Endo-GS*G*: Endoscopic Gaussian Splatting with Geometry-Awareness for Dynamic Tissue Reconstruction via Single-View Monocular Knowledge

Chao He[1] , **Kuangji Chen**[2] , **Bruce X.B. Yu**[2,*] , **Bo Lu**[1,*]
[1]Soochow University
[2]Zhejiang University
ansel.c.he@outlook.com, ckjjeang@zju.edu.cn, xinboyu@intl.zju.edu.cn, blu@suda.edu.cn

Abstract

Dynamic 3D reconstruction of surgical scenes plays a critical role in robotic-assisted surgery. Gaussian Splatting (GS), while effective for novel view synthesis, struggles to recover accurate surface from a monocular view due to the implicit multi-Gaussian representation of the surface. Specifically, (1) the vertical overlap of Gaussians leads to floating artifacts, and (2) the random orientation of Gaussians affects the smoothness of the reconstructed surface. Consequently, these fragmented and misaligned Gaussians hinder downstream applications, e.g., geometry-aware endoscopic navigation and physics-integrated tissue mechanics simulation. To this end, we propose Endo-GS*G*, a unified framework that couples dynamic Gaussian splatting with an SDF field for dynamic surface-aware tissue reconstruction. To further enhance geometric fidelity, we design geometry-informed regularization losses that constrain Gaussian density and spatial positioning. The entire pipeline is jointly supervised by RGB images and predicted depth maps, enabling high-quality reconstruction and rendering even with sparse or monocular input. Experiments on public datasets demonstrate that Endo-GS*G* outperforms state-of-the-art methods in both rendering quality and geometric surface accuracy.

1 Introduction

The rapid advancement of Artificial Intelligence (AI) is driving a global technological revolution [Ahmed *et al.*, 2022], with surgical robotics emerging as a fast-growing domain. Modern surgical robots are increasingly capable of precise task execution [Kim *et al.*, 2022], automated manipulation [Yip *et al.*, 2023], and human–robot collaboration [Fosch-Villaronga *et al.*, 2021] during specific surgical phases. To realize more intelligent image-guided operations and skill training, dynamic 3D surgical scene perception remains a long-standing challenge, particularly in providing robots with precise 3D awareness within the confined abdominal cavity.

*Corresponding author.

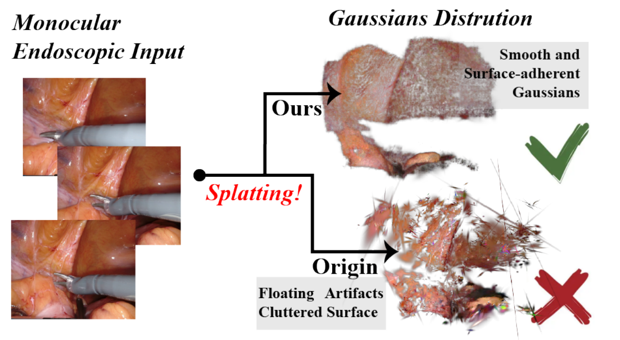

Figure 1: Dynamic scene reconstruction with instrument occlusion based on 3D-GS theory. The original method has numerous floating artifacts and unrealistic tissue structures, while our method produces smooth surface reconstruction.

3D Gaussian Splatting (3D-GS) has recently shown remarkable potential for high-fidelity view synthesis by representing scenes as collections of Gaussian ellipsoids optimized via multi-view image supervision [Kerbl *et al.*, 2023]. Its rasterization-based rendering pipeline enables real-time performance with high visual quality, making it well-suited for task guidance in robotic surgery. Nevertheless, 3D-GS relies solely on photometric supervision, resulting in poor 3D reconstruction quality. Although incorporating depth supervision offers partial improvements, the rendered surface—formed by aggregating multiple Gaussian ellipsoids—still produces inaccurate point clouds and exhibits noticeable floating artifacts. While dense-view supervision helps mitigate this issue, the inherently sparse-view setting in surgical applications fundamentally constrains 3D perception in dynamic scenes.

Recent studies have explored physics-integrated Gaussian representations [Xie *et al.*, 2023] [Zhang *et al.*, 2024] [Yang *et al.*, 2024a] [Zhong *et al.*, 2024], in which each Gaussian ellipsoid is regarded as an infinitesimal element of a physical object. Although these methods enable realistic simulation of soft tissue mechanics, they are still based on the inaccurate geometry produced by classical 3D-GS pipelines. Misaligned Gaussian ellipsoids lead to unrealistic physical responses, limiting their applicability to precise tissue modeling in surgical scenarios.

Method	Single View	Real-time	Dy-namic	Sur-face
NeuS [Wang *et al.*, 2021]	✗	✗	✗	✓
2D-GS [Huang *et al.*, 2024]	✗	✓	✗	✓
3D-GSR [Lyu *et al.*, 2024]	✗	✓	✗	✓
D-3D-GS [Yang *et al.*, 2024b]	✗	✓	✓	✗
Endo-GS [Zhu *et al.*, 2024]	✓	✓	✓	✗
Ours	✓	✓	✓	✓

Table 1: Comparison with SOTA methods regarding critical performance indexes for 3D reconstruction.

To resolve these problems, several state-of-the-art (SOTA) works have introduced surface regularization [Guédon and Lepetit, 2024], 2D-GS-based optimization [Huang *et al.*, 2024], and Gaussian surfels [Dai *et al.*, 2024] to improve 3D-GS reconstruction quality. However, these methods sacrifice fine-grained texture details and remain constrained by the inherent flaws of reconstruction methods. Although 3D-GSR [Lyu *et al.*, 2024] integrates SDF fields and loose coupling to align Gaussian ellipsoids with surfaces, it still struggles to generalize to dynamic scenes. Endo-GS [Zhu *et al.*, 2024] further introduces depth-fused pseudo-SDF supervision, yet lacks explicit geometric constraints on Gaussian opacity, limiting surface accuracy. A comparative analysis of our method and current SOTA methods is provided in Table 1.

To address these challenges, we propose Endo-GS*G*, a geometry-aware dynamic tissue reconstruction method based on 3D Gaussian Splatting. We decouple dynamic surgical scenes using a spatiotemporal plane, separating them into a static frame and corresponding Gaussian deformations, mapping all Gaussian kernel attributes consistently across the entire dynamic sequence. We design an SDF-Net integrated with hash encoding to output the distance of each Gaussian kernel to the surface and further propose three geometry-informed loss functions to enforce SDF-depth consistency, proximity to surface projections, and alignment with surface normals, ensuring accurate surface fitting and well-distributed Gaussian ellipsoids. Finally, instrument segmentation masks are incorporated to guide occlusion-aware optimization, effectively removing surgical tool interference and producing clean laparoscopic scene renderings.

In this work, we propose a geometry-informed regularization method to achieve surface alignment of Gaussian ellipsoids in single-view dynamic scenes. Our main contributions can be summarized as follows:

- We construct a spatiotemporal cross-plane dynamic representation field and propose an SDF-Net integrated with hash encoding, which is tightly coupled with the Gaussian splatting framework, to achieve dynamic reconstruction and precise implicit surface representation.
- We propose geometry-informed loss functions to regularize the Gaussian density on surfaces and spatial distribution of gaussian primitives, thereby achieving high-fidelity reconstructions of 3D surfaces.
- We integrate a temporal instrument segmentation mechanism into our optimization scheme, which can remove instrument occlusions and reveal covered laparoscopic scenes with globally consistent geometry awareness.

2 Related Work

2.1 Dynamic Tissue Reconstruction

Neural radiance fields (NeRF) [Mildenhall *et al.*, 2021] provide an implicit, continuous scene representation for reconstruction and novel view synthesis via volumetric rendering. For dynamic modeling, D-NeRF [Pumarola *et al.*, 2021] introduces a deformation field, while NeuS [Wang *et al.*, 2021] integrates SDF-based surface priors to enable higher-quality geometry extraction. In surgical scenes, Endo-NeRF [Wang *et al.*, 2022] leverages ray masking and depth supervision for single-view dynamic tissue reconstruction, and subsequent works further improve mesh extraction and robustness to unreliable depth [Zha *et al.*, 2023; Lu *et al.*, 2024]. Despite strong performance, NeRF-style methods remain computationally expensive due to dense sampling along rays, limiting practical efficiency in training and rendering.

2.2 Gaussian Splatting

With its efficient discretized representation and rendering capabilities, 3D-GS [Kerbl *et al.*, 2023] has rapidly advanced as a compelling alternative to neural radiance fields. Subsequent works [Oh *et al.*, 2025] [Liu *et al.*, 2024a] [Zhu *et al.*, 2024] [Yang *et al.*, 2024b] extended 3D-GS to support 3D/4D generation and dynamic scene modeling. By representing a scene as a collection of Gaussian ellipsoids, GS adjusts the color, opacity, and shape of each ellipsoid to faithfully capture scene textures. However, due to the fundamental differences between Gaussian ellipsoids and point clouds, there remains a gap in accurately representing scene geometry. Even when Gaussian ellipsoids are aligned along surfaces, they often fail to represent the true object geometry. This discrepancy leads to significant challenges when converting Gaussian-based representations to widely used discrete formats, such as meshes, thereby hindering downstream editing tasks.

Moreover, with the rise of world models [Genesis-Embodied-AI, 2024], 3D reconstruction is expected to move beyond texture and geometry, toward embedding physical properties that support physically plausible manipulation and interaction. Thereby, several works [Xie *et al.*, 2023] [Zhang *et al.*, 2024] [Zhong *et al.*, 2024] [Zong *et al.*, 2023] were introduced to embed physical attributes into the Gaussian representation. When utilized in surgical scenes, SimEndo-GS [Yang *et al.*, 2024a] showcased the potential of physics-integrated Gaussians in surgical applications by extending physical simulation and rendering into the endoscopic domain. However, 3D Gaussians inherently struggle to represent object surfaces accurately due to the flawed assumption that “Gaussian primitives represent the volumetric radiance rather than the explicit surface geometry”, making surface-aligned modeling a critical challenge.

To address this, SuGaR [Guédon and Lepetit, 2024] proposed limiting the overlap of Gaussians during 3D-GS optimization. Subsequently, a mesh can be extracted using Poisson reconstruction [Kazhdan and Hoppe, 2013] , which is

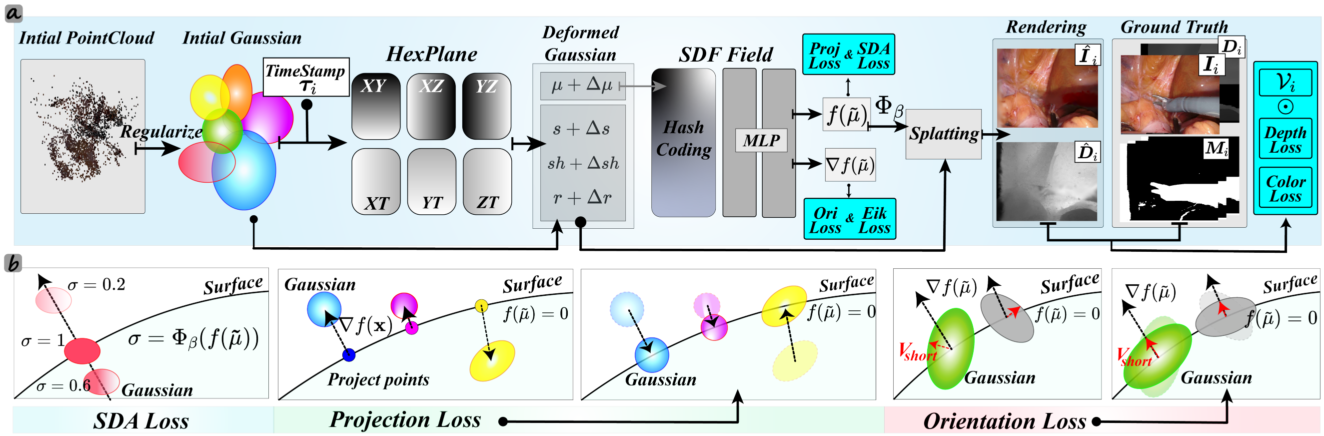

Figure 2: (a). The workflow of our model. The initial point cloud is regularized with depth maps and initialized as Gaussian ellipsoids. Using Hexplane, we model tissue deformation by predicting the temporal properties of the Gaussian ellipsoids, and map the coordinates to the SDF network for density prediction via hash encoding. Finally, the Gaussian ellipsoids are rasterized by calculating photometric and depth errors. Three geometry-informed loss terms jointly optimize the pipeline. (b). Sketches of SDA, projection, and orientation loss. SDA loss optimizes the SDF network's learning of the surface with a pseudo-SDF plane, refining the Gaussian sphere density. Projection and orientation losses pull the Gaussian ellipsoids to the learned surface, leading to a better geometry awareness of dynamic 3D scenarios.

then refined by binding flat Gaussians distributed on its surface and optimizing for accuracy. 2D-GS [Huang *et al.*, 2024] treats points as 2D surfaces and uses TSDF fusion to produce the final mesh. 3D-GSR [Lyu *et al.*, 2024] introduced the use of an SDF field with a loosely coupled strategy to regularize the position and orientation of Gaussian ellipsoids. While these methods have shown effectiveness, they are not applicable to dynamic scene surface reconstruction, especially in medical scenarios with limited viewpoints. Since the above methods heavily rely on multi-view information to drive surface learning, they tend to produce unsatisfactory reconstruction results in viewpoint-limited and unstructured medical scenarios, where multi-view data is unavailable and surgical instruments often occlude the scene.

In our framework, each Gaussian ellipsoid is modeled as a ***physical micro-element*** representing matter in the real world. Physical interactions among these elements are simulated through the Material Point Method (MPM) [Jiang *et al.*, 2016], enabling accurate modeling of deformation and contact dynamics. Concurrently, a rasterization-based pipeline facilitates real-time rendering of these interactions. This integration of physically grounded simulation and efficient rendering holds promise for advancing surgical robotics and other downstream applications that demand precise, real-time ***visuo-physical*** modeling.

3 Proposed Method

Figure 2 shows an overview of our proposed Endo-GS$\mathcal{G}$. We define input as $\{(\boldsymbol{I}_i, \boldsymbol{D}_i, \boldsymbol{M}_i, \mathbf{P}_i, \tau_i)\}_{i=1}^{T}$, where, $\mathbf{I}_i$ represents the RGB image of the i-th frame, $\boldsymbol{D}_i$ is the corresponding depth map estimated from monocular endoscopy, and $\mathbf{M}_i$ is a binary mask that filters out surgical tools. The projection matrix $\mathbf{P}_i$ encodes the transformation from world coordinates to image-plane coordinates, which remains fixed across all frames under the single-view setting. The timestamp τ_i denotes the temporal index of each frame and is normalized as i/T. To optimize the Gaussians more effectively, we use the tuple $(\boldsymbol{I}_i, \boldsymbol{D}_i, \boldsymbol{M}_i, \mathbf{P}_i, \tau_i)_{i=1}$ to regularize the initial point cloud from the first frame, which serves as the initialization for Gaussian modeling. The remainder of this section introduces detailed steps to achieve our tissue surface reconstruction pipeline.

3.1 Preliminaries

3D-GS [Kerbl *et al.*, 2023] models the scene as learnable 3D Gaussians $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}i, \alpha_i, c_i)$, which are projected to screen space and accumulated by front-to-back alpha compositing. The rendered color and depth at pixel u share the same compositing weights:

$$\begin{bmatrix} C(u) \\ D(u) \end{bmatrix} = \sum_i T_i\, \alpha_i \begin{bmatrix} c_i \\ d_i \end{bmatrix}, \qquad T_i = \prod_{j<i} (1-\alpha_j). \tag{1}$$

where d_i is the depth of $\boldsymbol{\mu}_i$. For completeness, we briefly summarize the standard 3D-GS rendering formulation; detailed derivations and implementation specifics can be found in [Kerbl *et al.*, 2023].

3.2 3D-GS for Dynamic Soft Tissues

To model dynamic soft tissue deformation induced by instrument traction, we represent the scene as a four-dimensional volume where a deformation network is introduced to capture temporal tissue motion. Inspired by HexPlane [Cao and Johnson, 2023], we employ six orthogonal spatio-temporal feature planes to jointly encode spatial and temporal information: three spatial planes (XY, XZ, YZ) and three spatio-temporal planes (XT, YT, ZT), denoted as $\mathcal{P} = \{(x, y), (y, z), (x, z), (x, \tau), (y, \tau), (z, \tau)\}$. Each plane has a resolution of $N_1 \times N_2$ and feature tensor $F \in \mathbb{R}^{d \times N_1 \times N_2}$, where d is the hidden feature dimension. For each Gaussian, we use its center $\boldsymbol{\mu}(x, y, z)$ and timestamp τ to bilinearly interpolate $\mathcal{B}(\cdot)$ on the four nearest voxels of each plane, aggregating the results as the final voxel feature:

$$\mathbf{f}_{\text{voxel}}(\boldsymbol{\mu}, \tau) = \text{Interp}\left(\{\mathcal{B}(F_{pq}, p, q) \mid (p, q) \in \mathcal{P}\}\right). \quad (2)$$

This feature is passed through a single-layer MLP to predict the temporal variations of Gaussian attributes: $\{(\Delta\mu, \Delta s, \Delta r, \Delta sh)\}$, which are treated as time-dependent terms. The deformed attributes are given by $\{(\mu + \Delta\mu,\ s + \Delta s,\ r + \Delta r,\ sh + \Delta sh)\}$, and we denote them as $\{(\tilde{\mu}, \tilde{s}, \tilde{r}, \tilde{sh})\}$ for brevity in later formulations.

3.3 Coupling Between Gaussian and SDF

3D-GS synthesizes novel views by directly rasterizing all Gaussian ellipsoids. However, this strategy lacks the ability to capture meaningful surface geometry. In single-viewpoint scenarios, the absence of multi-view constraints often leads to a disordered spatial distribution of Gaussians, despite the rendered images appearing visually plausible. To overcome this limitation, we integrate a neural implicit Signed Distance Function (SDF) field into the pipeline, which serves to regularize the spatial arrangement of Gaussians and ensure their alignment with the true tissue surface. The neural SDF is defined as a function $f : \mathbb{R}^3 \to \mathbb{R}$, which maps any 3D point in $\mathbb{R}^3$ to its signed distance from the target surface. This function is realized using a multi-resolution hash grid for efficient spatial encoding, coupled with a 2-layer MLP for regression. The underlying surface of the scene, denoted by $\mathcal{S}$, can be extracted as the zero level set of the SDF and is formally defined as:

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\}, \quad (3)$$

Next, we establish the coupling between the SDF field and the Gaussian ellipsoids. To enable effective surface learning for the SDF network, NeuS [Wang *et al.*, 2021] proposes converting the SDF values into densities of ray-sampled points in a neural radiance field, which are then integrated into a volumetric rendering framework for supervision. However, in single-view scenarios, the rendered color is not a reliable surface supervision signal due to its ambiguity. To address this, we evaluate the SDF at the Gaussian centers $\tilde{\mu}$, i.e., $f(\tilde{\mu}) \to$ SDF values, then convert these values into Gaussian densities using a differentiable bell-shaped function, and finally render the depth using Equation (5). The rendered depth $D'(u)$ is then compared with the ground-truth depth $D(u)$ to provide meaningful surface learning supervision. Specifically, the bell-shaped function $\Phi_\beta(f(\tilde{\mu}))$ is controlled by a learnable parameter β, which is:

$$\Phi_\beta(f(\tilde{\mu})) = 4 \cdot \frac{e^{-\beta f(\tilde{\mu})}}{\left(1 + e^{-\beta f(\tilde{\mu})}\right)^2}. \quad (4)$$

Through this function, we obtain the density of each Gaussian and encourage those closer to the zero level set to contribute more significantly to the overall density. This aligns with our core principle that Gaussians should be concentrated near the object surface. Moreover, the SDF field can be updated under the supervision of depth loss:

$$\mathcal{L}_d(u) = \boldsymbol{D}(u) - \hat{\boldsymbol{D}}(u), \quad (5)$$

While depth supervision facilitates Gaussian density attenuation near surface regions, the inherent structural ambiguity in monocular settings hinders stable optimization. To address this, we introduce pseudo ground-truth SDF values derived from depth maps to supervise the SDF network, ensuring SDF values are aligned with depth, which is:

$$\mathcal{L}_{\text{SDA}} = \left|\mathbf{S}_{\text{target}} - \mathbf{S}_{\text{pred}}\right|, \quad \mathbf{S}_{\text{target}} = \frac{\text{map}_z - z_{\text{cam}}}{\|\text{map}_z - z_{\text{cam}}\|}, \quad (6)$$

where, $\mathbf{S}_{\text{pred}}$ is the normalized predicted SDF value, while $\mathbf{S}_{\text{target}}$ is the target SDF direction computed as the normalized depth difference between the ground truth depth map map_z and the Gaussian center's depth z_{cam} in the camera coordinate system. The predicted SDF value $\mathbf{S}_{\text{pred}}$ is trained to approximate $\mathbf{S}_{\text{target}}$, encouraging SDF consistency in depth alignment. This strategy effectively regularizes the SDF distribution under limited multi-view information in laparoscopic scenes. Combined with Equation (5), it suppresses the density of Gaussians far from the surface.

To further mitigate floating artifacts caused by off-surface Gaussians and ensure smooth distribution along the surface [Lyu *et al.*, 2024], we leverage the SDF field to both attract Gaussian centers toward the surface and adjust their orientations, ensuring that all ellipsoids are distributed in a scale-like pattern along the zero level set. Specifically, we perform a first-order Taylor expansion of the SDF field at the Gaussian center $\tilde{\mu}$ to calculate the nearest projection point $\mathbf{x}_g$ onto the surface:

$$\mathbf{x}_{\text{proj}} = \tilde{\mu} + f(\tilde{\mu}) \cdot \nabla f(\tilde{\mu}), \quad (7)$$

where $f(\tilde{\mu})$ denotes the SDF value and $\nabla f(\tilde{\mu})$ is the gradient of the SDF at $\tilde{\mu}$, normalized to a unit vector. To ensure that all Gaussians lie exactly on the surface, we aim to enforce $\mathbf{x}_{\text{proj}} = \mathbf{x}_g$. During optimization, to encourage directional gradient flow, we define a surface projection loss that effectively pulls Gaussian centers toward the surface:

$$\mathcal{L}_{\text{proj}} = \mathbb{E}_{\tilde{\mu}}\left[\|\mathbf{x}_{\text{proj}} - \tilde{\mu}\|_1\right]. \quad (8)$$

Due to their anisotropic nature with distinct long and short axes, neglecting orientation consistency may result in slender ellipsoids generating fuzzy artifacts on the surface. Therefore, to ensure that each Gaussian ellipsoid is aligned with the local surface normal, we define an orientation loss by enforcing the shortest principal axis of the ellipsoid to be parallel to the SDF gradient $\nabla f(\tilde{\mu})$. Specifically, given the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ derived from the quaternion representation, and the scale vector $\mathbf{s} = (s_x, s_y, s_z)$, we identify the local axis corresponding to the smallest scale:

$$\mathbf{v}_{\text{short}} = \mathbf{R}^\top \cdot \mathbf{e}_{\arg\min(s_x, s_y, s_z)}, \quad (9)$$

where $\mathbf{e}_k$ denotes the k-th canonical basis vector. Let $\mathbf{n} = \nabla f(\tilde{\mu}) / \|\nabla f(\tilde{\mu})\|$ be the unit surface normal at the Gaussian center. To avoid the singularity that may caused by directly enforcing the dot product between $\mathbf{v}_{\text{short}}$ and $\mathbf{n}$ to be 1 during optimization, we define the orientation loss as:

$$\mathcal{L}_{\text{ori}} = \mathbb{E}_{\mathbf{x}_g}\left[\left(\cos^{-1}(\langle \mathbf{v}_{\text{short}}, \mathbf{n}\rangle) - 0\right)^2\right], \quad (10)$$

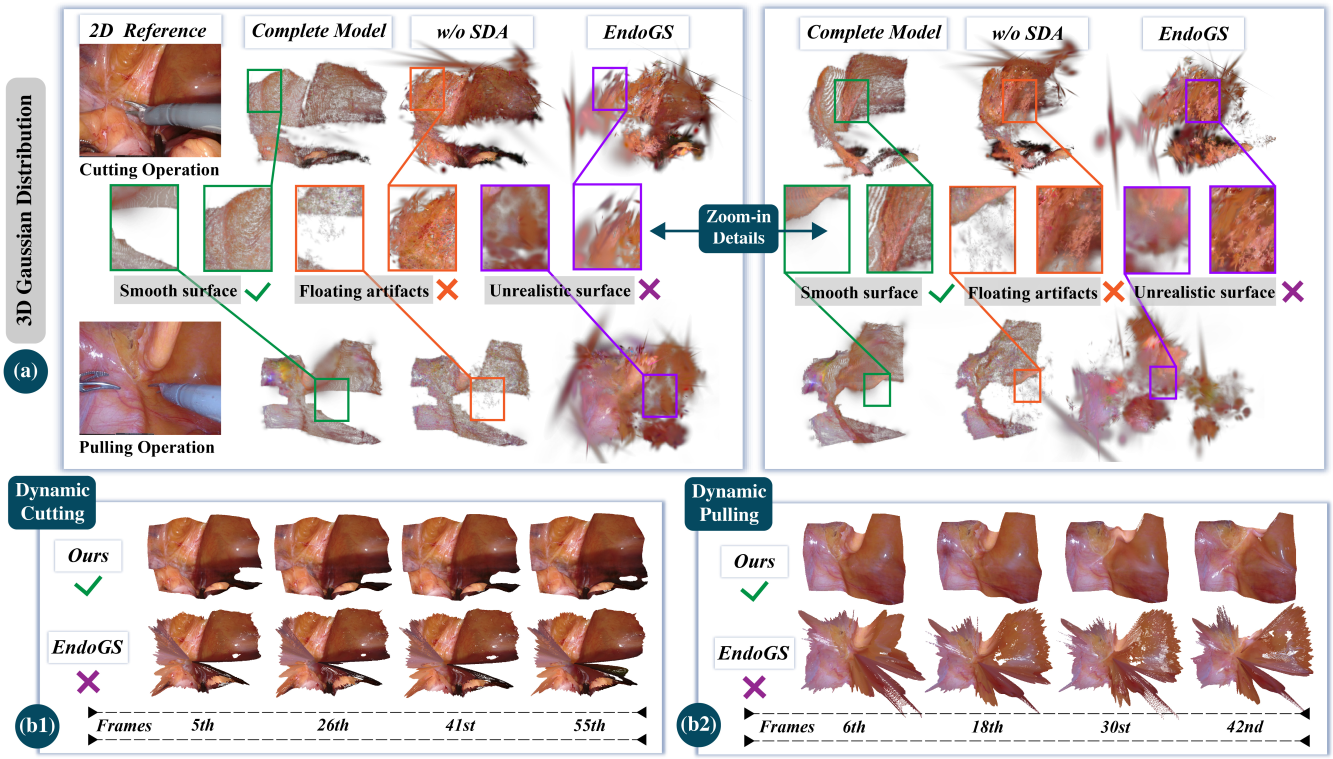

Figure 3: (a). Qualitative visualization of 3D Gaussian distributions on the Cutting and Pulling datasets. From left to right: our complete model, a set of ablations, and a comparison with Endo-GS reconstruction results. Our method exhibits precise surface structures, whereas Endo-GS generates numerous floating artifacts and unrealistic tissue surfaces. (b1)-(b2). Dynamic mesh and rendering of cutting and pulling procedures based on our Endo-GS*G* and the SOTA Endo-GS approach.

3.4 Optimization

Occlusions from surgical instruments pose a major challenge in tissue reconstruction. To mitigate this, we incorporate binary masks to identify occluded pixels and introduce a spatiotemporal importance sampling strategy that emphasizes frequently occluded regions. Let $\{\boldsymbol{M}_i\}_{i=1}^{T}$, where $\boldsymbol{M}_i \in \{0,1\}$, denote the binary occlusion masks across frames, with 0 for visible tissue and 1 for occluded areas. We define the corresponding importance maps $\{\boldsymbol{V}_i\}_{i=1}^{T}$ as:

$$\boldsymbol{V}_i = (1 - \boldsymbol{M}_i) \odot \left(1 + \gamma \cdot \frac{\sum_{j=1}^{T} \boldsymbol{M}_j}{\left\|\sum_{j=1}^{T} \boldsymbol{M}_j\right\|_F}\right), \tag{11}$$

where $\odot$ denotes element-wise multiplication. The term $1 - \boldsymbol{M}_i$ restricts optimization to visible regions, while the normalized sum encodes the temporal occlusion frequency, prioritizing tissue regions that are frequently hidden. The weighting strength is controlled by the hyperparameter γ.

To train the 3D-GS model, we begin by optimizing the Gaussian attributes via minimizing both photometric and depth losses. The color loss, denoted as $\mathcal{L}_c$, measures the difference between the rendered image and the ground-truth image. To suppress the influence of surgical instruments, we multiply both the photometric and depth losses by an importance map $\boldsymbol{V}_i$, which is derived from the instrument masks. Specifically, the photometric and depth losses are denoted as:

$$\begin{cases} \mathcal{L}_c(i) = \left|\boldsymbol{I}_i \odot \boldsymbol{V}_i - \hat{\boldsymbol{I}}_i \odot \boldsymbol{V}_i\right|, \\ \mathcal{L}_d(i) = \left|\boldsymbol{D}_i \odot \boldsymbol{V}_i - \hat{\boldsymbol{D}}_i \odot \boldsymbol{V}_i\right|. \end{cases} \tag{12}$$

where $\boldsymbol{I}_i$ is the ground-truth RGB image, $\hat{\boldsymbol{I}}_i$ is the rendered image, and $\odot$ denotes element-wise multiplication. The depth loss $\mathcal{L}_d$ is jointly optimized with the Eikonal loss $\mathcal{L}_{\text{eikonal}}$ to encourage the SDF field to learn accurate surface geometry. The Eikonal loss enforces the gradient of the SDF to remain unit-norm and is defined as:

$$\mathcal{L}_{\text{eikonal}} = \mathbb{E}_{\tilde{\mu}}\left[\left(\|\nabla f(\tilde{\mu})\|_2 - 1\right)^2\right], \tag{13}$$

Finally, to align the Gaussian ellipsoids precisely with the surface, we incorporate both the projection loss $\mathcal{L}_{\text{proj}}$ and orientation consistency loss $\mathcal{L}_{\text{ori}}$. The total loss for the framework is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_d + \lambda_3\mathcal{L}_{\text{eikonal}} + \lambda_4\mathcal{L}_{\text{proj}} + \lambda_5\mathcal{L}_{\text{ori}} + \lambda_6\mathcal{L}_{\text{SDA}} \tag{14}$$

4 Experiments

4.1 Implementation Details

We evaluate our method on two public datasets [Wang *et al.*, 2022]: Cutting, with tissue incisions by surgical forceps,

Methods	*Rendering and Speed↦Cutting*					*Rendering and Speed↦Pulling*				
	PSNR⇑	*SSIM⇑*	*LPIPS⇓*	*FPS⇑*	*Points*	*PSNR⇑*	*SSIM⇑*	*LPIPS⇓*	*FPS⇑*	*Points*
NeRF [Mildenhall *et al.*, 2021]	26.943	0.910	0.153	-	-	27.058	0.903	0.152	-	-
Endo-NeRF [Wang *et al.*, 2022]	34.186	0.932	0.151	<0.2	-	34.981	0.938	0.161	<0.2	-
Endo-Surf [Zha *et al.*, 2023]	34.981	0.953	0.106	<0.2	-	35.004	0.956	0.120	<0.2	-
3D-GS [Kerbl *et al.*, 2023]	27.044	0.829	0.151	-	141458	27.432	0.911	0.152	-	139278
D-3D-GS [Yang *et al.*, 2024b]	37.028	0.963	0.035	**98.2**	88130	37.224	0.961	0.036	**110.2**	79280
Endo-GS [Zhu *et al.*, 2024]	36.490	0.963	0.039	90.2	49538	37.729	**0.968**	0.038	101.8	43644
Ours *w/o* SDA	38.132	0.970	0.023	47.1	201069	37.970	0.967	0.036	62.2	168432
Ours (complete)	**38.857**	**0.974**	**0.015**	45.4	**231543**	**38.240**	**0.968**	**0.035**	61.9	**173839**

Table 2: Comparison of rendering quality (PSNR, SSIM, LPIPS), computational efficiency (FPS), and model size (Points) on Cutting and Pulling datasets. The best and second-best results are highlighted in pink and orange, respectively.

Methods	*Reconstruction↦Cutting*					*Reconstruction↦Pulling*				
	CD⇓	*P2S-M⇓*	*P2S-Md⇓*	*P2S-95⇓*	*F-score⇑*	*CD⇓*	*P2S-M⇓*	*P2S-Md⇓*	*P2S-95⇓*	*F-score⇑*
D-3D-GS [Yang *et al.*, 2024b]	4.41e-02	1.43e-01	1.78e-01	2.14e-01	0.89e-01	3.24e-02	2.01e-01	2.13e-01	2.41e-01	1.20e-01
Endo-GS [Zhu *et al.*, 2024]	1.15e-03	2.35e-02	2.40e-02	3.43e-02	1.64e-01	2.87e-03	2.87e-02	9.03e-03	1.13e-01	4.68e-01
Ours *w/o* SDA	3.00e-06	7.27e-04	2.23e-04	4.19e-03	9.97e-01	1.30e-05	1.13e-03	2.33e-04	9.67e-03	9.75e-01
Ours (complete)	**0.00e+00**	**1.86e-04**	**1.27e-04**	**4.27e-04**	**1.00e+00**	**0.00e+00**	**2.90e-05**	**2.50e-05**	**6.10e-05**	**1.00e+00**

Table 3: Comparison of geometric reconstruction on the Cutting and Pulling datasets. CD, P2S-M, P2S-Md, P2S-95, and F denote Chamfer distance, P2S mean, P2S median, P2S 95%, and F-score, respectively. All metrics are computed with a unified distance threshold of 1 mm. The best and second-best results are also highlighted.

Metric	*w/o* Depth	*w/o* Proj&Ori	*w/o* SDA	Complete
P2S Mean	2.01e-03	4.54e-04	7.27e-04	**1.86e-04**
P2S 95%	3.54e-02	2.25e-03	4.19e-03	**4.27e-04**
F-score	1.63e-01	9.97e-01	9.97e-01	**1.00e+00**
PSNR	36.429	37.836	38.132	**38.857**

Table 4: Ablation study of surface reconstruction and rendering quality. The best performance is highlighted in bold.

and Pulling, with upward tissue traction. Both monocular sequences contain instrument occlusions and complex non-rigid deformations. Following the dataset setup [Wang *et al.*, 2022], stereo-derived depth maps D_i and manually annotated binary masks M_i are used as pseudo supervision. The model is trained for 30,000 iterations on an RTX 4090 GPU with initial weights $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$, $\lambda_4 = 1$, $\lambda_5 = 0.5$, and $\lambda_6 = 0.1$.

4.2 Evaluation Analysis

We comprehensively compare our method with state-of-the-art approaches, including NeRF [Mildenhall *et al.*, 2021], Endo-NeRF [Wang *et al.*, 2022], Endo-Surf [Zha *et al.*, 2023], 3D-GS [Kerbl *et al.*, 2023], D-3D-GS [Yang *et al.*, 2024b], and Endo-GS [Zhu *et al.*, 2024], in terms of image rendering and 3D reconstruction. For rendering evaluation, we report PSNR, SSIM [Wang *et al.*, 2004], and LPIPS [Zhang *et al.*, 2018] to quantify visual fidelity, and measure inference speed using FPS. For Gaussian-based methods, we also track the number of 3D Gaussians to monitor splitting and attenuation, which affect rendering efficiency and model compactness. For 3D reconstruction, we assess the alignment between predicted Gaussian point clouds and RGB-D-derived ground-truth surfaces using Chamfer Distance [Fan *et al.*, 2017], point-to-surface (P2S) metrics [Ma *et al.*, 2020], and F-score [Tatarchenko *et al.*, 2017]. Chamfer Distance measures global geometric accuracy; P2S metrics (mean, median, 95th percentile) evaluate fidelity and robustness; and F-score captures surface-level correspondence through precision and recall. Table 2 and Table 3 present the quantitative results on rendering and reconstruction respectively. Figure 4 shows qualitative comparisons against Endo-GS, while Figure 3 visualizes multi-view Gaussian-based surface reconstructions, with the rendering outcomes under dynamic surgical scenes.

Our method, leveraging a dynamic representation network and a hash-table encoded SDF-Net, achieves superior rendering quality over existing SOTA methods, including Endo-GS [Zhu *et al.*, 2024]. This advantage mainly stems from the strong fitting capability of hash encoding and the improved geometric interpretability of SDF-based surface modeling. With SDA, orientation, and projection losses, our method aligns Gaussians with the underlying tissue surface while preserving texture details. Although Endo-GS [Zhu *et al.*, 2024] was trained for 60,000 iterations, our model achieves competitive or better performance with only 30,000 iterations, demonstrating faster convergence and lower training cost, as shown in Figure 5. Moreover, instrument segmentation masks enable dynamic instrument removal, and aligned comparisons show rendering quality comparable to or better than Endo-GS. In contrast, D-3D-GS [Yang *et al.*, 2024b] achieves fast rendering speed with its lightweight dynamic network,

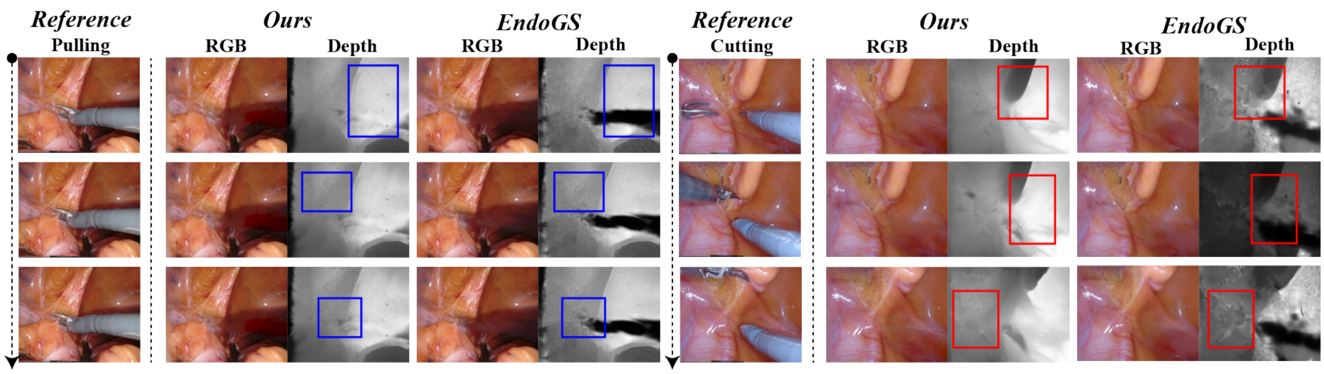

Figure 4: 2D rendering results of the model on the Cutting and Pulling datasets. The rendered RGB and depth maps are shown. Both methods achieve high-quality RGB renderings, but Endo-GS suffers from significant depth noise due to its lack of surface learning. In contrast, our method produces smooth and consistent depth renderings.

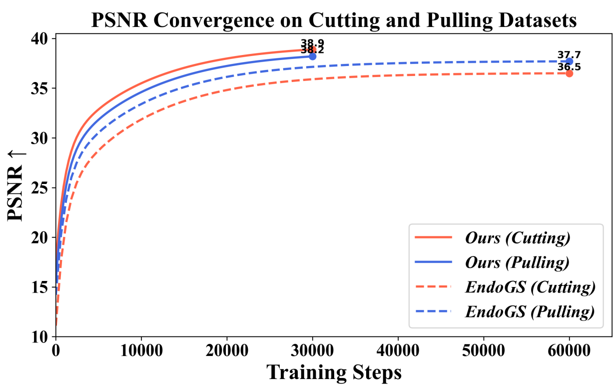

Figure 5: Curves showing the number of training iterations versus PSNR on the Cutting and Pulling datasets. Our model demonstrates excellent performance after 30,000 iterations, highlighting its stronger learning convergence capability.

but lacks depth guidance and effective geometric constraints, resulting in less consistent geometry and unstable surface reconstruction. 3D-GS [Kerbl *et al.*, 2023], originally designed for static scenes, struggles with highly dynamic tissue deformations, leading to degraded reconstruction performance. NeRF-based methods, limited by ray-tracing-based rendering, consistently underperform across all metrics.

Our method focuses on tightly adhering Gaussian ellipsoids to the surface, which may sacrifice rendering some high-frequency texture details. To capture finer details, our model generates additional Gaussian ellipsoids near the surface, resulting in an increased number of points, which causes slower rendering speed. Nevertheless, after optimization, our method still satisfies the commonly adopted real-time criterion in laparoscopic video analysis, i.e., $\geq$30 FPS [Liu *et al.*, 2024b]. In 3D reconstruction accuracy (Table 3), our method significantly outperforms D-3D-GS [Yang *et al.*, 2024b] and Endo-GS [Zhu *et al.*, 2024], owing to our specialized SDF-Net pipeline and joint loss function, which optimize Gaussian sphere representation for tissue surfaces, leading to superior accuracy in 3D structural reconstruction. D-3D-GS [Yang *et al.*, 2024b] lacks depth supervision, leading to more chaotic and weaker 3D structural reconstruction compared to Endo-GS [Zhu *et al.*, 2024].

Qualitative comparisons in Figure 3(a) show that our method tightly adheres Gaussian ellipsoids to the surface, reducing floating artifacts. Figures 3(b1)–(b2) demonstrate more accurate dynamic structure perception than SOTA methods under occlusion and sparse visual feedback, benefiting surgical navigation and physics-based simulation. Figure 4 further shows that our model handles instrument-induced occlusions effectively, recovering tissue textures and producing more accurate depth maps than Endo-GS [Zhu *et al.*, 2024].

4.3 Ablation Study

We performed an ablation study comparing ”Ours *w/o* SDA” (using only proj and ori losses) with the ”Complete model” incorporating all losses. As shown in Table 2, adding SDA loss enables SDF-Net to better capture surface information, improving rendering quality. Figure 3(a) and Table 3 show that ”Ours *w/o* SDA” exhibits artifacts and misalignment, while SDA loss effectively aligns the Gaussian distribution with the true surface geometry. Table 4 further demonstrates that the complete model achieves optimal performance in both rendering and 3D structure.

5 Conclusion

This work proposed Endo-GS*G*, a dynamic 3D reconstruction framework that couples 3D-GS with an SDF field to improve Gaussian distributions for tissue structures in unstructured surgical scenes. With a spatiotemporal deformation field, implicit SDF alignment, and geometry-informed regularization losses, Endo-GS*G* captures tissue dynamics and aligns Gaussians with the underlying surface under sparse-view endoscopic conditions. Experiments on complex real-world scenarios demonstrate superior rendering quality and surface accuracy over SOTA methods. Endo-GS*G* provides a surface-faithful representation for postoperative surgical training and physics-integrated GS with MPM in future surgical intelligence systems, while extreme topology-breaking events and broader clinical validation remain future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62573310) and (62502445), in part by the Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems, in part by the Innovation and Entrepreneurship Leading Talent Program of Suzhou City (ZXL2023156), and in part by the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001). The authors would also like to thank Suzhou Jupiter Semi Co., Ltd. for its support of this work.

Contribution Statement

Chao He and KuangJi Chen contributed equally to this work and are considered co-first authors. Bruce X.B. Yu and Bo Lu are the corresponding authors.

References

[Ahmed *et al.*, 2022] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE transactions on industrial informatics*, 18(8):5031–5042, 2022.

[Cao and Johnson, 2023] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023.

[Dai *et al.*, 2024] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.

[Fan *et al.*, 2017] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[Fosch-Villaronga *et al.*, 2021] Eduard Fosch-Villaronga, Pranav Khanna, Hadassah Drukarch, and Bart HM Custers. A human in the loop in surgery automation. *Nature Machine Intelligence*, 3(5):368–369, 2021.

[Genesis-Embodied-AI, 2024] Genesis-Embodied-AI. Genesis: A generative and universal physics engine for robotics and beyond. https://github.com/Genesis-Embodied-AI/Genesis, December 2024. Accessed: 2026-05-18.

[Guédon and Lepetit, 2024] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5354–5363, 2024.

[Huang *et al.*, 2024] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH '24, page 1–11. ACM, July 2024.

[Jiang *et al.*, 2016] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*, pages 1–52. 2016.

[Kazhdan and Hoppe, 2013] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

[Kim *et al.*, 2022] Yoonho Kim, Emily Genevriere, Pablo Harker, Jaehun Choe, Marcin Balicki, Robert W Regenhardt, Justin E Vranic, Adam A Dmytriw, Aman B Patel, and Xuanhe Zhao. Telerobotic neurovascular interventions with magnetic manipulation. *Science Robotics*, 7(65):eabg9907, 2022.

[Liu *et al.*, 2024a] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. Endogaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

[Liu *et al.*, 2024b] Yuzhang Liu, Yuichiro Hayashi, Masahiro Oda, Takayuki Kitasaka, and Kensaku Mori. Yolov7-repfpn: Improving real-time performance of laparoscopic tool detection on embedded systems. *Healthcare Technology Letters*, 11(2-3):157–166, 2024.

[Lu *et al.*, 2024] Bo Lu, Chao He, Wenjie Hou, Lijun Han, Hesheng Wang, Lining Sun, Zhaolei Jiang, and Peng Qi. Endo-4srf: Learning radiance field for dynamic surface reconstruction of surgical tissues with obstacle stealth under single-view and depth-free monocular endoscopy. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[Lyu *et al.*, 2024] Xiaoyang Lyu, Yang-Tian Sun, Yi-Hua Huang, Xiuzhe Wu, Ziyi Yang, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. 3dgsr: Implicit surface reconstruction with 3d gaussian splatting. *ACM Transactions on Graphics*, 43(6):1–12, November 2024.

[Ma *et al.*, 2020] Baorui Ma, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. *arXiv preprint arXiv:2011.13495*, 2020.

[Mildenhall *et al.*, 2021] 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*, 65(1):99–106, 2021.

[Oh *et al.*, 2025] Seungjun Oh, Younggeun Lee, Hyejin Jeon, and Eunbyung Park. Hybrid 3d-4d gaussian splatting for fast dynamic scene representation, 2025.

[Pumarola *et al.*, 2021] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[Tatarchenko *et al.*, 2017] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE international conference on computer vision*, pages 2088–2096, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2021] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[Wang *et al.*, 2022] 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*, pages 431–441. Springer, 2022.

[Xie *et al.*, 2023] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023.

[Yang *et al.*, 2024a] Zhenya Yang, Kai Chen, Yonghao Long, and Qi Dou. Simendogs: Efficient data-driven scene simulation using robotic surgery videos via physics-embedded 3d gaussians. *arXiv preprint arXiv:2405.00956*, 2024.

[Yang *et al.*, 2024b] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024.

[Yip *et al.*, 2023] Michael Yip, Septimiu Salcudean, Ken Goldberg, Kaspar Althoefer, Arianna Menciassi, Justin D Opfermann, Axel Krieger, Krithika Swaminathan, Conor J Walsh, He Huang, et al. Artificial intelligence meets medical robotics. *Science*, 381(6654):141–146, 2023.

[Zha *et al.*, 2023] 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*, pages 13–23. Springer, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2024] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2024.

[Zhong *et al.*, 2024] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pages 407–423. Springer, 2024.

[Zhu *et al.*, 2024] Lingting Zhu, Zhao Wang, Jiahao Cui, Zhenchao Jin, Guying Lin, and Lequan Yu. Endogs: Deformable endoscopic tissues reconstruction with gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 135–145. Springer, 2024.

[Zong *et al.*, 2023] Zeshun Zong, Xuan Li, Minchen Li, Maurizio M Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural stress fields for reduced-order elastoplasticity and fracture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.