang Xiao[7], Feiwei Qin[1](✉), and Huoling Luo[2,3](✉)

[1] Hangzhou Dianzi University, Hangzhou, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3] Shenzhen Institute of Information Technology, Shenzhen, China
[4] Shenzhen Research Institute of Big Data, Shenzhen, China
[5] Sun Yat-Sen University, Shenzhen, China
[6] The Chinese University of Hong Kong, Shenzhen, China
[7] Beijing Institute of Technology, Beijing, China
qinfeiwei@hdu.edu.cn, luohl@sziit.edu.cn

Abstract. Limited perspectives and complex tissue deformations pose significant challenges in accurately reconstructing monocular dynamic surgical scene. Many existing methods fail to fully exploit inter-frame relationships, resulting in suboptimal performance in processing complex tissue deformations and synthesizing novel views. To address these challenges, we propose Endo-GSMT, an accurate and high-quality method for dynamic endoscopic reconstruction from monocular surgical videos. Our method begins by comprehensively extracting both intra-frame information and inter-frame relationships from the raw monocular videos. We incorporate monocular depth priors and dense displacement field priors to generate the pixel-wise 3D trajectories during the training phase. Then, we design a set of compact and low-dimensional Sim(3) motion bases, with each point's motion represented as a weighted combination of these motion bases. Furthermore, we develop a novel depth loss function to address the scale inconsistency inherent in monocular depth priors. We evaluate our method using two distinct evaluation strategies, the experimental results demonstrate that our method achieves state-of-the-art reconstruction quality. The code is available at https://github.com/M11pha/Endo-GSMT.

Keywords: 3D Gaussian Splatting · Monocular Dynamic Novel View Synthesis · Surgical Scene Reconstruction

1 Introduction

Dynamic 3D reconstruction of deformable surgical scenes from endoscopic videos plays a critical role in modern medical procedures [7,16,14]. High-quality visual

⋆ Co-first authors.

reconstruction significantly facilitates downstream clinical applications, including robotic-assisted minimally invasive surgery and augmented reality surgical navigation [25], by providing surgeons with enhanced spatial understanding of tissue structures. This understanding not only improves the effectiveness of surgical procedures, but also improves the quality of medical training [12,9]. However, achieving this task is challenging due to limitations such as limited camera perspectives, occlusions caused by surgical instruments, and the inherent difficulty in accurately modeling non-rigid tissue deformations.

Traditional 3D reconstruction methods typically rely on complex, multi-step workflows [13]. Some studies [20,25,22] leverage Neural Radiance Field (NeRF) [11] to optimize the process and improve the quality of the reconstruction. Despite their innovation, these methods face slow training and rendering speed, while their implicit representations constrain applicability in subsequent tasks.

3D Gaussian Splatting (3D-GS) [5] combines the flexibility of implicit representations with the structure of explicit methods, enabling high-speed, high-quality rendering through its parallelizable pipeline. Recent works [26,10,4,24] leverage 3D-GS to overcome the limitations of NeRF-based methods, employing 3D Gaussian representations in a canonical space and integrating deformation fields to model deformable surgical scenes. EndoGS [26] utilizes Structure-from-Motion (SfM) [13] to initialize 3D Gaussian point cloud. However, the reliance on multi-view consistency of SfM often leads to suboptimal initialization, which prolongs training times and hampers accurate model fitting. To address this shortcomings, several studies [10,4,24] introduce depth priors to back-project the first-frame endoscopic capture into 3D space to initialize the 3D Gaussians. Besides, EndoGaussian [10] and Endo-4DGS [4] leverage HexPlane [1,21] to construct the deformation fields, while Deform3DGS [24] explicitly models the deformation of 3D Gaussians as linear combinations of Gaussian functions. These methods markedly reduce training times and increase rendering speeds while enhancing reconstruction quality. However, they rely solely on single-frame information with depth supervision and do not fully exploit the inter-frame relationships. Additionally, depth priors based on single-frame images suffer from inter-frame scale inconsistency. These limitations hinder existing methods from learning accurate time-varying deformation fields and capturing long-range 3D motion trajectories in videos. We argue that incorporating inter-frame relationships and maintaining depth stability could substantially enhance the training of accurate deformation fields, thereby providing more diverse and accurate information and enabling more effective interaction with the surgical scene.

In this paper, we present Endo-GSMT, a highly accurate and high-quality framework designed for reconstructing deformable surgical scenes. We model the dynamic scene using a set of canonical 3D Gaussians, integrating depth priors and dense displacement field priors to guide the reconstruction process and track pixel-wise trajectories. The motion of the scene is represented by a set of compact and low-dimensional Sim(3) motion bases. Moreover, we propose a novel depth loss function to address the scale inconsistency between depth priors. We evaluate

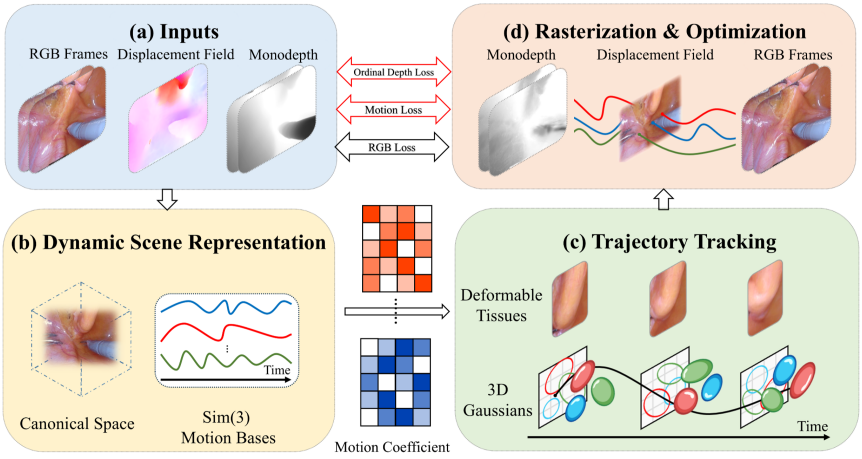

Fig. 1. The overview of our proposed Endo-GSMT pipeline. The framework comprises four sequential components: (a) Inputs, (b) Dynamic Scene Representation (c) Trajectory Tracking, and (d) Rasterization and Optimization.

our method on the EndoNeRF [20] and StereoMIS [3] datasets, the experimental results demonstrate that Endo-GSMT achieve state-of-the-art performance.

2 Method

Pipeline. As shown in Fig. 1, we first introduce depth priors and dense displacement field priors as auxiliary inputs (Sec. 2.1). We represent a dynamic scene using 3D Gaussians. To capture the scene motion, we use the two complementary priors to generate a set of compact and low-dimensional Sim(3) motion bases, and model each 3D Gaussian's motion by a weighted combination of these bases (Sec. 2.2). Additionally, we design a novel depth loss function to resolve the scale inconsistency between depth priors (Sec. 2.3). Finally, we optimize the framework by comparing the rendered results (Sec. 2.4).

2.1 Preliminaries: 3D Gaussian and Introduction of Priors

Gaussians in the canonical space. We model a dynamic surgical scene using global 3D Gaussians in a canonical frame. Each 3D Gaussian is defined by parameters $g \equiv (\mu, r, s, o, c)$, where $\mu \in \mathbb{R}^3$, $r \in \mathbb{R}^4$ and $s \in \mathbb{R}^3$ are the 3D mean, orientation and scale in the canonical frame, and $o \in \mathbb{R}$ the opacity, $c \in \mathbb{R}^3$ the color, are consistent across all time steps.

Introducing complementary priors. We use depth priors $\{D_t \in \mathbb{R}^{H\times W\times T}\}$ [23] and dense displacement field priors [2] $\{F_{t_i\to t_j} \in \mathbb{R}^{H\times W\times 2} | i, j \in \{1, \ldots, T\}\}$ to generate 3D optical flow. We first use depth priors to lift the 2D image pixels into 3D space. Then, we apply dense displacement fields to link 3D points across different frames, producing noisy initial 3D trajectories $\{\mathbf{X}(t) \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$.

2.2 Dynamic Scene Representation for Deformable Tissues

Dynamic Scene Representation. To reconstruct a dynamic surgical scene, we define a set of 3D Gaussians $\{\mathcal{G}\}$ in the canonical frame, and control their positions, orientations and scales over time. Inspired by [19], the 3D Gaussians at any time step t is computed based on $\{\mathcal{G}\}$ via $\mathbf{T}_{c\to t} = \begin{bmatrix} \mathbf{R}_{c\to t} & \mathbf{t}_{c\to t} \\ 0 & s_{c\to t}^{-1} \end{bmatrix} \in \mathrm{Sim}(3)$:

$$\boldsymbol{\mu}_t = s_{c\to t}(\mathbf{R}_{c\to t}\mu_c + t_{c\to t}), \quad \mathbf{R}_t = \mathbf{R}_{c\to t}\mathbf{R}_c, \quad s_t = s_{c\to t}s_c, \tag{1}$$

we choose the frame with the most visible 3D trajectories as the canonical frame $\mathbf{I}_c$ and randomly sample N 3D trajectories from $\mathbf{I}_c$ to initialize 3D Gaussians.

Modeling each Gaussian's 3D motion trajectory independently would be computationally expensive. Instead, we use the initial 3D trajectories $\{\mathbf{X}(t)\}$ to derive a set of globally shared, learnable basis trajectories $\{\mathbf{T}_{0\to t}^{(b)}\}_{b=1}^{B}$. The transformation $\mathbf{T}_{c\to t}$ for each Gaussian at any time step t is computed as a weighted combination of these basis trajectories. Specifically, we conduct vectorized velocity analysis on the noisy trajectories $\{\mathbf{X}(t)\}$ and apply K-means clustering to group them into B trajectory sets $\{\mathbf{X}(t)\}_{b=1}^{B}$. For the trajectory set $\{\mathbf{X}(t)\}_b$ in the b-th cluster, we align the canonical frame point set $\{\mathbf{X}(c)\}_b$ with $\{\mathbf{X}(\tau)\}_b$ for all time steps $\tau = 0, \ldots, T$ using weighted Procrustes alignment, thereby obtaining the initial basis transformations $\mathbf{T}_{c\to\tau}^{(b)}$. Subsequently, we initialize the weights $\mathbf{w}^{(b)}$ for each Gaussian by applying an exponential decay based on the distance to the cluster center. At each time step t, we compute the transformation $\mathbf{T}_{c\to t}$ by weighting and combining the global basis trajectory set using the per-cluster motion coefficients $\mathbf{w}^{(b)}$:

$$\mathbf{T}_{c\to t} = \sum_{b=0}^{B} \mathbf{w}^{(b)}\mathbf{T}_{c\to t}^{(b)}. \tag{2}$$

Rasterizing 3D Trajectories. Based on this representation, we now describe how to track pixel-wise 3D motion trajectories at query time t with target time t'. We use $\{\mathcal{G}\}$ to establish a dense displacement field between the two time steps [18]. Specifically, given a pixel p at time t, we perform rasterization via α-blending [5] to compute the expected 3D world coordinates of pixel p at the target time t':

$$\hat{\mathbf{X}}_{t\to t'}(\mathbf{p}) = \sum_{i\in H(\mathbf{p})} T_i\alpha_i\boldsymbol{\mu}_{i,t'}, \tag{3}$$

where $H(\mathbf{p})$ is the set of Gaussians that intersect the pixel p at query time t.

Then, we project the 3D position into the 2D plane and compute the depth of pixel p at time t'. The 2D position $\hat{\mathbf{U}}_{t\to t'}$ and depth $\hat{\mathbf{D}}_{t\to t'}(\mathbf{p})$ are given by:

$$\hat{\mathbf{U}}_{t\to t'} = \mathit{\Pi}\left(\mathbf{K}_{t'}\mathbf{E}_{t'}\hat{\mathbf{X}}_{t\to t'}(\mathbf{p})\right), \quad \hat{\mathbf{D}}_{t\to t'}(\mathbf{p}) = \left(\mathbf{E}_{t'}\hat{\mathbf{X}}_{t\to t'}(\mathbf{p})\right)_{[3]}, \tag{4}$$

where $\mathbf{K}_{t'}$ and $\mathbf{E}_{t'}$ denote the camera's intrinsic and extrinsic parameters at time t'. The function Π represents the perspective projection operation, and $(\cdot)_{[3]}$ extracts the depth (third element) of the transformed 3D vector. This approach enables accurate tracking of both the 2D position and depth of pixels across frames.

2.3 Ordinal Depth Loss

Image-based monocular depth priors often provide detailed information but lack consistency between frames, which can lead to flickering in consecutive frames. A potential solution is to use video-based monocular depth estimation methods, which effectively address inter-frame scale consistency but tend to be computationally intensive. Drawing inspiration from Liu *et al.* [8], we note that while the depth values themselves may be inconsistent, the relative order of these values across different pixels remains stable over time. This observation lead us to propose an order-based depth loss function:

$$\mathcal{L}_{\text{ordinal}} = \left\| \min\Big(0,\ \text{sign}\big(\hat{D}_t(p_1) - \hat{D}_t(p_2)\big) \times \text{sign}\big(D_t(p_1) - (D_t(p_2))\big)\Big) \right\|, \quad (5)$$

where sign is the symbolic function, D_t represents the predicted depth values, and $\hat{D}_t$ denotes the rendered depth values. This function converts the depth difference between $\hat{D}_t(p_1)$ and $\hat{D}_t(p_2)$ into 1 or -1 using the sign function, then forces the order of depths in the rendered depth map $\hat{D}_t$ matches that of the predicted depth map D_t.

2.4 Optimization

We employ two sets of loss functions to guide the optimization of dynamic Gaussians. The first set focuses on reconstruction, ensuring that the predicted pixel-level colors and depth order align with the input for each frame. During each training step, we render images $\hat{\mathbf{I}}_t$ and depth maps $\hat{\mathbf{D}}_t$ using the training cameras $(\mathbf{K}_t, \mathbf{E}_t)$. We supervise these predictions by enforcing reconstruction losses on each individual frame:

$$\mathcal{L}_{\text{recon}} = \|\hat{\mathbf{I}} - \mathbf{I}\|_1 + \lambda_{\text{ordinal}} \mathcal{L}_{\text{ordinal-depth}}, \quad (6)$$

where the reconstruction loss consists of a pixel-wise $\mathcal{L}_1$ loss for texture matching and an ordinal depth loss $\mathcal{L}_{\text{ordinal-depth}}$ weighted by λ_{ordinal}, to enforce depth consistency.

The second set of losses supervises the Gaussians' motion across frames. Under the temporal smoothness constraints, we use an $\mathcal{L}_1$ loss to fit the observed 3D trajectories while optimizing the Gaussian positions $\boldsymbol{\mu}_c$, the motion coefficients $\mathbf{w}^{(b)}$, and the basis transformations $\{\mathbf{T}_{c\to t}^{(b)}\}_{b=1}^{B}$. Specifically, for a randomly sampled query time t and target time t', we render the corresponding 2D positions and depth values for each pixel at time t'. These predictions are then supervised using the introduced priors to reinforce accurate temporal correspondences:

$$\mathcal{L}_{\text{track-2d}} = \left\| \mathbf{U}_{t\to t'} - \hat{\mathbf{U}}_{t\to t'} \right\|_1, \quad \mathcal{L}_{\text{track-depth}} = \left\| \mathbf{D}_{t\to t'} - \hat{\mathbf{D}}_{t\to t'} \right\|_1. \quad (7)$$

3 Experiments and Results

3.1 Experiment Setting

Datasets. We evaluate the performance of our method on two stereo endoscopic video datasets: (1) EndoNeRF Dataset [20] contains six stereo surgical videos, each exhibiting moderate tissue deformations. (2) StereoMIS Dataset [3] contains eleven stereo surgical videos, featuring diverse scenes and complex tissue deformations. For our evaluation, we focus on two accessible scenes from the EndoNeRF dataset and five carefully selected segments from the StereoMIS dataset.

Evaluation Setting. We perform experiments using two evaluation strategies: (1) **Frame Extraction Evaluation**: following [25], we divide the video frames into training and testing sets in a 7:1 ratio. (2) **Novel View Synthesis (NVS) Evaluation**: we argue that the frame extraction evaluation using training views cannot determine if the model overfits to the training views. Therefore, we use the left view (the primary view) for training and the right view for testing, ensuring a more comprehensive performance assessment.

Unlike previous methods that assume a fixed camera viewpoint, we enhance the robustness of our approach by pre-estimating the camera's intrinsic and extrinsic parameters using Droid-SLAM [17]. Potential inaccuracies in these calibrated parameters may lead to slight spatial misalignments in the rendered results compared to the corresponding ground truth. To address this issue, we align the rendered outputs with the ground truth using a pretrained optical flow network PWC-Net [15] and compute evaluation metrics on the aligned results. Specifically, we employ the aligned Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) as evaluation metrics.

Implementation Details. We use Adam optimizer [6] for training. Specifically, we perform preliminary fitting for the first 1,000 iterations, followed by the main optimization for 500 epochs. We model the surgical scene as a globally dynamic environment and set the number of Sim(3) motion bases to 60. In the canonical space, we initialize 100,000 Gaussians and adopt the adaptive Gaussian control strategy from 3D-GS [5]. The batch size, λ_{ordinal}, $\lambda_{\text{track-2d}}$, and $\lambda_{\text{track-depth}}$ are set to 8, 0.5, 0.2, and 0.01, respectively. We implement our approach using PyTorch and train it on a single NVIDIA RTX 4090 GPU.

3.2 Experiment Results

Comparison with State-of-the-art Methods. We evaluate our proposed method against two NeRF-based methods [20,22] and four 3DGS-based methods [26,10,4,24] using frame extraction and NVS strategies. Comparison of NVS is made only among 3DGS-based methods.

As listed in Table 1, our method outperforms all baseline methods in both frame extraction and NVS evaluations. For the EndoNeRF dataset, our method achieves modest improvements in PSNR and SSIM, and delivers a notable 24%

Table 1. Quantitative evaluation of our proposed framework against existing two NeRF-based methods and four 3DGS-based methods. The best results are in **bold**.

Dataset	Method	Frame Extraction Eva.			NVS Eva.		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
EndoNeRF [20]	EndoNeRF [20]	28.355	0.918	0.090	-	-	-
	LerPlane-32k [22]	38.238	0.948	0.055	-	-	-
	EndoGS [26]	35.616	0.952	0.059	24.566	0.882	0.115
	EndoGaussian [10]	35.522	0.957	0.103	27.949	0.905	0.096
	Endo-4DGS [4]	36.945	0.957	0.037	28.318	0.909	0.092
	Deform3DGS [24]	38.259	0.960	0.062	30.469	0.921	0.083
	Endo-GSMT (Ours)	**38.783**	**0.968**	**0.028**	**30.735**	**0.928**	**0.063**
StereoMIS [3]	EndoNeRF [20]	31.922	0.857	0.146	-	-	-
	LerPlane-32k [22]	31.679	0.845	0.113	-	-	-
	EndoGS [26]	32.819	0.907	0.099	20.714	0.755	0.200
	EndoGaussian [10]	29.191	0.827	0.181	23.098	0.721	0.227
	Endo-4DGS [4]	31.580	0.862	0.124	27.461	0.802	0.156
	Deform3DGS [24]	32.209	0.863	0.124	22.131	0.702	0.214
	Endo-GSMT (Ours)	**34.703**	**0.917**	**0.060**	**29.699**	**0.863**	**0.091**

Table 2. Ablation study of different motion representations and depth losses. The best results are in **bold**.

Method	PSNR↑	SSIM↑	LPIPS↓
Per-Gaussian motion + $\mathcal{L}_1$ depth loss	37.014	0.926	0.039
Per-Gaussian motion + ordinal loss	37.620	0.934	0.035
Sim(3) motion base + $\mathcal{L}_1$ depth loss	38.202	0.956	0.031
Sim(3) motion base + ordinal loss	**38.783**	**0.968**	**0.028**

Table 3. Quantitative comparison with different numbers of Sim(3) motion base.

Number of Motion base	PSNR↑	SSIM↑	LPIPS↓
10	38.106	0.965	0.032
30	39.591	0.966	0.030
60	**38.783**	**0.968**	**0.028**
80	38.782	**0.968**	**0.028**

improvement in LPIPS over the second-best method. For the more challenging StereoMIS dataset, the superiority of our method is even more evident: it improves PSNR by 1.9 dB and 2.2 dB, SSIM by 1% and 7%, and LPIPS by 39% and 41%, respectively, compared to the second-best method. These results demonstrate that our method better captures non-rigid motions and texture details in dynamic surgical scenes, improving both reconstruction fidelity and synthesis quality.

To further demonstrate the effectiveness of our method, we provide qualitative visual comparisons. Fig. 2 shows the frame extraction results, while Fig. 3 shows the novel view synthesis results. For a clearer assessment of reconstruction quality, key areas are highlighted in green boxes. In the frame extraction comparisons (Fig. 2), our method accurately restores regions with complex textures and tissue deformations, avoiding the blurred artifacts often associated with 3DGS methods. In the more challenging novel view synthesis comparisons (Fig. 3), our approach consistently preserves excellent texture details and geometric structures. These results further confirm the robust performance of our method in enhancing both 3D reconstruction quality and dynamic scene representation.

Ablation Study. To validate the effectiveness of our proposed method, we conduct ablation study on the EndoNeRF dataset using frame extraction evaluation: (1) "Per-Gaussian motion": replacing Sim(3) motion base with naive per-Gaussian motion trajectories, and (2) "$\mathcal{L}_1$ depth loss": replacing ordinal depth

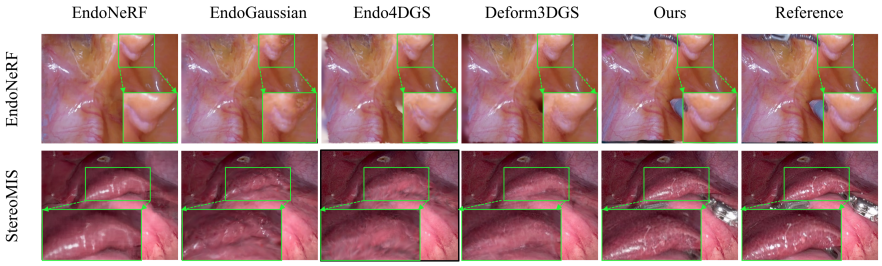

Fig. 2. Visualization of the frame extraction results.

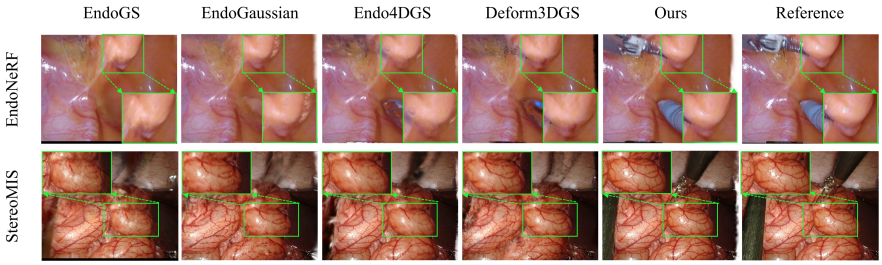

Fig. 3. Visualization of the novel view synthesis results.

loss with $\mathcal{L}_1$ depth loss. As shown in Table 2, any other combination leads to suboptimal results, thereby confirming the effectiveness of our designed Sim(3) motion base and ordinal depth loss. Furthermore, we quantitatively assess the effect of the number of Sim(3) motion bases on scene representation capability. Table 3 shows that as the number of Sim(3) motion bases increases, the scene representation capability improves, reaching a peak at a certain point. This suggests that a higher number of motion bases allows more accurate capture of non-rigid motion as well as texture details, reaching a capacity limit at a certain point, thereby enhancing both reconstruction accuracy and synthesis quality.

4 Conclusion

We present Endo-GSMT, a novel 3DGS-based framework for dynamic endoscopic reconstruction from monocular surgical videos. Our method fully utilizes both intra-frame information and inter-frame relationships by incorporating depth priors and dense displacement field priors, and employs canonical 3D Gaussians with their parameters controlled by a set of compact Sim(3) motion bases to accurately model the dynamic surgical scene with high quality. Additionally, we develop a novel depth loss function to address the scale inconsistency inherent in monocular depth priors. Extensive evaluations on two datasets demonstrate that our method significantly improves dynamic reconstruction of surgical scenes.

Acknowledgments. This study was supported in part by the Shenzhen Medical Research Fund (No. D2402006), the Fundamental Research Funds for the Provincial Uni-

versities of Zhejiang (No. GK259909299001-006), the Project (No. 20232ABC03A25), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617), Shenzhen Medical Research Fund (No. C2401036), and Anhui Provincial Joint Construction Key Laboratory of Intelligent Education Equipment and Technology (No. IEET202401).

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

References

1. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
2. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10061–10072 (2023)
3. Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., Sznitman, R.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery **18**(7), 1185–1192 (2023)
4. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2024)
5. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 139–1 (2023)
6. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Li, Q., Yang, S., Shen, D., Jin, Y.: Free-dygs: Camera-pose-free scene reconstruction based on gaussian splatting for dynamic surgical videos. arXiv preprint arXiv:2409.01003 (2024)
8. Liu, Q., Liu, Y., Wang, J., Lv, X., Wang, P., Wang, W., Hou, J.: Modgs: Dynamic gaussian splatting from causually-captured monocular videos. arXiv preprint arXiv:2406.00434 (2024)
9. Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2020)
10. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. arXiv preprint arXiv:2401.12561 (2024)
11. 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)
12. Pelanis, E., Teatini, A., Eigl, B., Regensburger, A., Alzaga, A., Kumar, R.P., Rudolph, T., Aghayan, D.L., Riediger, C., Kvarnström, N., et al.: Evaluation of

a novel navigation platform for laparoscopic liver surgery with organ deformation compensation using injected fiducials. Medical Image Analysis **69**, 101946 (2021)
13. 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)
14. Shan, J., Cai, Z., Hsieh, C.T., Cheng, S.S., Wang, H.: Deformable gaussian splatting for efficient and high-fidelity reconstruction of surgical scenes. arXiv preprint arXiv:2501.01101 (2025)
15. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8934–8943 (2018)
16. Taş, M., Yılmaz, B.: Super resolution convolutional neural network based preprocessing for automatic polyp detection in colonoscopy images. Computers & Electrical Engineering **90**, 106959 (2021)
17. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in Neural Information Processing Systems **34**, 16558–16569 (2021)
18. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19795–19806 (2023)
19. Wang, Q., Ye, V., Gao, H., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. arXiv preprint arXiv:2407.13764 (2024)
20. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022)
21. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
22. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2023)
23. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 21875–21911. Curran Associates, Inc. (2024)
24. Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., Jin, Y.: Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 132–142. Springer (2024)
25. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2023)
26. Zhu, L., Wang, Z., Cui, J., Jin, Z., Lin, G., Yu, L.: Endogs: deformable endoscopic tissues reconstruction with gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 135–145. Springer (2024)