

SAGS: Self-Adaptive Alias-Free Gaussian Splatting for Dynamic Surgical Endoscopic Reconstruction

Wenfeng Huang^{a,b}, Xiangyun Liao^b, Yinling Qian^b, Hao Liu^c, Yongming Yang^c, Wenjing Jia^a, Qiong Wang^{b,*}

^a*University of Technology Sydney, Australia*

^b*Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China*

^c*Shenyang Institute of Automation, Chinese Academy of Sciences, China*

Abstract

Surgical reconstruction of dynamic tissues from endoscopic videos is a crucial technology in robot-assisted surgery. The development of Neural Radiance Fields (NeRFs) has greatly advanced deformable tissue reconstruction, achieving high-quality results from video and image sequences. However, reconstructing deformable endoscopic scenes remains challenging due to aliasing and artifacts caused by tissue movement, which can significantly degrade visualization quality. The introduction of 3D Gaussian Splatting (3DGS) has improved reconstruction efficiency by enabling a faster rendering pipeline. Nevertheless, existing 3DGS methods often prioritize rendering speed while neglecting these critical issues. To address these challenges, we propose SAGS, a self-adaptive alias-free Gaussian splatting framework. We introduce an attention-driven, dynamically weighted 4D deformation decoder, leveraging 3D smoothing filters and 2D Mip filters to mitigate artifacts in deformable tissue reconstruction and better capture the fine details of tissue movement. Experimental results on two public benchmarks, EndoNeRF and SCARED, demonstrate that our method achieves superior performance in all metrics of PSNR, SSIM, and LPIPS compared to the state of the art while also delivering better visualization quality.

Keywords: Surgical Reconstruction, Endoscopic Reconstruction, 3D Gaussian Splatting, Self-adaptive Decoder

1. Introduction

3D reconstruction of deformable tissue structures from dynamic endoscopic videos represents an essential cornerstone in modern robotic-assisted surgical interventions, significantly enhancing navigation, precision, and overall patient outcomes [48]. High-quality, real-time 3D reconstructions facilitate numerous

*Corresponding author: Yinling Qian and Qiong Wang (Email: wangqiong@siat.ac.cn)

intraoperative clinical applications, including augmented reality (AR)-based visualization, robotic surgery automation, immersive surgical training, and precise surgical planning [18, 19, 52].

Early developments in the medical scene reconstruction primarily relied on conventional depth estimation techniques and simultaneous localization and mapping (SLAM)-based frameworks [31]. Classic methods such as E-DSSR [20] and Surfelwarp [8] established preliminary successes by integrating stereo depth cues and dynamic surface tracking. However, these approaches struggle to accurately handle severe non-rigid tissue deformations, significant occlusion by surgical instruments, and complex dynamics typically encountered in real surgical environments [52, 48]. This fundamental limitation motivated further exploration into more robust and scalable representations.

The emergence of Neural Radiance Fields (NeRFs) [22] significantly shifted the paradigm toward implicit volumetric representations, achieving photorealistic quality in novel-view synthesis and continuous 3D scene modelling [37]. NeRF leverages multi-layer perceptrons (MLPs) to implicitly encode volumetric densities and radiance, thereby attaining markedly higher visual fidelity than traditional discrete approaches. Dynamic extensions—such as Dynamic NeRF (D-NeRF) [28], Neural Volumes [28], and Temporal-Interpolation NeRF (TiNeuVox) [6]—further generalise this framework to temporally evolving scenes, while LerPlane [39] reduces complexity by factorising the volume into a set of explicit planes, accelerating optimisation and improving near-real-time applicability. Within the medical domain, EndoNeRF [33] represents the first attempt to apply NeRF to 3D surgical reconstruction. By integrating a static radiance field with a temporal deformation field, EndoNeRF can jointly encode geometry and temporal motion from a limited set of images, thereby enabling flexible 3D scene synthesis in dynamic surgical settings. EndoSurf [48] embeds signed-distance fields within a radiance-field backbone to impose explicit geometric constraints, yielding smoother and more precise surfaces that are crucial for surgical visualisation. However, NeRF-style approaches are inherently limited by their requirement to sample numerous points along each viewing ray and to perform an MLP evaluation at every sample. This high computational multiplicity leads to long training cycles, substantial memory footprints, and rendering latencies that are incompatible with the real-time requirement of intra-operative surgical guidance, thereby motivating the exploration of explicit, real-time representations such as Gaussian Splatting [12, 37].

Recent advancements introduced explicit representations via 3D Gaussian Splatting (3DGS), overcoming critical limitations of implicit models. 3DGS represents scenes explicitly with anisotropic Gaussian primitives optimized through differentiable rasterization, enabling rapid inference speeds and real-time rendering capabilities [12, 37]. Groundbreaking studies like 3DGS demonstrated substantial performance improvements, achieving real-time frame rates while maintaining visual fidelity competitive with state-of-the-art NeRF-based methods [12]. Extending 3DGS to dynamic scenes, 4D Gaussian Splatting (4DGS) integrates temporal deformation fields directly into Gaussian primitives, offering an efficient representation for dynamic scenes by using lightweight neural

deformation networks to model Gaussian trajectories over time [36]. This development significantly improves rendering speed and storage efficiency compared to previous methods, thus proving highly suitable for dynamic surgical scene reconstruction.

Motivated by these limitations, a growing body of work has adopted Gaussian-Splatting to accelerate the reconstruction of dynamic surgical scenes. EndoGS [52] boosts monocular performance by combining depth-guided spatio-temporal weighting with surface-aligned regularisation, thus alleviating severe occlusions. EndoGaussian [18, 19] introduces holistic Gaussian initialisation from depth estimation and incorporates a lightweight spatio-temporal tracker to cope with large deformations. Although GS methods can achieve real-time rendering, endoscopic surgery still presents unresolved challenges: large non-rigid tissue motion, instrument-induced occlusions, and an uneven or sparse Gaussian distribution often give rise to aliasing artefacts and inaccurate geometry. Existing GS frameworks primarily optimise for speed and do not explicitly address alias suppression or deformation robustness.

To overcome these shortcomings, we introduce **SAGS**, an attention-driven, alias-free Gaussian-splatting framework specifically designed for dynamic endoscopic reconstruction. SAGS suppresses high-frequency artefacts while employing a self-adaptive deformation decoder to capture complex tissue motion, thereby delivering high-fidelity 3D reconstructions suitable for the reconstruction of deformable endoscopic tissues.

The contributions of this paper include:

- (1) We propose a self-adaptive weighted deformation decoder, a multi-head attention-based mechanism capable of dynamically weighting Gaussian attributes, significantly enhancing the ability to model deformations in complex endoscopic scenes.
- (2) We employ 3D smoothing filters and 2D Mip filters to achieve alias-free processing, effectively reducing artifacts in the reconstruction of deformable tissues.
- (3) Experimental results on the public benchmarks EndoNeRF [33] and SCARED [1] show that our method performs better than state-of-the-art approaches in PSNR, SSIM, and LPIPS metrics while also delivering enhanced visual reconstruction quality.

2. Related Works

2.1. 3D Reconstruction

3D reconstruction has experienced rapid development [24], driven by various methods and applications ranging from traditional computer vision techniques to advanced neural rendering frameworks. Early classical methods such as Structure-from-Motion (SfM) [35], and traditional volumetric rendering have laid fundamental groundwork for subsequent innovations [4, 42]. Real-time non-rigid capture emerged with DynamicFusion (2015) [25], which fused depth maps while estimating dense deformation fields. Li et al. [14] proposed a weighted 3D volume reconstruction method from slice data based on a modified Allen-Cahn

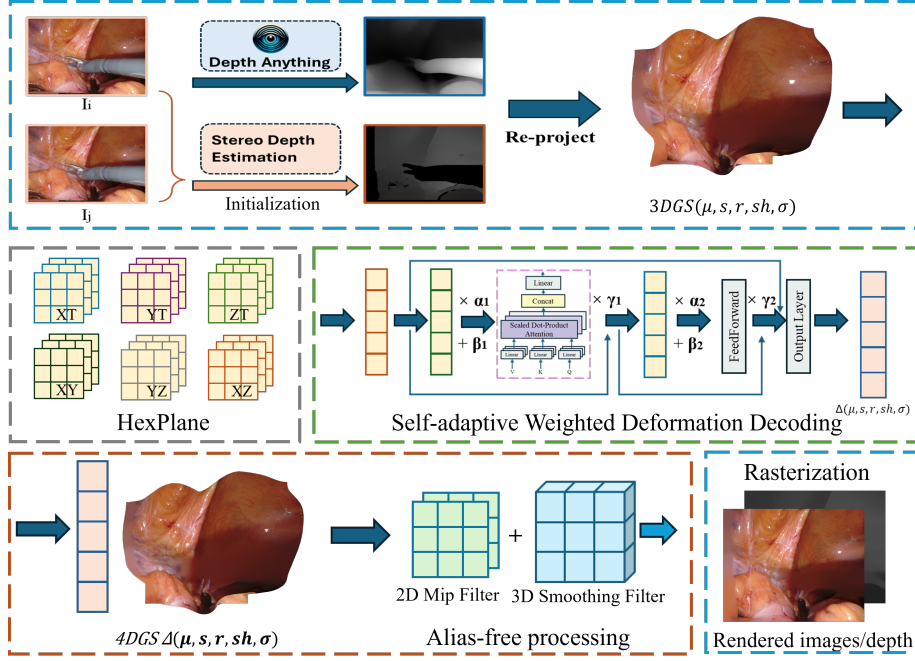


Figure 1: The overall pipeline of the proposed SAGS framework. Depth maps from monocular and stereo estimation are re-projected to initialize 3D Gaussians, which are then refined using HexPlane encoding and a self-adaptive weighted deformation decoder MLP for deformation modeling. Alias-free filtering with Mip and smoothing filters is subsequently applied, and finally, rasterization generates high-fidelity rendered images and depth.

equation. Shi et al. [30] introduced an edge-guided framework that improves 3D reconstruction from RGB images and sketches.

More recently, neural implicit representations, particularly NeRF [22], have significantly advanced the quality and realism of scene reconstructions by representing 3D scenes implicitly through continuous volumetric functions optimized via differentiable rendering techniques [28]. And more recent developments, such as D-NeRF [28] and NeRFies [26], integrated deformation fields to capture dynamic scenes effectively. To tackle the computational complexity and slow inference speed inherent to NeRF, several follow-up methods emerged. Instant neural graphics primitives employ a multi-resolution hash grid to drastically speed up both the optimization and rendering phases of Neural Radiance Field models [23, 9], while explicit representations, such as Plenoxels, provide real-time rendering capabilities without relying on neural networks [7]. H₂O-NeRF reconstructs radiance fields for two-hand-held objects by combining SDF-based semantic cues with view-dependent visibility masks, improving completeness and view consistency under severe hand occlusion [17]. MS-NeRF enhances NeRF performance in reflective and refractive scenes by introducing parallel sub-space radiance fields, achieving higher PSNR with minimal computational overhead [43]. STGC-NeRF enforces spatial-temporal geometric consistency for

dynamic LiDAR scenes, improving reconstruction accuracy from sparse, low-frequency input data [44]. However, most of the NeRF-based methods remain burdened by substantial computational demands and a substantial memory footprint.

Recently, 3DGS has emerged as a promising technique, explicitly representing scenes with optimized anisotropic Gaussian primitives. This method enabled rapid rendering and improved reconstruction quality by directly optimizing Gaussian attributes from scene images [12]. Subsequent enhancements to 3DGS further improved performance, including 3DGS for anti-aliased rendering via multi-scale design [38], and Mip-Splatting for alias-free representation [45]. SpotLessSplats improves the robustness of 3D Gaussian Splatting by suppressing transient distractors using pre-trained features and robust optimisation, enabling high-quality reconstruction in casual capture settings [29]. EDGS improves efficiency in dynamic Gaussian Splatting by modeling sparse, time-variant attributes, significantly reducing redundancy and accelerating rendering for monocular videos [13]. Extending these concepts into 4D scenarios, 4DGS was proposed to handle temporal variations explicitly, providing the capabilities that can render dynamic scenes in real-time [36]. Instruct-4DGS enables efficient dynamic scene editing by separating static and dynamic components within a 4D Gaussian framework, significantly reducing editing time while preserving visual fidelity [13]. Overall, these methods collectively pushed the boundaries of 3D and 4D reconstruction, significantly enhancing realism, computational efficiency, and adaptability to dynamic scenes. Jiang et al [11]. proposed a real-time point-splatting framework for dynamic hand reconstruction with photorealistic rendering.

2.2. 3D Reconstruction for Medical Applications

3D reconstruction techniques have been increasingly important in medical scenarios, particularly for dynamic surgical environments. Traditional reconstruction approaches such as SfM [35] and SLAM-based methods [3, 51] have facilitated initial explorations in reconstructing medical scenes. SfM reconstructs 3D models by estimating camera poses and sparse point clouds from image collections, but it typically struggles with dynamic and textureless scenes common in endoscopic procedures. In contrast, SLAM-based methods integrate camera localization and dense mapping simultaneously, offering more robust performance in minimally invasive surgery [3, 51]. Machucho-Cadena et al. [21] applied geometric algebra methods for ultrasound probe tracking, tumor segmentation, and 3D reconstruction in medical imaging. Zhang et al. [24] proposed a domain-adaptive method for 3D microvascular reconstruction from OCT angiography images.

Recent developments in neural rendering, especially NeRF [22], have sparked significant interest in reconstructing dynamic surgical scenes. Methods such as EndoNeRF [33] and D-NeRF [28] have effectively modeled dynamic tissues by training neural fields for deformation and canonical density. EndoNeRF specifically adapts NeRF to endoscopic contexts, integrating dynamic deformation fields with neural radiance for realistic scene reconstruction [5, 27]. To improve

rendering speed and training efficiency, LerPlane [39] utilizes 4D representations by extending 3D spaces with temporal dimensions, significantly accelerating the reconstruction process. However, NeRF-based pipelines are fundamentally constrained by lengthy training cycles, slow inference, and considerable memory footprints, which collectively hinder their practical deployment in the operating theatre.

As an explicit representation approach, 3DGS has recently emerged, demonstrating impressive real-time rendering capabilities and high-quality reconstructions through anisotropic Gaussians [12]. Adaptations of 3DGS for dynamic scenes, such as EndoGS [52] and EndoGaussian [18, 19], have achieved superior reconstruction quality by addressing deformation tracking and spatial-temporal coherence. EndoGS employs surface-aligned regularization to reduce artifacts and enhance surface consistency, proving robust against occlusions common in surgical procedures [52]. Similarly, EndoGaussian introduces holistic Gaussian initialization and spatio-temporal tracking for effective real-time performance [18, 19]. HFGS specifically targets high-frequency reconstruction issues, enhancing both spatial and temporal fidelity in endoscopic videos [50]. Other notable methods such as Deform3DGS [41] integrate deformation fields and surface alignment into the 3D Gaussian framework to enhance reconstruction accuracy and surface details. Despite these advancements, existing methods still face challenges related to artifact reduction, spatial-temporal coherence, and computational efficiency, motivating continuous development in the field. Nevertheless, most existing GS-based pipelines are still optimised primarily for speed and therefore devote limited capacity to learning alias-suppression mechanisms and fine-grained deformation cues; consequently, specular ringing, texture drift, and subtle folding artefacts remain visible in challenging frames, revealing a persistent gap in alias-aware, deformation-adaptive modelling that motivates the proposed *SAGS* framework.

3. Methodology

Endoscopic surgical scenes are characterised by rapid, non-rigid tissue motion, severe occlusions from instruments, and highly specular, spatially varying illumination. These factors impose stringent requirements on any 3D representation used for intra-operative guidance: it must preserve fine geometric detail without introducing aliasing artefacts, and remain robust to large, topology-changing deformations. Traditional NeRF-based methods struggle to satisfy these constraints. And secondly, their implicit volumetric formulation incurs substantial computational latency, whereas classical SLAM pipelines fail to model the continuous tissue motion observed in laparoscopy. Consequently, there is a growing interest in explicit, point-based encodings that can be updated and rendered faster while still providing photorealistic quality.

3.1. Preliminary of 3D Gaussian Splatting

3DGS [12] provides fast rendering capabilities and superior 3D representation performance. It represents scenes explicitly through point clouds, which models

each point cloud as a 3DGS characterized by a center point μ (*a.k.a.*, the mean), as well as a covariance matrix Σ as:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \quad (1)$$

When projecting 3-D Gaussians onto the image plane, each Gaussian’s covariance is transformed into a 2D covariance matrix via

$$\Sigma' = J W \Sigma W^T J^T,$$

where Σ is the original 3D covariance, W represents the view-dependent rigid transformation, and J is the Jacobian matrix of the projection operation.

The covariance matrix Σ is expressed as: $\Sigma = R S S^T R^T$, where R defines the rotation, and S specifies the scale, to ensure positive semi-definiteness. Rendering pixel colors $C(p)$ is achieved through point-based volume rendering, which combines color contributions and opacities of Gaussians along the ray:

$$C(p) = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (2)$$

Here, the opacity α_i for each Gaussian is computed as:

$$\alpha_i = \sigma_i e^{-\frac{1}{2}(p-\mu_i)^T \Sigma'^{-1}(p-\mu_i)}. \quad (3)$$

In this framework, μ_i specifies the Gaussian’s position, c_i represents its color, and σ_i indicates its opacity. To account for view-dependent effects, spherical harmonics are employed for color modeling. Each explicit 3D Gaussian is modeled by a set of characteristics: its position $\mu \in \mathbb{R}^3$, scaling factor $s \in \mathbb{R}^3$, rotation factor $r \in \mathbb{R}^4$, spherical harmonic (SH) coefficients $sh \in \mathbb{R}^k$ (where k denotes the number of SH functions), and opacity $\sigma \in \mathbb{R}$. Collectively, these attributes define the Gaussian as (μ, s, r, sh, σ) .

3.2. Point Cloud Information Acquisition

Accurate depth cues are indispensable for dynamic endoscopic reconstruction, where narrow baselines, specular highlights, and rapid non-rigid tissue motion make reliable correspondence estimation particularly challenging. In such confined surgical scenes, point clouds must therefore be initialised from either monocular or binocular cues before they can be refined by our Gaussian-splatting pipeline.

Monocular Depth: Recent advances in large-scale depth pre-training have markedly improved single-frame depth estimation, even under the extreme lighting and texture conditions of minimally invasive surgery. Inspired by previous work [18, 19], we adopt *Depth Anything* [40], which is optimised on billions of images with synthetic and sparsely supervised depth, and has demonstrated strong zero-shot generalisation to laparoscopic footage. Given an endoscopic image I_i from time step T , the network predicts a dense depth map:

$$D_i = \text{DepthAnything}(I_i),$$

which we then back-project through the intrinsic calibration to generate an initial partial point cloud P_i following the reprojection scheme in EndoNeRF [33]. **Binocular Depth:** Building on the previous work [18, 19], binocular depth estimation is achieved by using adjacent stereo inputs I_i and I_j to compute the depth of the remaining frame D_i using stereo depth prediction methods [15]. The depth D_i is then processed through the same re-projection pipeline described in Monocular Depth to generate the corresponding point cloud P_i .

3.3. 4D Representation for Deformable Tissue Reconstruction

In dynamic endoscopic surgery, tissue motion is continuous and highly non-rigid, necessitating a 4D representation that can evolve over time. While standard 3DGS captures static geometry efficiently, it lacks the temporal expressiveness needed for live surgical scenes. 4DGS [36] tackles this gap by proposing a deformable time-series representation in which each Gaussian primitive can change its position, shape, and appearance across frames.

The 4D representation in the proposed SAGS models the Gaussian deformations of deformable tissues. This involves not only learning the Gaussian attributes—position $\mu \in \mathbb{R}^3$, scaling factor $s \in \mathbb{R}^3$, rotation factor $r \in \mathbb{R}^4$, spherical harmonic (SH) coefficients $sh \in \mathbb{R}^k$, and opacity $\sigma \in \mathbb{R}$ —but also tracking their deformations, which are defined as a set ΔGS . To drive the deformation fields, we employ the HexPlane [2] with a resolution of D_1 and D_2 . The HexPlane comprises six planes: $P_{XY}, P_{XZ}, P_{YZ}, P_{XT}, P_{YT}$, and P_{ZT} , where the first three are spatial planes and the latter three are spatiotemporal planes.

The HexPlane encodes Gaussian information I , where $I \in \mathbb{R}^{h \times D_1 \times D_2}$ and h represents the hidden space. The encoded voxel information I_{voxel} for point (μ, t) can be extracted as:

$$I_{\text{voxel}}(\mu, t) = \mathcal{F}(I_{XY}, x, y) \odot \mathcal{F}(I_{XZ}, x, z) \odot \mathcal{F}(I_{YZ}, y, z) \odot \mathcal{F}(I_{XT}, x, t) \odot \mathcal{B}(I_{YT}, y, t) \odot \mathcal{F}(I_{ZT}, z, t). \quad (4)$$

Here, $\mathcal{F}$ denotes the bilinear interpolation operation used to obtain the nearest voxel Gaussian information, and $\odot$ represents the element-wise multiplication. This voxel encoding mechanism ensures the integration of spatial and temporal features, which are critical for accurately reconstructing the deformation fields. However, the aforementioned 4DGS variant that relies solely on the HexPlane representation is unable to capture the complex, non-rigid deformations present in dynamic surgical scenes. To overcome this problem, we proposed a *Self-adaptive Weighted Deformation Decoder*.

3.4. Self-adaptive Weighted Deformation Decoding

Endoscopic scene reconstruction is uniquely challenging: tissues undergo large, non-rigid deformations, surgical tools create severe and view-dependent occlusions, lighting is highly specular and spatially varying, and camera motion

is restricted to narrow baselines. These factors hinder stable correspondence estimation and make it difficult for conventional neural encoders to predict temporally coherent geometry and appearance. To address these issues, and inspired by recent advances of MLPs [32, 16], we introduce a dynamically weighted multi-head self-attention module integrated with a deformation-aware decoder, named the self-adaptive deformation decoder, to model and decode the Gaussian attribute deformations. Unlike traditional methods that rely solely on fixed-weight MLPs, our approach introduces learnable dynamic weights to adaptively focus on spatial-temporal features, enabling more accurate and robust deformation predictions for each attribute.

The dynamic weight mechanism assigns adaptive importance to different features during the self-attention computation. A fixed combination of attention and MLP outputs would fail to account for the varying demands of endoscopic scenes, where large-scale motions demand global coherence while fine-scale tissue variations require local refinement. Specifically, for each Gaussian attribute deformation, the dynamic weights γ_1 and γ_2 adjust the contributions of the self-attention output and the MLP output, respectively. These weights are learnable and initialized with small values to allow gradual learning during training. The self-attention approach can be defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V. \quad (5)$$

Here, Q (Query), K (Key), and V (Value) are matrices obtained from the encoded voxel features $I_{\text{voxel}}(\mu, t)$, and each key vector has dimensionality d_k . The dimensionality of the key vectors is d_k . The attention is applied across multiple heads: $\text{MSA}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O$, where head_i denotes the output of the i -th attention head, and W^O is a learnable projection matrix.

As shown in Fig. 1, we employ a dynamically weighted mechanism to adaptively combine the outputs of the self-attention and MLP branches. In this design, the learnable deformable attributes, denoted as y , are obtained by aggregating the contributions from both pathways under dynamic weights (*i.e.*, α_1 , β_1 , α_2 , β_2 , γ_1 , and γ_2). This mechanism allows the model to flexibly balance global consistency captured by self-attention and local detail refinement provided by MLPs, ensuring that the deformation representation adapts effectively to varying tissue motions.

This self-adaptive decoding is formulated as:

$$y = \text{Affine}_{\text{post}}(y') + \gamma_2 \cdot \text{MLP}(\text{Affine}_{\text{post}}(y')) + x, \quad (6)$$

where

$$y' = \text{Affine}_{\text{pre}}(x) + \gamma_1 \cdot \text{MSA}(\text{Affine}_{\text{pre}}(x)). \quad (7)$$

Here, $\text{Affine}_{\text{pre}}(x)$ and $\text{Affine}_{\text{post}}(x)$ denote two learnable Affine transformations applied to the input features, formulated as $\text{Affine}(x) = \alpha \cdot x + \beta$, where α and β represent the scaling and shifting parameters, respectively. The

Affine_{pre} transformation serves to normalize and re-scale the features before they enter the attention block, while Affine_{post} adjusts the outputs after feature aggregation, ensuring stable training and effective feature fusion. The two-stage pre-post Affine transformation stabilizes feature scaling and shifting before and after each branch, which regularizes feature magnitudes and facilitates residual learning. The weighted residual formulation, governed by γ_1 and γ_2 , provides a controllable trade-off between global and local cues, rather than enforcing a rigid fusion. This operation serves as a lightweight linear adaptation layer, ensuring that the input features are properly normalized and rescaled before entering the self-attention and MLP branches. Learning the affine parameters α_1 , α_2 , β_1 , and β_2 further increases the flexibility of the decoder by allowing feature-level adjustments that facilitate stable training and improve the expressiveness of deformation modeling.

This design leverages the complementary strengths of the two modules: self-attention is particularly effective at capturing long-range dependencies and preserving global geometric consistency, whereas the MLP component is more adept at modeling local nonlinear variations and fine-grained tissue deformations. Following the previous work [18, 19], we use four small MLPs in the output layers. Instead of assigning fixed contributions, the dynamically learnable weights (*i.e.*, α_1 , α_2 , β_1 , β_2 , γ_1 , and γ_2) regulate the relative influence of these two information pathways, enabling the network to emphasize global coherence under substantial movements and viewpoint shifts, while simultaneously prioritizing localized corrections when detailed tissue structures undergo fine-scale motion.

Within the Self-adaptive Weighted Deformation Decoding module, the residual outputs from the attention and MLP branches are not only combined through dynamically learned weights but are also propagated to update the Gaussian parameters directly. In this way, the adaptive weighting directly governs how global coherence and local deformation cues translate into geometry refinement.

As the final step of the self-adaptive weighted deformation decoding, the residuals Δ derived from the attention-MLP aggregation are applied to the initial Gaussian parameters. In this way, the Gaussian attributes are iteratively refined and updated as:

$$\mu' = \mu + \Delta\mu, \quad s' = s + \Delta s, \quad r' = r + \Delta r, \quad (8)$$

$$sh' = sh + \Delta sh, \quad \sigma' = \sigma + \Delta\sigma. \quad (9)$$

Here, $\mu \in \mathbb{R}^3$ denotes the Gaussian mean, representing the 3D spatial position of the primitive. $s \in \mathbb{R}^3$ encodes the anisotropic scaling factors, which control the spatial extent of the Gaussian along the three principal axes. $r \in \mathbb{R}^4$ corresponds to the quaternion rotation, defining the orientation of the Gaussian ellipsoid in 3D space. $sh \in \mathbb{R}^k$ represents the spherical harmonic coefficients that model view-dependent color variations, enabling photorealistic appearance representation. Finally, $\sigma \in \mathbb{R}$ denotes the opacity term, governing the transparency and blending behavior of the Gaussian during rasterization.

This novel dynamic attention-based deformation decoder significantly enhances the representation and reconstruction quality for deformable tissues by

combining attention-based global feature aggregation and local feature refinement, allowing the model to effectively capture fine-grained details and large-scale deformations in a highly adaptive manner.

3.5. Alias-Free Processing

Surgical dynamic reconstruction is particularly prone to visual artefacts: specular highlights generated by the endoscope light, narrow stereo baselines, and rapid non-rigid tissue motion combine to produce strong aliasing, ringing, and frame-to-frame flicker. In our 4D reconstruction framework, these aliasing and high-frequency artefacts pose a major obstacle to both quantitative fidelity and clinical usability. To mitigate them, we integrate 3D smoothing filters together with 2D Mip filters, drawing on the anti-aliasing principles of Mip-Splatting [46, 47]. Acting in tandem, the volumetric (3D) and image-space (2D) filters attenuate high-frequency noise during Gaussian optimisation and during the final projection step, respectively, thereby yielding temporally consistent, artefact-free reconstructions.

The 3D smoothing filter [46] is applied to each Gaussian primitive to constrain its high-frequency components according to the Nyquist sampling theorem. Given the maximal sampling rate $\hat{\nu}_k$ of a Gaussian primitive k , we apply a filter for Gaussian low-pass G_{low} with variance Σ_{low} to regularize the Gaussian, defined as: $G_k(\mathbf{x})_{\text{reg}} = (G_k * G_{\text{low}})(\mathbf{x})$, where $*$ denotes the convolution operation. The convolution of two Gaussians results in another Gaussian, with the new covariance matrix given by $\Sigma_k + \Sigma_{\text{low}}$. The regularized Gaussian can be expressed as:

$$\mathcal{G}_k(\mathbf{x})_{\text{reg}} = \sqrt{\frac{|\Sigma_k|}{|\Sigma_k + \frac{s}{\hat{\nu}_k} \cdot \mathbf{I}|}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{p}_k)^T (\Sigma_k + \frac{s}{\hat{\nu}_k} \cdot \mathbf{I})^{-1} (\mathbf{x} - \mathbf{p}_k)}, \quad (10)$$

where s is a scalar hyperparameter controlling the filter size, $\hat{\nu}_k$ is the maximal sampling rate for primitive k , and $\mathbf{p}_k$ represents the center of the Gaussian. By applying this filter, high-frequency artifacts in the volumetric domain are effectively reduced.

While 3D smoothing filters suppress high-frequency artifacts in the volumetric representation, aliasing can still occur during these Gaussians onto the 2D image plane. To handle this, we use 2D Mip filters [46] to approximate a box filter for each pixel in screen space, defined as:

$$\mathcal{G}_k^{2D}(\mathbf{x})_{\text{mip}} = \sqrt{\frac{|\Sigma_k^{2D}|}{|\Sigma_k^{2D} + s\mathbf{I}|}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{p}_k)^T (\Sigma_k^{2D} + s\mathbf{I})^{-1} (\mathbf{x} - \mathbf{p}_k)}, \quad (11)$$

where Σ_k^{2D} represents the Gaussian’s covariance projected into 2D screen space, and s is chosen to cover a single pixel. This filter mimics the integration of photons over a pixel’s area, ensuring the alignment of the rendered Gaussians and the pixel resolution, significantly reducing aliasing during zoom-out views or varying camera distances.

By integrating 3D smoothing and 2D Mip filters, we ensure that the voxel features $I_{\text{voxel}}(\mu, t)$ in our 4D endoscopic reconstruction pipeline remain robust to both spatial and temporal aliasing. The smoothed and filtered voxel information can be computed as:

$$I'_{\text{voxel}}(\mu, t) = \mathcal{F}(G'_{XY}, x, y) \odot \mathcal{F}(G'_{XZ}, x, z) \odot \mathcal{F}(G'_{YZ}, y, z) \\ \odot \mathcal{F}(G'_{XT}, x, t) \odot \mathcal{F}(G'_{YT}, y, t) \odot \mathcal{F}(G'_{ZT}, z, t), \quad (12)$$

where $G'_{XY}, G'_{XZ}, \dots$ represent Gaussians after 3D smoothing and 2D Mip filtering. This unified approach effectively suppresses artifacts in the 4D deformation fields, ensuring visually coherent and high-fidelity endoscopic scene reconstruction.

Following the previous work [18, 19], we adopt similar loss functions to optimize our framework, combining rendering constraints and spatio-temporal smoothness constraints into a unified objective: $L = \lambda_1 L_{\text{color}} + \lambda_2 L_{\text{depth}} + \lambda_3 L_{\text{spatial}} + \lambda_4 L_{\text{temporal}}$, where $\lambda_{1,2,3,4}$ are balancing weights for color rendering, depth consistency, spatial regularization, and temporal smoothness, respectively.

4. Experiments

4.1. Datasets

Dynamic 3D reconstruction in invasive surgery must cope with centimetre-scale tissue deformations, strong specular reflections from the endoscope light, frequent occlusions by forceps or scissors, and a very limited camera baseline. To study these challenges systematically, we adopt two public benchmarks that have become standard in endoscopic reconstruction research. In line with previous works [33, 48, 18, 19, 52], we trained and evaluated our method using two public benchmark datasets: EndoNeRF [33] and SCARED [1].

EndoNeRF [33]: This dataset was captured during in-vivo prostatectomy surgeries using stereo cameras. It includes two cases showcasing two distinct scenarios: tissue pulling and tissue cutting. The dataset presents significant challenges due to the irregular tissue deformation and the movement caused by surgical tools.

SCARED [1]: The SCARED dataset was captured using a DaVinci endoscope and a projector, providing RGB-D data of porcine cadaver abdominal anatomies. The dataset provides seven training sequences and two hidden test sequences. We follow the previous work [18, 19, 48] to split the dataset for training and testing.

4.2. Evaluation Metrics

Precise visual feedback is vital in minimally invasive surgery, where even subtle geometric or textural errors can mislead a surgeon’s perception of tissue boundaries or tool–tissue interaction. To capture both pixel-level accuracy and perceptual plausibility in this high-risk setting, we employ three complementary

Table 1: Performance comparison on the EndoNeRF dataset with SOTAs.

Dataset	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
EndoNeRF [33]	EndoNeRF [33]	36.062	0.933	0.089
	EndoSurf [48]	36.529	0.954	0.074
	LerPlane-9k [39]	34.988	0.926	0.080
	EndoGS [52]	36.990	0.961	0.038
	LerPlane-32k [39]	37.384	0.950	0.047
	EndoGaussian (Monocular) [18]	37.464	0.960	0.052
	SAGS (Monocular)	37.711	0.962	0.043
	EndoGaussian (Binocular) [18]	38.088	0.962	0.048
	SAGS (Binocular)	39.164	0.970	0.025

metrics: Peak Signal-to-Noise Ratio (PSNR) [10], which quantifies reconstruction accuracy by measuring pixel-wise differences between the reconstructed and ground-truth images; Structural Similarity Index (SSIM) [34], which assesses perceived structural similarity by comparing luminance, contrast, and structural consistency; and Learned Perceptual Image Patch Similarity (LPIPS) [49], which evaluates perceptual similarity using deep feature embeddings.

Following the previous work [18, 19], we randomly sample 0.1% of the points during the initialization stage to reduce redundancy and improve computational efficiency. We use the Adam optimizer for training, and during training, we use an initial learning rate of 1.6×10^{-3} . A warm-up strategy is employed, where the Gaussians are optimized for 1,000 iterations, followed by the optimization of the entire framework for an additional 3,000 iterations. The frequency of pruning and densification in point clouds depends on depth types and varies across different tasks. All experiments were conducted using an NVIDIA RTX 4090 GPU.

4.3. Comparison with the State-of-the-Art Methods

We comprehensively compared our proposed SAGS framework with several SOTA methods across two benchmark datasets: EndoNeRF [33] and SCARED [1], under both binocular and monocular depth settings.

On the **EndoNeRF dataset**, SAGS achieves top performance in both binocular and monocular configurations. In the binocular setting, SAGS attains a PSNR of **39.164**, an SSIM of **0.970**, and an LPIPS of **0.025**, surpassing all previous approaches, including EndoGaussian (PSNR: 38.008, SSIM: 0.962, LPIPS: 0.048). In the monocular setup, SAGS also outperforms others with a PSNR of **37.711**, an SSIM of **0.962**, and an LPIPS of **0.043**, consistently achieving the best results across all metrics. These improvements reflect the model’s robustness in reconstructing fine-grained structural and appearance details even under sparse input constraints.

On the **SCARED dataset**, which includes challenging scenarios with diverse lighting conditions and significant deformation, SAGS maintains consistent superiority. As reported in Table 2, our method outperforms EndoNeRF,

Table 2: Performance comparison of SAGS with SOTA methods on EndoNeRF [33] and SCARED [1] datasets using binocular depths.

Dataset	Task/Scene	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
EndoNeRF [33]	Pulling	EndoNeRF [33]	34.21	0.938	0.161
		EndoSurf [48]	35.00	0.956	0.120
		EndoGaussian [18]	37.21	0.957	0.061
		SAGS (Ours)	38.30	0.964	0.033
	Cutting	EndoNeRF [33]	34.19	0.932	0.151
		EndoSurf [48]	34.98	0.953	0.106
		EndoGaussian [18]	38.44	0.968	0.043
		SAGS (Ours)	39.51	0.972	0.022
SCARED [1]	d1k1	EndoNeRF [33]	24.37	0.763	0.326
		EndoSurf [48]	24.40	0.769	0.319
		EndoGaussian [18]	29.75	0.864	0.143
		SAGS (Ours)	30.23	0.875	0.102
	d2k1	EndoNeRF [33]	25.73	0.828	0.240
		EndoSurf [48]	26.24	0.829	0.254
		EndoGaussian [18]	30.90	0.871	0.189
		SAGS (Ours)	33.53	0.915	0.070
	d3k1	EndoNeRF [33]	19.00	0.599	0.467
		EndoSurf [48]	20.04	0.649	0.441
		EndoGaussian [18]	18.82	0.609	0.493
		SAGS (Ours)	20.11	0.619	0.426
	d6k1	EndoNeRF [33]	24.04	0.833	0.464
		EndoSurf [48]	24.09	0.866	0.461
		EndoGaussian [18]	25.69	0.871	0.372
		SAGS (Ours)	25.73	0.856	0.304
	d7k1	EndoNeRF [33]	22.64	0.813	0.312
		EndoSurf [48]	23.42	0.861	0.282
		EndoGaussian [18]	24.97	0.855	0.239
		SAGS (Ours)	26.54	0.862	0.168
	Average	EndoNeRF [33]	26.31	0.815	0.303
		EndoSurf [48]	26.88	0.840	0.283
		EndoGaussian [18]	29.40	0.857	0.220
		SAGS (Ours)	30.56	0.866	0.161

EndoSurf, and EndoGaussian across five different SCARED sequences. SAGS achieves the highest PSNR in four out of five sub-datasets (e.g., 33.53 on d2k1, 30.23 on d1k1), and its average performance across all sequences reaches a PSNR of **30.56**, SSIM of **0.866**, and LPIPS of **0.161**. This highlights the framework’s effectiveness in generalizing to highly dynamic and diverse surgical scenes beyond a single dataset.

Across datasets, SAGS consistently perform better than previous methods in terms of both pixel-level accuracy (PSNR/SSIM) and perceptual quality (LPIPS). Notably, even compared with recent high-performing models like EndoGS and LerPlane-32k, SAGS achieves better structural similarity and lower

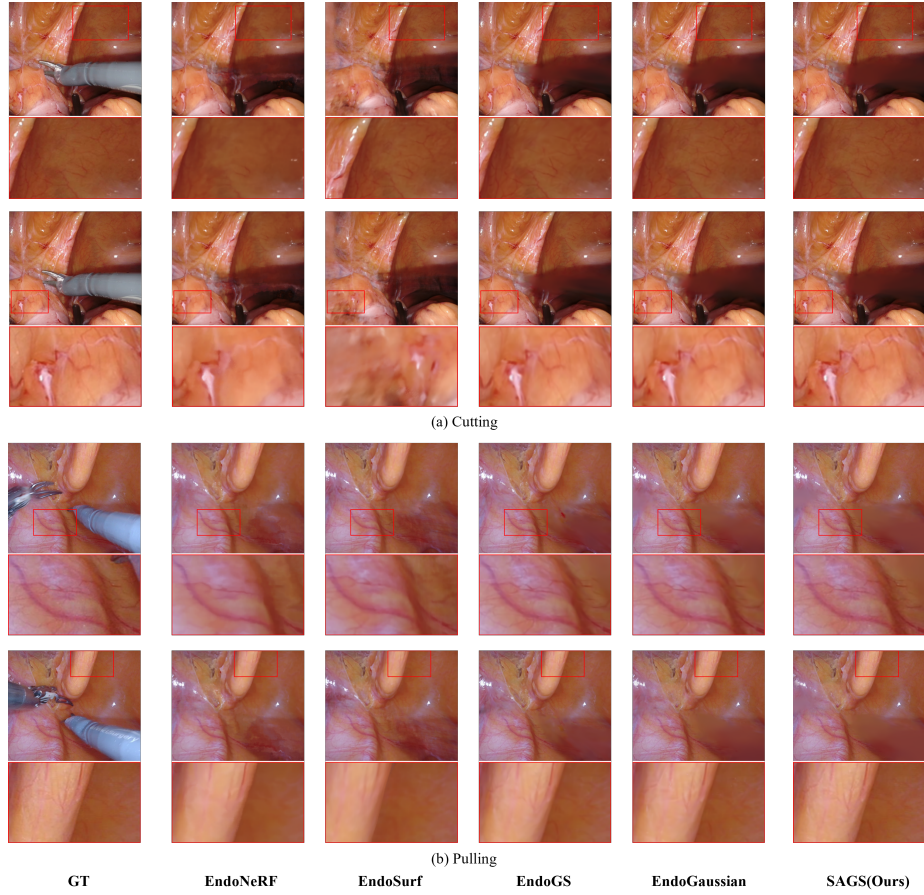


Figure 2: The qualitative result comparison between SOTAs and our proposed SAGS.

perceptual error, demonstrating its capacity to preserve geometric continuity and texture integrity under both sparse and dense depth supervision.

Qualitative comparisons further confirm the effectiveness of our framework, as shown in Figure 2. Figure 2 presents visual results for four challenging frames from the EndoNeRF dataset with binocular depth that involve strong specular highlights, rapid non-rigid tissue motion, and instrument-induced occlusions. In each case, the top row shows the full rendering, whereas the bottom row enlarges the red region of interest to reveal fine-grained differences. Methods adapted from NeRF (*EndoNeRF* and *EndoSurf*) suffer from noticeable blur and ringing around specular highlights, and the high-frequency vascular textures on the peritoneum become smeared once the camera viewpoint changes. The two existing Gaussian-splatting baselines (*EndoGaussian* and *EndoGS*) improve sharpness but still exhibit aliasing along instrument edges and faint contours on dynamically deforming tissue; in addition, colour consistency across adjacent

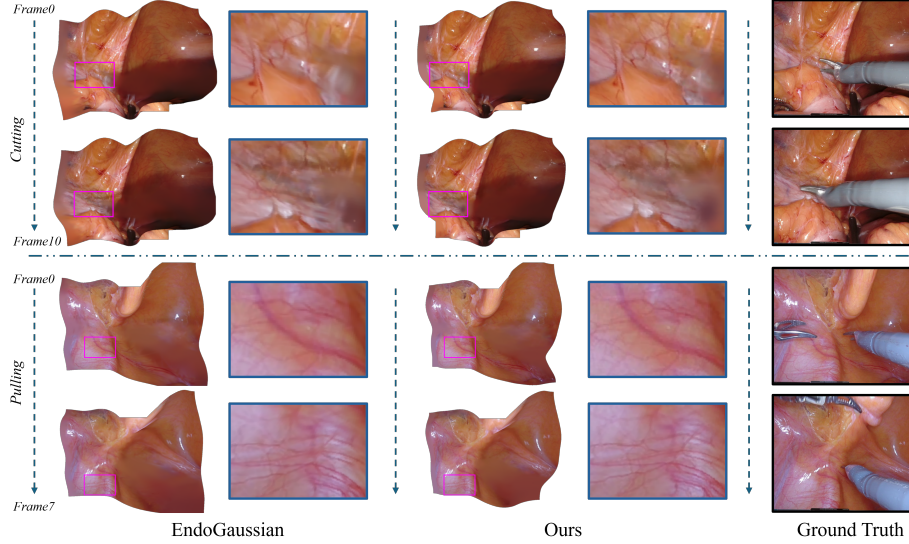


Figure 3: The qualitative result comparison between EndoGaussian and our proposed SAGS.

frames is occasionally lost, producing flicker artifacts.

By contrast, the proposed **SAGS** reconstruction is visually closest to the reference video. Vessel bifurcations and subtle surface folds remain crisp, specular reflections are neither over-sharpened nor haloed, and instrument boundaries appear well-defined without stair-step artefacts. The alias-free rasteriser suppresses moire patterns that are visible in the EndoGaussian results (see third frame), while the self-adaptive deformation decoder prevents the texture tearing seen in EndoGS when the grasper lifts tissue (row 3). Across all test frames, SAGS delivers more coherent shading, fewer high-frequency artefacts, and superior geometric integrity, qualitatively corroborating the quantitative gains reported in Tables 2 and 1.

Figure 3 compares the reconstructed 3D meshes of our SAGS pipeline with the most closely related baseline, EndoGaussian. Across both sequences, the EndoGaussian reconstructions exhibit noticeable texture drift: vascular patterns become blurred, and high-frequency highlights bleed across neighbouring. These artefacts are particularly evident in the “Cutting” sequence at Frames 0 and 10, where the specular streak along the liver surface spreads beyond its anatomical boundary. By contrast, SAGS preserves crisp vessel bifurcations and maintains a stable highlight footprint, indicating that the alias-free rasteriser successfully suppresses high-frequency noise. Geometric fidelity also improves: in the “Pulling” sequence, the surface around the grasper tip flattens in the EndoGaussian model, whereas our deformation-adaptive decoder reconstructs the local indentation, matching the ground-truth depth cue.

Table 3: Ablation study evaluating the efficacy of each proposed module on the EndoNeRF-Pulling [33] dataset.

Ablation Study	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Baseline	37.18	0.9577	0.0632
w/o Alias-Free	37.33	0.9578	0.0627
w/o SAD	38.08	0.9629	0.0434
SAGS (Full)	38.34	0.9642	0.0326

4.4. Ablation Study

To measure the contributions of each proposed component in the SAGS framework, *i.e.*, the 3D smoothing filters and the self-adaptive deformation decoder (dubbed as “SAD”) module, we designed an ablation study on the EndoNeRF-Pulling dataset. Table 3 presents the ablation study results on the EndoNeRF-Pulling dataset, evaluated using PSNR, SSIM, and LPIPS metrics.

4.4.1. Effectiveness of Alias-Free Processing

Endoscopic videos contain strong specular highlights, sharp tissue boundaries, and rapid motion, all of which amplify high-frequency content and make dynamic reconstructions especially susceptible to ringing, Moire patterns, and shimmer across frames. Hence, suppressing aliasing is critical for delivering clinically reliable 3D visualisation. We first evaluate the effectiveness of the Alias-Free processing module. As illustrated in Table 3, adding Alias-Free processing to the baseline model results in improvements in both PSNR (from 37.18 to 37.33) and LPIPS (from 0.0632 to 0.0627). These improvements, although subtle, demonstrate that the Alias-Free module effectively mitigates high-frequency noise and aliasing artifacts, contributing to clearer and smoother visual reconstructions. This validates the importance of the Alias-Free processing in enhancing visual quality and reducing perceptual artifacts in dynamic surgical scene reconstruction.

4.4.2. Effectiveness of SAD Module

Modelling surgical tissue motion is particularly challenging: organs undergo centimetre-scale, non-linear deformations, and the interaction with graspers or scissors can introduce abrupt, topology-changing displacements. Static or fixed-weight networks often fail to track these rapid, heterogeneous motions, leading to texture drift and geometric distortion over time. To quantify the benefit of our *Self-Adaptive Weighted Deformation (SAD)* module—equipped with dynamic multi-head attention and learnable per-head weights—we replace it with a single-layer MLP of comparable parameter count. Removing SAD causes a consistent drop across all metrics: PSNR falls from 38.34 to 38.08, SSIM from 0.9642 to 0.9629, while LPIPS rises from 0.0326 to 0.0434. These degradations confirm that the SAD module is crucial for capturing the intricate spatial-temporal variations of soft tissue, enabling perceptually faithful and geometrically accurate reconstructions in dynamic endoscopic scenes.

4.4.3. Effectiveness of Combined Modules

Lastly, we examine the combined impact of integrating both the Alias-Free processing and the SAD module within the full SAGS framework. The complete model achieves the best results, achieving a PSNR of 38.34, SSIM of 0.9642, and LPIPS of 0.0326. This substantial performance gain illustrates a clear synergistic effect, where the complementary functions of artifact suppression by the Alias-Free module and dynamic modeling capabilities by the SAD module effectively enhance the overall reconstruction quality. Thus, this final evaluation confirms the integral roles and synergistic relationship of these two key components, jointly addressing challenges posed by dynamic endoscopic scenes.

5. Conclusion

In this work, we introduced SAGS, a novel self-adaptive alias-free Gaussian splatting framework for dynamic endoscopic scene reconstruction. Leveraging a dynamically weighted deformation decoder with multi-head attention and advanced alias-free processing through 3D smoothing and 2D Mip filters, SAGS effectively reduces artifacts and captures fine-grained tissue details under complex deformations. Comprehensive evaluations on EndoNeRF [33] and SCARED [1] datasets demonstrated that our SAGS outperforms state-of-the-art methods in terms of PSNR, SSIM, and LPIPS, highlighting its ability to preserve geometric and texture fidelity. Beyond quantitative improvements, qualitative results further validated the superiority of SAGS in reconstructing sharp details, mitigating aliasing, and maintaining temporal consistency in challenging scenarios involving dynamic tissue deformations and surgical tool interactions. These results emphasize the potential of our method for robotic-assisted surgery, where precise 3D modeling is essential for navigation and intervention planning.

References

- [1] Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al., 2021. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133.
- [2] Cao, A., Johnson, J., 2023. Hexplane: A fast representation for dynamic scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 130–141.
- [3] Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J., 2018. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer methods and programs in biomedicine* 158, 135–146.
- [4] Chen, R., Han, S., Xu, J., Su, H., 2019. Point-based multi-view stereo network, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1538–1547.

- [5] Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P., Willcocks, C.G., 2022. MedNeRF: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray, in: 2022 44th annual international conference of the IEEE engineering in medicine & Biology society (EMBC), IEEE. pp. 3843–3848.
- [6] Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q., 2022. Fast dynamic radiance fields with time-aware neural voxels, in: SIGGRAPH Asia 2022 Conference Papers.
- [7] Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A., 2022. Plenoxels: Radiance fields without neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5501–5510.
- [8] Gao, W., Tedrake, R., 2019. SurfelWarp: Efficient non-volumetric single view dynamic reconstruction. arXiv preprint arXiv:1904.13073 .
- [9] Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J., 2022. HeadNeRF: A real-time nerf-based parametric head model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20374–20384.
- [10] Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. ssim, in: 2010 20th international conference on pattern recognition, IEEE. pp. 2366–2369.
- [11] Jiang, Z., Rahmani, H., Black, S., Williams, B., 2025. 3d points splatting for real-time dynamic hand reconstruction. Pattern Recognition 162, 111426.
- [12] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. 42, 139–1.
- [13] Kong, H., Yang, X., Wang, X., 2025. Efficient gaussian splatting for monocular dynamic scene rendering via sparse time-variant attribute modeling, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4374–4382.
- [14] Li, Y., Song, X., Kwak, S., Kim, J., 2022. Weighted 3d volume reconstruction from series of slice data using a modified allen–cahn equation. Pattern Recognition 132, 108914.
- [15] Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M., 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6197–6206.
- [16] Liu, H., Dai, Z., So, D., Le, Q.V., 2021. Pay attention to MLPs. Advances in neural information processing systems 34, 9204–9215.

- [17] Liu, X., Zhang, Q., Huang, X., Feng, Y., Zhou, G., Wang, Q., 2025a. H₂-o-nerf: Radiance fields reconstruction for two-hand-held objects. *IEEE Transactions on Visualization and Computer Graphics* .
- [18] Liu, Y., Li, C., Liu, H., Yang, C., Yuan, Y., 2025b. Foundation model-guided gaussian splatting for 4D reconstruction of deformable tissues. *IEEE Transactions on Medical Imaging* .
- [19] Liu, Y., Li, C., Yang, C., Yuan, Y., 2024. Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. *arXiv:2401.12561*.
- [20] Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q., 2021. E-DSSR: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception, in: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV* 24, Springer. pp. 415–425.
- [21] Machucho-Cadena, R., Rivera-Rovelo, J., Bayro-Corrochano, E., 2014. Geometric techniques for 3d tracking of ultrasound sensor, tumor segmentation in ultrasound images, and 3d reconstruction. *Pattern recognition* 47, 1968–1987.
- [22] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65, 99–106.
- [23] Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* 41, 1–15.
- [24] Neupane, R.B., Li, K., Mao, Z., 2025. High-fidelity 3d reconstruction via unified nerf-mesh optimization with geometric and color consistency. *Pattern Recognition* , 112071.
- [25] Newcombe, R.A., Fox, D., Seitz, S.M., 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 343–352.
- [26] Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R., 2021. Nerfies: Deformable neural radiance fields, in: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5865–5874.
- [27] Psychogyios, D., Vasconcelos, F., Stoyanov, D., 2023. Realistic endoscopic illumination modeling for nerf-based data generation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 535–544.

- [28] Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F., 2021. D-nerf: Neural radiance fields for dynamic scenes, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10318–10327.
- [29] Sabour, S., Goli, L., Kopanas, G., Matthews, M., Lagun, D., Guibas, L., Jacobson, A., Fleet, D., Tagliasacchi, A., 2025. Spotlessplats: Ignoring distractors in 3D gaussian splatting. *ACM Transactions on Graphics* 44, 1–11.
- [30] Shi, W., Yin, A., Li, Y., Wen, Y., 2025. Edge-guided 3d reconstruction from multi-view sketches and rgb images. *Pattern Recognition* 163, 111462.
- [31] Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G., 2017. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters* 3, 155–162.
- [32] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al., 2022. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence* 45, 5314–5321.
- [33] Wang, Y., Long, Y., Fan, S.H., Dou, Q., 2022. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 431–441.
- [34] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity 13, 600–612.
- [35] Westoby, M.J., Brasington, J., Glasser, N.F., Hambrey, M.J., Reynolds, J.M., 2012. ‘structure-from-motion’ photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* 179, 300–314.
- [36] Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X., 2024. 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.
- [37] Xia, W., Xue, J.H., 2023. A survey on deep generative 3d-aware image synthesis. *ACM Computing Surveys* 56, 1–34.
- [38] Yan, Z., Low, W.F., Chen, Y., Lee, G.H., 2024. Multi-scale 3D gaussian splatting for anti-aliased rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20923–20931.
- [39] Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W., 2023. Neural lerplane representations for fast 4D reconstruction of deformable tissues, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 46–56.

- [40] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data, in: CVPR.
- [41] Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., Jin, Y., 2024b. Deform3DGS: Flexible deformation for fast surgical scene reconstruction with gaussian splatting, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 132–142.
- [42] Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1790–1799.
- [43] Yin, Z.X., Jiao, P.Y., Qiu, J., Cheng, M.M., Ren, B., 2025. MS-NeRF: Multi-space neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [44] Yu, S., Sun, X., Li, W., Xu, Q., Yuan, Z., Wang, S., She, R., Wang, C., 2025. STGC-NeRF: Spatial-temporal geometric consistency for lidar neural radiance fields in dynamic scenes, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9644–9652.
- [45] Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A., 2024a. Mip-splatting: Alias-free 3D gaussian splatting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 19447–19456.
- [46] Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A., 2024b. Mip-splatting: Alias-free 3D gaussian splatting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19447–19456.
- [47] Yu, Z., Sattler, T., Geiger, A., 2024c. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (TOG)* 43, 1–13.
- [48] Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z., 2023. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 13–23.
- [49] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in: Proceedings of the IEEE Computer Society CVPR. doi:10.1109/CVPR.2018.00068.
- [50] Zhao, H., Zhao, X., Zhu, L., Zheng, W., Xu, Y., 2024. HFGS: 4D Gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction, in: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25–28, 2024, BMVA. URL: <https://papers.bmvc2024.org/0039.pdf>.

- [51] Zhou, H., Jagadeesan, J., 2019. Real-time dense reconstruction of tissue surface from stereo optical video. *IEEE transactions on medical imaging* 39, 400–412.
- [52] Zhu, L., Wang, Z., Cui, J., Jin, Z., Lin, G., Yu, L., 2024. EndoGS: deformable endoscopic tissues reconstruction with gaussian splatting, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 135–145.