D^{4}Recon: Dual-stage Deformation and Dual-scale Depth Guidance for Endoscopic Reconstruction

Hritam Basak[1], Zhaozheng Yin[1,2]

[1]Dept. of Computer Science, [2]Dept. of Biomedical Informatics
Stony Brook University, NY, USA
{hbasak,zyin}@cs.stonybrook.edu

Abstract. Deformable tissue reconstruction in endoscopy is vital for surgery, yet current methods struggle with high-fidelity reconstruction of irreversible tissue deformations. To this end, we present D^{4}Recon, a novel framework for real-time and high-fidelity endoscopic reconstruction, addressing crucial challenges in surgical applications. A Dual-stage Deformation modeling and a Dual-scale Depth guidance (D^4) are proposed in a dynamic 3D Gaussian Splatting paradigm along with lightweight multi-layer perception (MLP) to model dynamics in endoscopic scenes. In the dual-stage deformation modeling, we introduce a spatial deformation model to correct multiview inconsistencies, accompanied by a temporal deformation model that accurately represents tissue distortion and dynamic tissue interaction with surgical tools in the reference frames. In the dual-scale depth guidance, we propose to balance local error correction with absolute depth consistency, enabling precise depth refinement while preserving fine-grained color accuracy. D^{4}Recon generates accurate 3D reconstructions with superior PSNR, SSIM, and LPIPS scores, outperforming existing methods in terms of geometric coherence and photorealism with real-time rendering speed, as demonstrated by extensive experiments on diverse endoscopic datasets. Reconstruction videos are in the supplementary file. Website.

Keywords: 3D Reconstruction · Endoscopy · Gaussian Splatting.

1 Introduction

Endoscopic reconstruction (ER) is pivotal in minimally invasive surgery, providing enhanced intraoperative visualization and surgical guidance while reducing patient trauma and recovery time. High-fidelity ER also enables downstream applications such as simulation, AR/VR-based training, and robotic automation. However, achieving accurate reconstruction remains challenging due to the constrained nature of endoscopic environments, characterized by limited fields of view, occlusions, and complex tissue deformations induced by physiological motion and surgical interactions. These factors significantly hinder traditional ER methods, necessitating robust and efficient solutions.

Early approaches used explicit discrete representations such as point clouds [33] and surfels [16] to model scene geometry, but they struggled with complex

tissue deformations due to sparse warp fields. Neural Radiance Fields (NeRF) [18] shifted the paradigm by offering continuous representations that capture high-fidelity geometry and appearance—evidenced by EndoSurf [31] that leverages SDF [1] for surface reconstruction. LerPlane [28] further improved efficiency by encoding spatiotemporal features via orthogonal 2D planes [6]. However, NeRF-based methods require extensive ray sampling for complex surgical dynamics, incurring high computational overhead even with optimizations [13], hindering real-time rendering—a critical intraoperative requirement.

3D Gaussian Splatting (3DGS) [11] has recently gained attention for surgical scene reconstruction due to its efficient differentiable rendering scheme using anisotropic 3D Gaussians and tile rasterization, which significantly enhances training and rendering speed than NeRF [7, 12, 15]. However, static Gaussians in 3DGS are inadequate for dynamic scenes, prompting the extension to 4D Gaussian Splatting (4DGS) for temporal modeling [25]. Dynamic scene modeling with 3DGS typically involves a deformation field combined with efficient voxel encoding and lightweight decoders, as demonstrated in LGS [14] and Endo-4DGS [10]. Endo-4DGS [10] utilizes monocular depth priors from Depth-Anything [29] with a confidence-guided learning mechanism to address uncertainties in monocular depth estimation, while Deform3DGS [30] employs learnable basis functions for improved representation efficiency. Other methods integrate deformation fields to model tissue motion, leveraging efficient encodings such as orthogonal feature planes [27], regularization in MLPs [26], or Gaussian life-cycle mapping [22]. Although these methods improve the reconstruction speed, their limitations in motion hierarchy modeling and geometric accuracy in terms of precise topology and texture mapping hinder surgical applications.

Recent advancements in 3DGS and 4DGS have pushed the boundaries of ER, yet critical limitations remain in dynamic surgical environments: **(1)** existing frameworks struggle to compensate for *spatial and temporal inconsistencies and to accurately model the complex, nonrigid deformations* inherent in dynamic tissue interactions; and **(2)** Gaussian radiance fields are highly sensitive to minor depth inaccuracies, resulting in artifacts in textured regions and unstable primitive distributions [4]. Conventional scale-invariant depth losses prioritize *global alignment, often sacrificing the local geometric precision needed for fine structural details*. These challenges underscore the need for a more robust approach that we try to propose. Specifically, our contributions are as follows: **(1)** We propose a **dual-stage deformation modeling** that robustly addresses both multiview inconsistencies and dynamic tissue deformation via dual Score Distillating Sampling (SDS) losses, enabling stable 3D reconstructions; **(2)** We introduce a novel **dual-scale hard and soft depth guidance** framework that enforces a dual-scale loss: a hard constraint component that anchors global geometric consistency using absolute depth priors, and a soft constraint component that adaptively weighs local depth gradients to refine fine-grained structural details, thereby mitigating the sensitivity of Gaussian radiance fields to minor depth inaccuracies; **(3)** Upon evaluation on two dynamic real endoscopic benchmarks and three static colonoscopy benchmarks, our proposed **D^{4}Recon** produces superior reconstruction qualities with a real-time rendering efficacy.

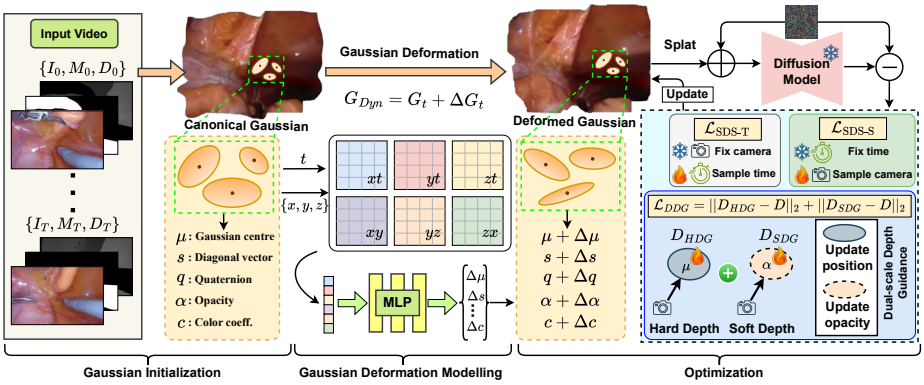

Fig. 1: Overall workflow of D[4]Recon: Gaussians are initialized from input video, followed by Gaussian deformation modeling (subsection 2.2). Finally, the Gaussians are updated using our spatiotemporal SDS losses $\mathcal{L}_{\text{SDS-S}}, \mathcal{L}_{\text{SDS-T}}$ (subsection 2.4) and dual-scale depth guidance loss $\mathcal{L}_{\text{DDG}}$ (subsection 2.3).

2 Proposed Method

Given an input endoscopic video $V = \{I_t, M_t, D_t : t \in [0, T]\}$, where for the t^{th} RGB frame I_t, M_t is its surgical tool mask and D_t is its depth map, we generate a dynamic 3D Gaussian representation G_{Dyn} as ER output. In subsection 2.1, we outline necessary preliminaries, followed by describing dynamic scene representation in subsection 2.2. We then detail the dual-scale depth guidance in subsection 2.3 and present our dual-stage spatiotemporal deformation modeling in subsection 2.4. An overview of our workflow is shown in Figure 1.

2.1 Preliminaries of 3DGS

We model 3D static scenes using 3D Gaussian Splatting (3DGS) [11], representing the scene as a set of Gaussian primitives $\{G\}$, each defined by its center μ, opacity α, color coefficients c, and a covariance matrix Σ. The Gaussian at position p is formulated as $G(p) \approx \exp[-\frac{1}{2}(p-\mu)^T \Sigma^{-1}(p-\mu)]$. To ensure positive semi-definiteness, Σ is decomposed as $\Sigma = RSS^T R^T$, where S is a scaling matrix and R is a rotation matrix, stored as diagonal vector s and quaternion vector q, respectively. Thus, each Gaussian is represented as $G = \{\mu, s, q, \alpha, c\}$. For rendering, 3D Gaussians are projected to the 2D image plane using viewing transformation V and the Jacobian J of the projective transformation, resulting in 2D covariance $\Sigma' = JV\Sigma V^T J^T$. The color C and depth D for pixel p are then computed by blending the Gaussians in depth (d) order:

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

where α'_i is the projected opacity of the i^{th} Gaussian.

2.2 Dynamic 3D Scene Representation

Gaussian Initialization: 3DGS typically relies on Structure from Motion (SfM) point clouds for Gaussian initialization, yet in endoscopic videos—with limited viewpoints, sparse textures, and dynamic lighting—SfM often produces inaccurate point clouds that hinder precise initialization. Unlike [10, 15, 26], which rely solely on SfM-derived data, we project tissue pixels into 3D space from frame 0 to create an initial point cloud. To address occlusions due to surgical tools, we update tissue pixel information from subsequent frames, generating refined image I', depth map D', and surgical tool mask M' as follows: for each pixel p, if $M_0(p) = 1$ but $M_t(p) = 0$ for frame t, we update I' and D' with values from frame t, otherwise retaining the data from frame 0. The refined point cloud P' is computed as:

$$P' = \{D' K_e^{-1} K_i^{-1} (I' \odot (1 - M'))\}, \quad M' = \cap_{\tau=0}^{t} M_\tau, \tag{2}$$

where K_i and K_e are camera intrinsics and extrinsics, respectively.

Dynamic Gaussian Representation: We then model the deformable surgical scene using a dynamic 3D Gaussian representation $G_{Dyn} = G_t + \Delta G_t$, where G_t is the static 3D Gaussian component at time t and ΔG_t encodes spatiotemporal deformations. For this deformation encoding, we utilize a multi-resolution encoder $\mathcal{E}$ comprising hex-planes $\Omega(u, v)$ and an MLP $\psi_{\mathcal{E}}$, formulated as $\mathcal{E} = \{\Omega(u, v), \psi_{\mathcal{E}}\}$ where $(u, v) \in \{(x, y), (x, z), (y, z), (x, t), (y, t), (z, t)\}$, yielding t^{th}-frame features $f_t \leftarrow \mathcal{E}(G_t)$. Here hex-planes serve as learnable 2D feature grids defined over pairs of dimensions, and when combined with MLP, form a multi-resolution encoder. We then deploy a multi-head decoder $\mathcal{D} = \{\psi_\mu, \psi_s, \psi_q, \psi_\alpha, \psi_c\}$, composed of five MLPs for deformed position, scaling, rotation, opacity, and color. The deformed 3D Gaussian is represented as:

$$\begin{aligned} G_{Dyn} &= \{\mu + \psi_\mu(f_t), s + \psi_s(f_t), q + \psi_q(f_t), \alpha + \psi_\alpha(f_t), c + \psi_c(f_t)\} \\ &= \{\mu + \Delta\mu, s + \Delta s, q + \Delta q, \alpha + \Delta\alpha, c + \Delta c\}. \end{aligned} \tag{3}$$

2.3 Dual-scale Depth Guidance

3D Gaussian Splatting (3DGS) optimizes four parameters $\{\mu, s, q, \alpha\}$ that collectively influence the reconstructed depth. However, applying uniform regularization to all parameters may lead to overfitting and blur, as monocular depth is smoother than color. To preserve both geometric fidelity and visual clarity, we selectively regularize only μ and α, which are the primary determinants of spatial position and occupancy, while keeping s and q fixed to avoid introducing color reconstruction artifacts. To enhance the shaping of Gaussian fields, we introduce hard depth guidance (HDG) that leverages the global depth cues encoded in the Gaussian centers μ. We assign a high opacity value β to all Gaussians and render a "hard depth" D_{HDG} primarily from the Gaussians nearest to the camera center ρ along rays cast from pixel p, enforcing global depth consistency:

$$D_{HDG}(p) = \sum_i \beta(1 - \beta)^{i-1} \|\rho - \mu_i\|_2. \tag{4}$$

Hard depth guidance alone is insufficient as it lacks opacity optimization, potentially leading to semitransparent surfaces and hollow structures. To address

this, we freeze μ to prevent undesirable shifts and introduce soft depth guidance (SDG) to refine α while maintaining geometry. This leverages α's role in governing local depth cues through opacity modulation:

$$D_{SDG}(p) = \sum_i \alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j)\|\rho - \mu_i\|_2. \tag{5}$$

We enforce alignment with monocular depth D using a norm-2 similarity loss:

$$\mathcal{L}_{DDG}(p) = ||D_{HDG}(p) - D(p)||_2 + ||D_{SDG}(p) - D(p)||_2. \tag{6}$$

2.4 Dual-stage Deformation Modeling

Traditional 3DGS often relies on hand-crafted depth heuristics, which struggles to resolve inherent scene ambiguity with multiview inconsistencies (e.g., non-Lambertian surfaces, transient occlusions). To address them, we propose a deformation framework that disentangles geometric and temporal refinements via two novel Score Distillation Sampling (SDS) objectives, leveraging the semantic and structural priors of a pre-trained 2D diffusion model. To mitigate multiview conflicts and temporal flickering, we introduce two deformation fields: **(1)** Spatial deformation field $\mathcal{D}_s$ that adjusts Gaussian positions to resolve static multiview inconsistencies and **(2)** Temporal deformation field $\mathcal{D}_t$ that models dynamic scene variations across time. These fields are jointly optimized with our SDS losses, aligning with diffusion prior and preserving physical plausibility.
Multiview Consistency: For spatial refinement, we sample a camera pose $\hat{P}_i$ from the canonical trajectory distribution, freeze temporal dynamics (fixing t), and render the scene via differentiable splatting: $\mathcal{I}_{\hat{P}_i} \xleftarrow[\text{splat}]{\hat{P}_i} G_{Dyn}\{\mu + \mathcal{D}_s(\mu), s, q, \alpha, c\}$. We then perturb $\mathcal{I}_{\hat{P}_i}$ with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ to obtain $\mathcal{I}^{\epsilon}_{\hat{P}_i}$ and compute noise residual ϵ_ϕ (ϕ denotes frozen diffusion weights). The spatial SDS loss $\mathcal{L}_{\text{SDS-S}}$ backpropagates gradients through deformation field $\mathcal{D}_s$:

$$\nabla_{\theta_s}\mathcal{L}_{\text{SDS-S}} = \mathbb{E}_{\hat{P}_i,\epsilon,\sigma}\left[w(\sigma)(\epsilon_\phi - \epsilon)\frac{\partial \mathcal{I}_{\hat{P}_i}}{\partial \theta_s}\right];\ \epsilon_\phi = \text{Diffusion}(\mathcal{I}^{\epsilon}_{\hat{P}_i}, \hat{P}_i, \sigma), \tag{7}$$

where $w(\sigma)$ is a noise-level-dependent weighting, and θ_s parameterizes $\mathcal{D}_s$.
Temporal Consistency: To prevent degenerated spatial solutions (e.g., flatness), we apply a temporal SDS loss $\mathcal{L}_{\text{SDS-T}}$ that enforces coherence across sampled time steps $t \sim [t_0 - \Delta t, t_0 + \Delta t]$. Here, $\mathcal{D}_t$ deforms Gaussians to positions $\mu + \mathcal{D}_t(\mu, t)$, and similar to Equation 7, $\mathcal{L}_{\text{SDS-T}}$ penalizes deviations from the diffusion prior when rendering dynamic sequences. Crucially, the spatial and temporal fields are optimized alternately, decoupling high-frequency geometric details (handled by $\mathcal{D}_s$) from low-frequency motion (handled by $\mathcal{D}_t$).

Unlike traditional SDS [19] that naively distills single-view semantics, our disentangled formulation (1) explicitly models the static-dynamic duality of real-world scenes and (2) leverages camera pose conditioning in the diffusion model to resolve the inconsistency inherent in SDS. Finally, defining $\mathbb{1}_k{=}k \bmod 2$, we

formulate the overall loss as $\mathcal{L}_{\text{total}} = \mathcal{L}_{DDG} + \mathbb{1}_k \mathcal{L}_{\text{SDS-S}} + (1 - \mathbb{1}_k)\ \mathcal{L}_{\text{SDS-T}}$. G_{Dyn} at iteration k is updated as: $G_{Dyn}^{k+1} \leftarrow G_{Dyn}^{k} - \eta \nabla \mathcal{L}_{total}$ with learning rate η.

Unlike 4DGS [10, 25] that parameterizes space-time in a single volumetric representation, our work initializes per-frame Gaussians and aligns them via localized spatiotemporal updates with dual-scale depth guidance for efficient optimization. Hence, we categorize it as **Dynamic3DGS**, emphasizing its flexible 3D representation that incrementally adapts to temporal variations.

3 Experiments and Results

Datasets We evaluate D^{4}Recon on five benchmark datasets, consisting of two surgical datasets: StereoMIS [8] and EndoNeRF [24], and three static datasets: Simulation [32], In-Vivo [17], and Phantom [2]. StereoMIS consists of 11 surgical sequences captured with the da Vinci Xi system on in-vivo porcine subjects. Following [22], we utilize two segments from videos P2_1 and P2_2. EndoNeRF comprises two prostatectomy cases with stereo-matched depth maps, comprising challenges like tool occlusion and non-rigid deformations. Both datasets are split in 7:1 training and validation ratio, following [10]. We follow [3] for the static datasets: Simulation with Unity-rendered colonoscopy sequences using RNNSLAM for depth and pose, In-Vivo with real colonoscopy videos at 270×216 resolution, and Phantom from C3VD with high-resolution sequences ("cecum_t4_b", "desc_t4_a", "transt_t1_a").

Implementation Details We adopt Adam optimizer for 3000 iterations with learning rate $\eta = 1.6 \times 10^{-3}$ and implement the pipeline using Python environment on an NVIDIA RTX4090 GPU with 24GB RAM. ArSDM [5] is utilized as the pretrained diffusion network. β in Equation 4 is set to 0.95 following validation. We follow previous works [10, 15, 30] for other settings to maintain a fair comparison. Subsection 2.2 is replaced with standard 3DGS representation for static scene reconstruction. Our performance is evaluated in terms of photorealism (PSNR and SSIM) and geometric consistency (LPIPS).

3.1 Comparison with State-of-the-art (SoTA)

Table 1 summarizes the comparison of D^{4}Recon with the existing SoTA on two surgical datasets. Notably, LerPlane [28] exhibits suboptimal performance, likely due to its limited NeuralPlane-based representation which inadequately captures the intricate, dynamic tissue deformations. NeRF-based methods like EndoSurf [31] achieve moderate metrics but are impractical for real-time use due to slow inference and extended training. LGS [14] improves time using lightweight 4DGS, but records poor metrics, likely due to oversimplified Gaussian representation without fine deformation details. SurgicalGaussian [26] improves quality but suffers from low FPS and longer times, indicating heavy computational overhead. Other 4DGS-based methods like Endo-4DGS [10] and EndoGaussian [15] although deliver satisfactory performance, their LPIPS and speed remain suboptimal. Deform3DGS [30] offers competitive quality using deformable 3DGS, yet its higher training time reflects inefficiencies in dynamic tissue handling. EH-SurGS [22] also provides high photometric quality in parts, but overall computational demands and inconsistent performance reduce its practical appeal. In contrast, our approach integrates dual-stage spatiotemporal deformation modeling and

Table 1: Quantitative evaluation of D[4]Recon on EndoNeRF and StereoMIS datasets. The best & second-best performances are highlighted in red & blue.

Method	Category	EndoNeRF-Cutting			EndoNeRF-Pulling			Average		StereoMIS			Average	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Time(s)↓	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Time(s)↓
LerPlane-32K [28]	NeuralPlane	34.66	0.923	0.071	31.77	0.910	0.071	100	240	24.12	0.814	0.327	100	255
EndoSurf [31]	NeRF	34.98	0.953	0.106	35.00	0.956	0.120	0.04	2.5e4	30.78	0.856	0.294	0.05	2.5e4
LGS [14]	4DGS	36.21	0.937	0.088	35.89	0.930	0.089	188	122	24.47	0.831	0.301	190	145
Endo-4DGS [10]	4DGS	36.56	0.955	0.032	37.85	0.959	0.043	100	240	33.85	0.894	0.165	100	420
EndoGaussian [15]	4DGS	38.29	0.962	0.058	37.31	0.958	0.070	193	120	34.37	0.899	0.158	190	130
SurgicalGaussian [26]	3DGS	37.51	0.961	0.062	38.78	0.970	0.049	82	165	30.09	0.845	0.309	86	182
Deform3DGS [30]	3DGS	37.86	0.958	0.059	37.94	0.959	0.061	335	71	34.71	0.904	0.163	332	79
EH-SurGS [22]	3DGS	39.91	0.972	0.034	38.72	0.963	0.062	383	101	34.91	0.906	0.166	365	120
Ours	Dynamic3DGS	**40.13**	**0.978**	**0.029**	**39.98**	**0.986**	**0.049**	**336**	**122**	**35.03**	**0.910**	**0.155**	**335**	**120**

Table 2: Quantitative evaluation of D[4]Recon on three static datasets.

Method	Category	Simulation			In-Vivo			Phantom		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [18]	NeRF	35.29	0.92	0.14	18.93	0.67	0.43	32.10	0.81	0.39
REIM-NeRF [20]	NeRF	32.22	0.82	0.33	18.94	0.65	0.45	31.66	0.78	0.22
Nice-SLAM [34]	SLAM	35.61	0.84	0.31	20.37	0.77	0.32	28.08	0.88	0.29
Endo-Depth [21]	DepthCNN	38.88	0.93	0.13	23.51	0.79	0.25	30.18	0.86	0.26
Endo2DTAM [9]	SLAM+3DGS	35.62	0.85	0.22	23.19	0.76	0.28	29.93	0.81	0.28
EndoGSLAM [23]	SLAM+3DGS	39.48	0.92	0.10	25.59	0.81	0.19	32.63	0.89	0.21
GPancake [3]	RNNSLAM+3DGS	40.34	0.97	0.05	26.25	0.83	0.21	32.31	0.90	0.20
Ours	Dynamic3DGS	**46.79**	**0.99**	**0.02**	**30.63**	**0.92**	**0.14**	**37.82**	**0.94**	**0.15**

dual-scale depth guidance to achieve high-fidelity and geometrically consistent outputs. These findings are well-supported by our results in Figure 2.

In Table 2, our D[4]Recon framework demonstrates a large improvement margin compared to all the latest approaches, across the three static benchmark datasets. NeRF [18] is limited by its static scene assumptions and slow volumetric rendering, while REIM-NeRF [20] extends NeRF to dynamic settings but remains vulnerable to depth estimation errors under tissue deformations. Nice-SLAM [34] provides robust geometric fidelity yet falls short in capturing fine texture details, leading to suboptimal perceptual quality. Endo-Depth [21] and Endo2DTAM [9] offer reasonable reconstructions but compromise between computational efficiency and fidelity, particularly in dynamic environments. Although EndoGSLAM [23] and GPancake [3] leverage 3DGS+SLAM, they struggle with irreversible tissue deformations and noise, especially in phantom data.

In contrast, as evident in Figure 2, our approach integrates dual-stage deformation modeling to enforce spatiotemporal consistency and address flickering artifacts (row 1), enhances depth accuracy in occluded or deformable regions (row 2,3) and refines structural coherence to eliminate surface distortions (row 4,5). Collectively, we yield high-fidelity and geometrically consistent reconstructions, as evidenced by the sharper tissue boundaries and stable geometry in Figure 2.

3.2 Ablation Results

We conduct an ablation study (Table 3) to assess the contributions of our proposed components. Experiment (a) represents the baseline without any enhancements, resulting in the lowest performance. In (b), adding the standard SDS loss [19] alone yields modest PSNR and SSIM gains, but with degraded perceptual quality (i.e., higher LPIPS), suggesting that SDS loss in isolation is insufficient. The incorporation of temporal supervision through $\mathcal{L}_{\text{SDS-T}}$ in (c) leads to significant improvements in metrics, emphasizing the importance of capturing dynamic tissue interactions. Further adding spatial supervision via $\mathcal{L}_{\text{SDS-S}}$ in (d)

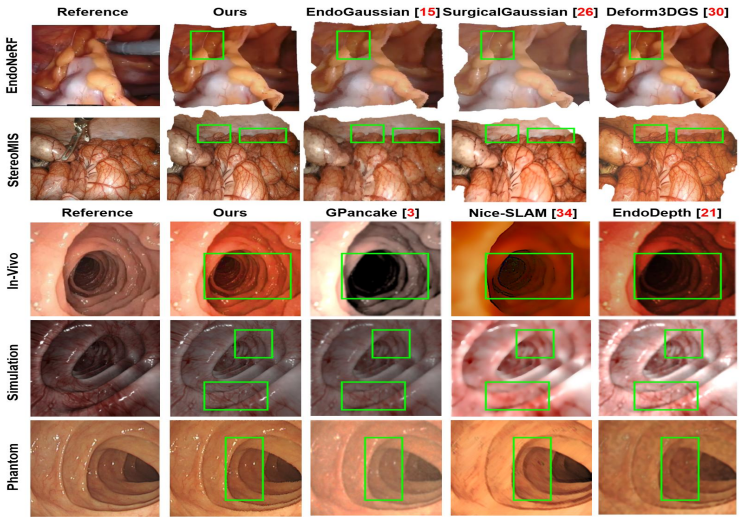

Fig. 2: Qualitative comparison of reconstruction quality.

Table 3: Ablation experiment of D^{4}Recon on surgical reconstruction datasets. Mean values of cutting and pulling scenes are reported for EndoNeRF dataset.

Exp#	D_{HDG}	D_{SDG}	D_{Any}	$\mathcal{L}_{\text{SDS-S}}$	$\mathcal{L}_{\text{SDS-T}}$	$\mathcal{L}_{\text{SDS}}$	EndoNeRF			StereoMIS		
							PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
(a)	✗	✗	✗	✗	✗	✗	31.03	0.886	0.129	25.63	0.766	0.331
(b)	✗	✗	✗	✗	✗	✓	32.19	0.903	0.853	26.71	0.781	0.305
(c)	✗	✗	✗	✗	✓	✗	36.67	0.945	0.077	29.27	0.820	0.214
(d)	✗	✗	✗	✓	✓	✗	38.42	0.968	0.052	31.38	0.886	0.176
(e)	✗	✗	✓	✓	✓	✗	38.89	0.970	0.049	32.13	0.890	0.171
(f)	✗	✓	✗	✓	✓	✗	39.69	0.973	0.045	33.91	0.902	0.168
(g)	✓	✓	✗	✓	✓	✗	**40.06**	**0.982**	**0.039**	**35.03**	**0.910**	**0.155**

refines the structural fidelity of the reconstruction, as indicated by a notable performance boost, justifying the importance of dual-stage deformation modeling. Experiment (e) shows that integrating depth cues from Depth-Anything [29], combined with spatiotemporal supervision leads to further quality gains. Subsequent introduction of SDG-based depth guidance (D_{SDG}) in (f) stabilizes the reconstruction substantially by providing robust local depth cues, particularly in areas affected by deformations. Finally, the full integration D_{HDG} in (g) achieves the best performance, demonstrating the efficacy of our proposed components.

4 Conclusion

We introduce D^{4}Recon—a novel endoscopic reconstruction framework that combines 3D Gaussian Splatting with dual-stage spatiotemporal deformation modeling and dual-scale depth guidance. Our method significantly enhances reconstruction fidelity with a high rendering speed across both static and dynamic scenarios by employing a dual-stage SDS loss that integrates spatial deformations with temporal dynamics and robust hard and soft depth guidance, demonstrating strong potential for intraoperative surgical applications. Future work will focus on incorporating physics-informed priors for biomechanical consistency and optimizing real-time adaptability for diverse endoscopic domains.

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

References

[1] Víctor M Batlle et al. "Lightneus: Neural surface reconstruction in endoscopy using illumination decline". In: *MICCAI*. 2023, pp. 502–512.

[2] Taylor L Bobrow et al. "Colonoscopy 3D video dataset with paired depth from 2D-3D registration". In: *Medical image analysis* 90 (2023), p. 102956.

[3] Sierra Bonilla et al. "Gaussian pancakes: geometrically-regularized 3D gaussian splatting for realistic endoscopic reconstruction". In: *MICCAI*. 2024, pp. 274–283.

[4] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. "Depth-regularized optimization for 3d gaussian splatting in few-shot images". In: *CVPR*. 2024, pp. 811–820.

[5] Yuhao Du et al. "ArSDM: colonoscopy images synthesis with adaptive refinement semantic diffusion models". In: *MICCAI*. 2023, pp. 339–349.

[6] Sara Fridovich-Keil et al. "K-planes: Explicit radiance fields in space, time, and appearance". In: *CVPR*. 2023, pp. 12479–12488.

[7] Jiaxin Guo et al. "Free-SurGS: SfM-Free 3D Gaussian Splatting for Surgical Scene Reconstruction". In: *MICCAI*. 2024, pp. 350–360.

[8] Michel Hayoz et al. "Learning how to robustly estimate camera pose in endoscopic videos". In: *International journal of computer assisted radiology and surgery* 18.7 (2023), pp. 1185–1192.

[9] Yiming Huang et al. "Advancing Dense Endoscopic Reconstruction with Gaussian Splatting-driven Surface Normal-aware Tracking and Mapping". In: *arXiv:2501.19319* (2025).

[10] Yiming Huang et al. "Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting". In: *MICCAI*. 2024, pp. 197–207.

[11] Bernhard Kerbl et al. "3d gaussian splatting for real-time radiance field rendering." In: *ACM Trans. Graph.* 42.4 (2023), pp. 139–1.

[12] Chenxin Li et al. "Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting". In: *MICCAI*. 2024, pp. 252–262.

[13] Ruilong Li et al. "Nerfacc: Efficient sampling accelerates nerfs". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 18537–18546.

[14] Hengyu Liu et al. "Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction". In: *MICCAI*. 2024, pp. 660–670.

[15] Yifan Liu et al. "Endogaussian: Gaussian splatting for deformable surgical scene reconstruction". In: *arXiv preprint arXiv:2401.12561* (2024).

[16] Yonghao Long et al. "E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception". In: *MICCAI*. 2021, pp. 415–425.

[17] Ruibin Ma et al. "RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy". In: *Medical image analysis* 72 (2021), p. 102100.
[18] Ben Mildenhall et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* (2021), pp. 99–106.
[19] Ben Poole et al. "Dreamfusion: Text-to-3d using 2d diffusion". In: *arXiv preprint arXiv:2209.14988* (2022).
[20] Dimitrios Psychogyios, Francisco Vasconcelos, and Danail Stoyanov. "Realistic Endoscopic Illumination Modeling for NeRF-Based Data Generation". In: *MICCAI*. 2023, pp. 535–544.
[21] David Recasens et al. "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints". In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 7225–7232.
[22] Jiwei Shan et al. "Deformable Gaussian Splatting for Efficient and High-Fidelity Reconstruction of Surgical Scenes". In: *arXiv:2501.01101* (2025).
[23] Kailing Wang et al. "Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting". In: *MICCAI*. 2024, pp. 219–229.
[24] Yuehao Wang et al. "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery". In: *MICCAI*. 2022, pp. 431–441.
[25] Guanjun Wu et al. "4d gaussian splatting for real-time dynamic scene rendering". In: *CVPR*. 2024, pp. 20310–20320.
[26] Weixing Xie et al. "Surgicalgaussian: Deformable 3d gaussians for high-fidelity surgical scene reconstruction". In: *MICCAI*. 2024, pp. 617–627.
[27] Chen Yang et al. "Efficient deformable tissue reconstruction via orthogonal neural plane". In: *IEEE Transactions on Medical Imaging* (2024).
[28] Chen Yang et al. "Neural lerplane representations for fast 4d reconstruction of deformable tissues". In: *MICCAI*. 2023, pp. 46–56.
[29] Lihe Yang et al. "Depth anything v2". In: *Advances in Neural Information Processing Systems* 37 (2025), pp. 21875–21911.
[30] Shuojue Yang et al. "Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting". In: *MICCAI*. 2024.
[31] Ruyi Zha et al. "Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos". In: *MICCAI*. 2023, pp. 13–23.
[32] Shuai Zhang et al. "A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images". In: *IEEE Transactions on Medical Robotics and Bionics* 3.1 (2020), pp. 85–95.
[33] Haoyin Zhou and Jagadeesan Jayender. "Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos". In: *MICCAI*. 2021, pp. 331–340.
[34] Zihan Zhu et al. "Nice-slam: Neural implicit scalable encoding for slam". In: *CVPR*. 2022, pp. 12786–12796.