FeatureEndo-4DGS: Real-Time Deformable Surgical Scene Reconstruction and Segmentation with 4D Gaussian Splatting

Kai Li* KAII.LI@MAIL.UTORONTO.CA
University of Toronto, Canada

Junhao Wang* HRY.WANG@MAIL.UTORONTO.CA
University of Toronto, Canada

William Han WJHAN@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Ding Zhao DINGZHAO@CMU.EDU
Carnegie Mellon University, USA

Abstract

Minimally invasive surgery (MIS) requires high-fidelity, real-time visual feedback of dynamic and low-texture surgical scenes. To address these requirements, we introduce FeatureEndo-4DGS (FE-4DGS), the first real-time pipeline leveraging feature-distilled 4D Gaussian Splatting for simultaneous reconstruction and semantic segmentation of deformable surgical environments. Unlike prior feature-distilled methods restricted to static scenes, and existing 4D approaches that lack semantic integration, FE-4DGS seamlessly leverages pre-trained 2D semantic embeddings to produce a unified 4D representation—where semantics also deform with tissue motion. This unified approach enables the generation of real-time RGB and semantic outputs through a single, parallelized rasterization process. Despite the additional complexity from feature distillation, FE-4DGS sustains real-time rendering **(287.95 FPS)** with a compact footprint, achieves state-of-the-art rendering fidelity on EndoNeRF **(39.1 PSNR)** and SCARED **(27.3 PSNR)**, and delivers competitive EndoVis18 segmentation, matching or exceeding strong 2D baselines for binary segmentation tasks **(0.93 DSC)** and remaining competitive for multi-label segmentation **(0.77 DSC)**.

Keywords: Surgical Scene Reconstruction, Feature Distillation, Gaussian Splatting

Data and Code Availability This study did not involve the collection of new data from human subjects; all data is obtained from publicly available datasets. Specifically, we used the EndoNeRF dataset for endoscopic 4D reconstruction (Wang et al., 2022), the SCARED dataset for robotic surgery depth estimation (Allan et al., 2020), and the EndoVis18 dataset for surgical image segmentation (Allan et al., 2020). Our code for reproducing all experiments, including pre-processing and model training will be available at the following link: https://github.com/kailathan/FE-4DGS.

Institutional Review Board (IRB) This research did not involve the collection of new data from human subjects. All data used in this study were obtained from publicly available, de-identified datasets, and therefore IRB approval was not required.

1. Introduction

Recent advancements in artificial intelligence (AI) and the decreasing cost of computational resources have enabled the development of automated methods to assist minimally invasive surgery (MIS) in real-time (Ali et al., 2023). Tasks such as image classification (Subedi et al., 2024), object detection (Yu et al., 2022), semantic segmentation (Zhu et al., 2024), tissue tracking (Wang et al., 2024), and surgical scene reconstruction (Zha et al., 2023) have become increasingly sophisticated, significantly benefiting robotic-assisted MIS.

Surgical scene reconstruction has recently attracted considerable attention (Zha et al., 2023; Long et al., 2021; Liu et al., 2020, 2024b). This technique provides surgeons with a comprehensive real-time view that enhances navigation and instrument

* These authors contributed equally

control, and has potential to enable robotic surgery automation. Early methods were dominated by traditional techniques such as simultaneous localization and mapping (SLAM) (Chen et al., 2018; Song et al., 2018; Zhou and Jayender, 2021a), but the field has since transitioned to neural network–based approaches (Long et al., 2021; Li et al., 2020) and more recently, to neural radiance fields (NeRFs) (Mildenhall et al., 2020; Zha et al., 2023). However, NeRF-based methods require large volumes of data and suffer from slow rendering speeds, motivating the development of 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) as a more efficient alternative. Landmark studies such as EndoGaussian (Liu et al., 2024b) and LGS (Liu et al., 2024a) have successfully applied 3DGS to surgical scene reconstruction.

Feature field distillation techniques, which were initially developed for the NeRF domain (Zhi et al., 2021; Siddiqui et al., 2022; Kobayashi et al., 2022), have been extended to 3DGS (Zhou et al., 2025, 2024) beyond their application in surgical scene reconstruction. Inspired by these advances, we propose FeatureEndo-4DGS (FE-4DGS). FE-4DGS integrates semantic features of 2D foundation models (Ravi et al., 2024; Zhu et al., 2024) into 4DGS to support real-time segmentation and rendering of deformable surgical scenes. Crucially, online segmentation enables safe surgical augmented reality (AR) renderings by highlighting critical anatomy and instruments to guide surgical actions (Doornbos et al., 2024). We introduce a unified deformation module to simultaneously update both per-Gaussian geometric properties including position, scale, rotation, and opacity, in addition to per-Gaussian semantic features. The module extracts deformation features, which are learned deformation cues that enable semantic features and geometry to be updated consistently and efficiently. During optimization, we further employ a CNN-based semantic decoder to align the rendered semantic feature map with features extracted by 2D segmentation models.

In summary, our contributions are as follows:
1. We develop FE-4DGS, a real-time pipeline that jointly reconstructs deformable scenes and semantic feature fields in MIS in a single pass.
2. We achieve state-of-the-art performances on rendering fidelity, while maintaining real-time rendering speeds.
3. We compare FE-4DGS and 2D segmentation foundation models on binary and multi-label segmentation tasks and achieve state-of-the-art performances on binary segmentation, while being competitive in multi-label segmentation.

2. Related Works

2.1. Surgical Scene Reconstruction

Reconstructing soft tissues from endoscopic stereo videos is critical for intraoperative navigation and robotic automation (Lu et al., 2021; Penza et al., 2017; Liu et al., 2020). Traditional approaches based on SLAM (Chen et al., 2018; Song et al., 2018; Zhou and Jayender, 2021a; Gao and Tedrake, 2019) degrade under non-rigid motion, specular reflections, and occlusions. Learning-based methods with CNNs and vision transformers improve stereo depth estimation and non-rigid modeling (Li et al., 2020), yet remain limited in dynamic, unstructured scenes with weak depth cues.

NeRFs (Mildenhall et al., 2020) offer powerful 3D representations of complex deformations and have been extended to dynamic objects in nonsurgical contexts (Niemeyer and Geiger, 2021; Martin-Brualla et al., 2021). EndoNeRF (Wang et al., 2022) adapts dynamic NeRFs for robotic surgery, achieving state-of-the-art 3D reconstruction and deformation tracking. However, NeRFs demand extensive point and ray sampling, leading to heavy computational costs that hinder real-time applications (Chen and Wang, 2024; Rabby and Zhang, 2024).

To address efficiency constraints, 3DGS (Kerbl et al., 2023) models scenes with anisotropic Gaussians and tile-based rasterization, enabling real-time rendering while preserving quality. Under deformable conditions, recent works have leveraged 4DGS for surgical reconstruction, demonstrating strong performance in real-time 3D video scene modeling (Xie et al., 2024; Chen et al., 2024b).

2.2. Surgical Scene Segmentation

In addition to real-time scene reconstruction, simultaneous segmentation of key anatomical structures and surgical instruments during intraoperative procedures is highly advantageous. Previous works have explored segmentation in endoscopic surgeries (Wang et al., 2023; Li et al., 2024; Chen et al., 2024a). With recent advances in segmentation foundation models, such as SAM (Kirillov et al., 2023) and SAM 2 (Ravi et al., 2024), and their adaptations to the medical domain (e.g., MedSAM (Ma et al., 2024), MedSAM

2 (Zhu et al., 2024)), many approaches have fine-tuned these models for surgical applications. A recent study, Feature 3DGS (Zhou et al., 2024), extends 3DGS to incorporate feature field distillation, enabling real-time segmentation, language-guided editing, and other interactive operations. Building on these developments, our work integrates 4DGS (i.e., deformable 3DGS) with feature field distillation to facilitate real-time rendering and segmentation of surgical scenes, thereby advancing intraoperative visualization.

3. Our Method

We propose FE-4DGS, a 4D Gaussian splatting framework that augments surgical scene reconstruction with dense, real-time semantic features distilled from 2D segmentation foundation models (denoted as SAM for simplicity). Our pipeline integrates semantics through two components: (1) a Gaussian deformation module with a motion-aware decoder F_{feat} that updates per-Gaussian features across frames, and (2) a CNN-based decoder that upsamples rendered semantic maps and aligns them with SAM outputs under a per-pixel L_1 loss. A parallelized rasterizer jointly renders color, depth, and semantic features, enabling comprehensive scene representation. An overview of the pipeline is shown in Figure 1.

3.1. 3D Gaussian Scene Representation and Initialization

Our method builds upon the 3D Gaussian Splatting (3DGS) framework (Kerbl et al., 2023), which represents a scene using a dense collection of 3D Gaussians. Each Gaussian is centered at a mean μ and characterized by a covariance matrix Σ that defines its spatial spread. Specifically, the contribution of a Gaussian at a 3D point x is given by:

$$G(\mathbf{x}) = \exp\Big(-\frac{1}{2}(x\ -\mu)^\top \Sigma^{-1}(x\ -\mu)\Big). \quad (1)$$

For both efficiency and interpretability, the covariance matrix is factorized as:

$$\Sigma = R\,S\,S^\top\,R^\top, \quad (2)$$

where R is a rotation matrix and S is a scaling matrix. In this representation, the scene is modeled by jointly optimizing the parameters of each Gaussian—its position μ, rotation R, scaling S, opacity o, and spherical harmonic (SH) coefficients.

Following Liu et al. (2024b), a holistic initialization is performed by re-projecting pixels from the input image sequence using estimated depth maps. For an image I_i with depth D_i and binary mask M_i, the corresponding 3D points are computed as:

$$P_i = K^{-1}T_i\,D_i\,\big(I_i \odot M_i\big), \quad P = \bigcup_{i=1}^{T} P_i, \quad (3)$$

where K and T_i denote the intrinsic and extrinsic parameters of the camera, respectively.

3.2. Feature-Spatiotemporal (FST) Deformation Module: Semantic Feature Integrated Deformation Decoder

To capture spatiotemporal deformation, each Gaussian is first encoded by a 4D voxel module that maps its center μ and time t to a latent feature f. Inspired by Yang et al. (2023); Wu et al. (2024) to represent the 4D voxel module as a multi-resolution HexPlane introduced by Cao and Johnson (2023), we leverage a multi-resolution HexPlane to efficiently learn features for Gaussians in spacetime. The HexPlane is paired with a light MLP decoder to produce a hidden representation $h = F_{\text{out}}(f) \in \mathbb{R}^W$ that is fed into smaller deformation branches. Here, W denotes size of the hidden dimension. Refer to Appendix A for more details.

Four-Branch Geometric Deformation Decoder: We utilize four lightweight decoder branches to update the position μ, rotation R, scaling S, and opacity o. Let $h = F_{\text{out}}(f) \in \mathbb{R}^W$ denote the intermediate feature representation. For each geometric property $g \in \{\mu, R, S, o\}$, a branch-specific feature extractor F_g^{feat} and prediction head F_g^{head} produce the corresponding update Δg:

$$h_g = F_g^{\text{feat}}(h), \qquad \Delta g = F_g^{\text{head}}(h_g) \in \mathbb{R}^{d_g}, \quad (4)$$

where the output dimensions are $d_\mu = 3$, $d_R = 4$, $d_S = 3$, and $d_o = 1$. The features h_g are also used to generate semantic updates, as described in the next section.

Feature-Based Semantic Updater: Within a 4D dynamic surgical scene, semantic embeddings derived from SAM undergo systematic shifts under changes in viewpoint and 3D deformation throughout time. To address this, we introduce a semantic update network

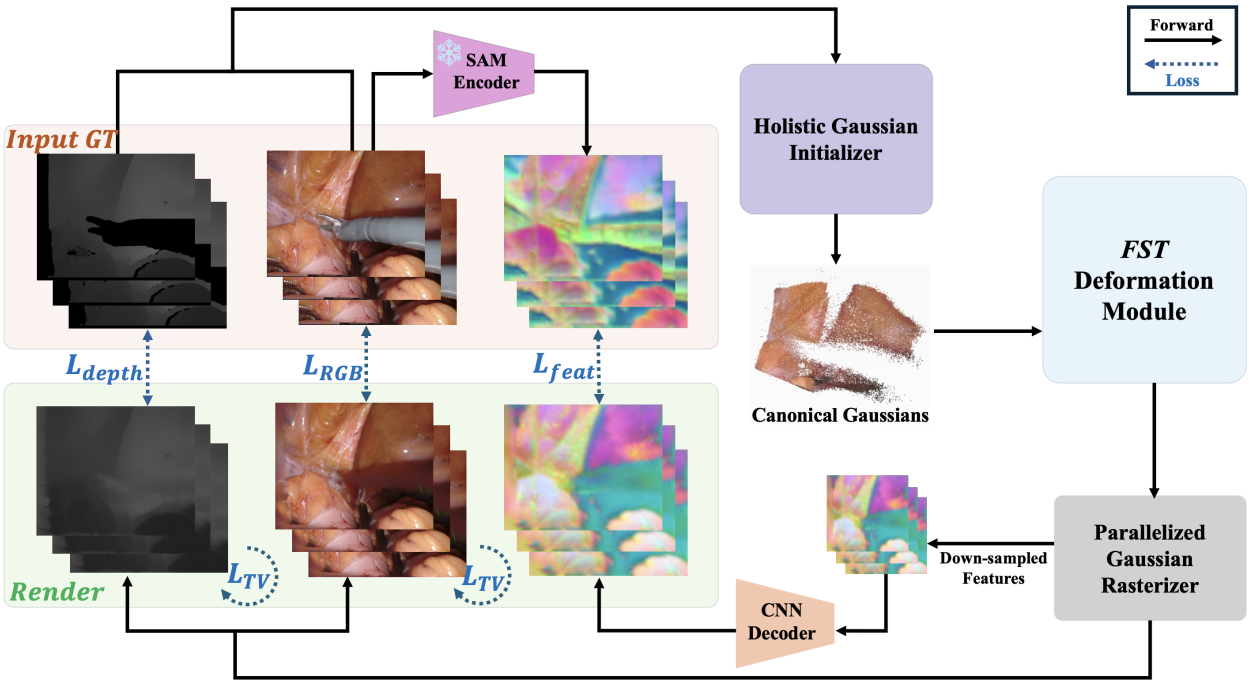

Figure 1: Overview of FE-4DGS. The pipeline begins with holistic Gaussian initialization, re-projecting image pixels into 3D Gaussians. In the FST deformation module, a 4D voxel encoder extracts latent features, which a deformation decoder refines by updating Gaussian parameters and semantics via a lightweight MLP (Section 3.2, Appendix A). A differentiable rasterizer renders the updated Gaussians into radiance and semantic maps, which a CNN decoder upsamples and aligns with features from a 2D segmentation model (SAM) to ensure semantic consistency.

F_{feat} which utilizes the aforementioned deformation features. We concatenate:

$$u_{\text{all}} = \left[h_\mu \,\|\, h_R \,\|\, h_S \,\|\, h_o\right] \in \mathbb{R}^{4W},$$

which is passed through our semantic-update network F_{feat} to obtain the per-Gaussian semantic feature update $\Delta z \in \mathbb{R}^N$ where N is the number of feature channels:

$$\Delta z = F_{\text{feat}}(u_{\text{all}}) \in \mathbb{R}^N \quad z' = z + \Delta z. \tag{5}$$

Finally, the fully augmented Gaussians are updated as follows:

$$G_t = \left(\mu + \Delta\mu,\ R + \Delta R,\ S + \Delta S,\ o + \Delta o,\ z'\right). \tag{6}$$

Semantic Feature Extraction: For all experiments, we used a pre-trained SAM (Kirillov et al., 2023) with a ViT-H encoder to extract semantic feature maps. Given an input image, SAM produces a high-dimensional semantic feature map with a spatial resolution of 64×64 with 256 channels. To match the aspect ratio of the input images, we adopt the cropping procedure from Zhou et al. (2024): an input image of size $H \times W$ (with $W > H$) yields a cropped feature map of size $64 \times \frac{64W}{H}$, preserving semantic information without introducing padding artifacts. The resulting feature maps serve as supervision signals during training. To improve computational efficiency without sacrificing semantic expressivity, we compress the rendered feature maps to 128 dimensions, which are updated through the semantic MLP branch F_{feat} as defined in Equation 5.

3.3. Differentiable Rendering and Loss Functions

During rendering, both the radiance (color) and semantic feature fields are computed via front-to-back alpha blending. For a pixel x, the rendered color $\hat{C}(x)$

and semantic feature $\hat{z}(x)$ are given by:

$$\hat{C}(x) = \sum_{i=1}^{N} c_i \, \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{7}$$

$$\hat{z}(x) = \sum_{i=1}^{N} z'_i \, \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{8}$$

where c_i and z'_i denote the color and updated semantic feature contributions from the ith Gaussian, and α_i is its effective opacity. The effective opacity is computed by evaluating a corresponding 2D covariance matrix:

$$\Sigma'_i = J \, W \, \Sigma_i \, W^\top \, J^\top, \tag{9}$$

with J representing the Jacobian of the affine approximation of the projective transformation and W the view transformation matrix.

CNN-Based Semantic Decoder: After differentiable rasterization, a CNN-based decoder is employed to upsample the rendered semantic feature map, matching the channel dimension to that of SAM. This decoder performs a simple pointwise convolution aligning the feature dimensionality to match SAM's output. The resulting semantic feature map is then compared, using a per-pixel L_1 loss, to the high-level semantic features f_{SAM}:

$$\mathcal{L}_{\text{feat}} = \frac{1}{HW} \sum_{x \in \Omega} \left\| \hat{z}(x) - z_{\text{SAM}}(x) \right\|_1, \tag{10}$$

where Ω denotes the set of pixel coordinates in an image of resolution $H \times W$.

In addition to the photometric loss $\mathcal{L}_{\text{rgb}}$ and depth loss $\mathcal{L}_{\text{depth}}$, as in Liu et al. (2024b), the overall loss is defined as:

$$\begin{aligned} \mathcal{L} = \lambda_{\text{rgb}} \, \mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}} \, \mathcal{L}_{\text{depth}} \\ + \lambda_{\text{feat}} \, \mathcal{L}_{\text{feat}} + \lambda_{\text{TV}} \, \mathcal{L}_{\text{TV}}. \end{aligned} \tag{11}$$

with $\mathcal{L}_{\text{TV}}$ enforcing spatiotemporal smoothness. The hyperparameters are fixed at $\lambda_{\text{rgb}} = 1$, $\lambda_{\text{depth}} = 0.01$, $\lambda_{\text{feat}} = 1$, and $\lambda_{\text{TV}} = 0.03$.

3.4. Training and Inference

The complete optimization is performed in a coarse-to-fine manner. In the coarse stage, only the basic Gaussian parameters are updated. In the fine stage, the semantic branch—comprising both the MLP F_{feat} within the deformation decoder and the CNN-based semantic decoder—is activated. This enables the network to jointly optimize the 3D Gaussian parameters and all MLP weights (including those of F_{feat}) using the Adam optimizer, while the CNN-based decoder ensures that the rendered semantic feature map is accurately aligned with the SAM features.

During inference, segmentation masks are generated by decoding the rendered and upsampled semantic features using SAM's pretrained decoder. To facilitate accurate multi-instance segmentation, class-specific bounding box prompts derived from ground truth annotations are provided as input. The inference process can be summarized as:

$$\hat{z}(x) = \text{FE-4DGS}_{\text{render}}(G_t, x), \tag{12}$$

$$\hat{z}'(x) = \text{FE-4DGS}_{\text{decoder}}(\hat{z}(x)), \tag{13}$$

$$\hat{M}(x) = \text{SAM}_{\text{decoder}}(\hat{z}'(x), B), \tag{14}$$

where $\hat{z}(x)$ represents the rendered semantic feature at pixel x, $\hat{z}'(x)$ is the upsampled feature map, B denotes bounding box prompts, and $\hat{M}(x)$ is the resulting segmentation mask.

We follow the experimental setup and configurations from EndoGaussian (Liu et al., 2024b) for evaluation consistency, noting that further hyperparameter tuning on specific datasets, such as the EndoVis18 dataset (Allan et al., 2020), may yield additional performance improvements.

4. Experiments

4.1. Experimental Setup

We optimize both the original Gaussian features and the deformation module—including parameters for HexPlane (Cao and Johnson, 2023), feature extractors, and decoders—using the Adam optimizer (Kingma and Ba, 2017). Following Liu et al. (2024b), we first optimize the canonical (coarse) Gaussians for 1000 iterations and then train the full FE-4DGS model for an additional 6000 iterations. In each iteration, we render one training camera view, compute losses, and apply an optimizer step. To ensure stable convergence, we apply an exponential learning rate decay schedule. All experiments are conducted on NVIDIA RTX A6000 48GB and NVIDIA RTX 5090 32GB GPUs unless specified otherwise. We provide a detailed breakdown of all hyperparameters used, including learning rates for different parameter groups, decay schedules, and other tuning adjustments in Appendix G.

Table 1: Baselines for surgical scene reconstruction on ENDONERF (Wang et al., 2022) and SCARED (Allan et al., 2021).

Dataset	Method	LPIPS (↓)	SSIM (↑)	PSNR (↑)
ENDONERF Wang et al. (2022)	EndoNeRF Wang et al. (2022)	0.09	0.93	36.06
	EndoSurf Zha et al. (2023)	0.07	0.95	36.53
	LerPlane-9k Yang et al. (2023)	0.08	0.93	34.99
	LerPlane-32k Yang et al. (2023)	0.05	0.95	37.38
	Endo-4DGS Huang et al. (2024)	0.04	0.96	37.21
	EndoGS Zhou and Jayender (2021b)	0.05	0.96	37.29
	EndoGaussian Liu et al. (2024b)	0.05	0.96	37.78
	LGS Liu et al. (2024a)	0.07	0.96	37.48
	FE-4DGS (Ours)	**0.03**	**0.97**	**39.08**
SCARED Allan et al. (2021)	EndoNeRF Wang et al. (2022)	0.40	0.77	24.35
	EndoSurf Zha et al. (2023)	0.36	0.80	25.02
	EndoGaussian Liu et al. (2024b)	0.27	**0.83**	26.89
	LGS Liu et al. (2024a)	0.30	**0.83**	27.05
	FE-4DGS (Ours)	**0.23**	**0.83**	**27.28**

Table 2: Comparison across different segmentation foundation models and two modes of FE-4DGS for binary segmentation on the EndoVis18 dataset (Allan et al., 2021). FE-4DGS (w/o F_{feat}) denotes the removal of the F_{feat} network, where semantic features do not receive additional temporal updates from the FST Deformation Module (i.e. semantic features only supervised by L1-feature loss). Values are reported as the mean ± half-width of the 95% bootstrap confidence interval (B = 10,000).

Model	IoU (↑)	DSC (↑)	Recall (↑)	Precision (↑)
SAM ViT-B	0.85 ± 0.02	0.92 ± 0.01	0.95 ± 0.01	0.89 ± 0.02
SAM ViT-H	0.86 ± 0.02	0.92 ± 0.01	0.94 ± 0.01	0.91 ± 0.02
SAM 2 (Hiera-T)	0.82 ± 0.03	0.90 ± 0.02	0.89 ± 0.01	0.91 ± 0.02
SAM 2 (Hiera-H)	0.84 ± 0.02	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.02
MedSAM ViT-B	0.84 ± 0.02	0.91 ± 0.01	0.92 ± 0.01	0.91 ± 0.02
FE-4DGS (w/o F_{feat})	0.79 ± 0.03	0.88 ± 0.02	0.86 ± 0.01	0.90 ± 0.03
FE-4DGS (Ours)	0.88 ± 0.03	0.93 ± 0.02	0.96 ± 0.01	0.91 ± 0.03

4.2. Datasets

For surgical scene reconstruction, we use the EndoNeRF (Wang et al., 2022) and SCARED (Allan et al., 2021). For segmentation tasks, we utilize the EndoVis18 (Allan et al., 2020).

We utilize the cutting and pulling sets from EndoNeRF, which consist of 2 in-vivo prostatectomy cases captured with stereo cameras from a single viewpoint. The pulling and cutting sets contain 63 and 156 frames, respectively. Following Zha et al. (2023), we select every eighth frame for testing.

The SCARED dataset comprises RGB-D scans of porcine cadaver abdominal anatomy and includes 7 sequences. In line with protocols from EndoGaussian (Liu et al., 2024b) and LGS (Liu et al., 2024a), we use a subset of 5 sequences—datasets 1, 2, 3, 6, and 7. For each selected sequence, we render the first keyframe and use all frames within that keyframe, denoted as d1k1 (197 frames), d2k1 (88 frames), d3k1 (329 frames), d6k1 (637 frames), and d7k1 (647 frames). A 7:1 train-test split is adopted, consistent with prior works (i.e., Liu et al. (2024b)).

For segmentation evaluation, we use the EndoVis18 dataset, which contains 4 sets of 149 frames of diverse surgical scenes. Sets 2-4 are used for training and set 1 is reserved for testing, including a 60-frame video to assess our integrated rendering and segmentation pipeline. Since EndoVis18 is designed for segmentation rather than scene reconstruction, it lacks per-frame camera calibrations and poses; motion blur and low frame rates further complicate reconstruction. We therefore report metrics between ground-truth semantic masks and decoded semantic features

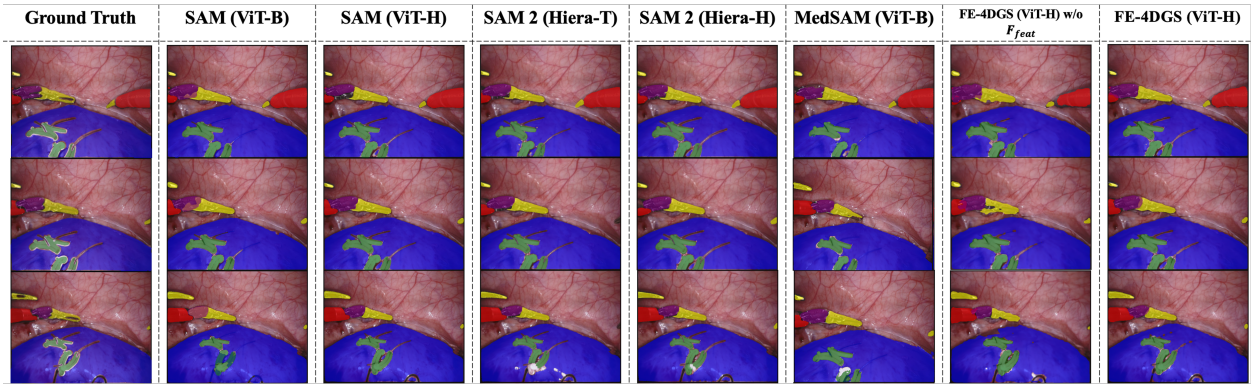

Figure 2: Qualitative segmentation comparisons. Colors correspond to the following classes: kidney (blue), small intestine (orange), instrument shaft (red), instrument clasper (yellow), instrument wrist (purple), and clamps (green). Notably, FE-4DGS (ViT-H) and SAM (ViT-H) exhibit clean multi-label segmentations, while other models struggle with finer labels such as clamps or instrument claspers.

Table 3: Weight-averaged scores for multi-label segmentation on the EndoVis18 dataset (Allan et al., 2021).

Model	IoU (↑)	DSC (↑)	Recall (↑)	Precision (↑)
SAM ViT-B	0.67 ± 0.03	0.73 ± 0.03	0.75 ± 0.02	0.79 ± 0.03
SAM ViT-H	0.68 ± 0.02	0.76 ± 0.02	0.74 ± 0.02	0.84 ± 0.01
SAM 2 (Hiera T)	0.77 ± 0.02	0.85 ± 0.01	0.88 ± 0.01	0.84 ± 0.01
SAM 2 (Hiera H)	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$
MedSAM ViT-B	0.66 ± 0.03	0.75 ± 0.02	0.86 ± 0.01	0.72 ± 0.02
FE-4DGS (w/o F_{feat})	0.63 ± 0.02	0.73 ± 0.01	0.72 ± 0.02	0.82 ± 0.01
FE-4DGS (Ours)	0.70 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.84 ± 0.01

rendered by FE-4DGS during training, and compare them against the zero-shot performance of 2D segmentation foundation models. 2 tasks are considered: multi-label segmentation with 6 classes (kidney, small intestine, instrument shaft, instrument clasper, instrument wrist, clamps) and binary segmentation with all foreground classes merged.

4.3. Main Results

We now discuss the performance of FE-4DGS in both rendering and segmentation, comparing our method against existing approaches and evaluating various ablations.

Rendering Results: Table 1 reports the quantitative results for surgical scene reconstruction. On the EndoNeRF dataset (averaged over both pulling and cutting sequences), FE-4DGS outperforms all baselines on SSIM, PSNR and LPIPS, while handling semantic feature rendering in real-time. On the SCARED dataset, FE-4DGS still achieves the highest PSNR and lowest LPIPS while matching SSIM, demonstrating robustness afforded by our semantic integration. Figure 3 shows example frames from the EndoNeRF cutting and pulling sets: the 1.3 dB PSNR gain over EndoGaussian corresponds to cleaner and more coherent textures in highly reflective regions. In the cutting example, there is also slightly improved rendering in the occluded regions which were missed in the EndoGaussian renders. Given these results, the incorporation of semantic features enhances geometric consistency during reconstruction. Additional renderings for EndoVis18 are found in Figure 5.

Rendering Speed: A key advantage of 3DGS methods (Kerbl et al., 2023) is the improved rendering speed relative to NeRF-based approaches (Mildenhall et al., 2020). Despite the additional complexity for semantic processing, FE-4DGS maintains competitive FPS and model size. On the NVIDIA RTX 5090 GPU, we achieve an average rendering speed of 287.95 FPS on the EndoNeRF dataset, at

Table 4: Ablation on removing different components of FE-4DGS.

Model	SSIM (↑)	PSNR (↑)	LPIPS (↓)	L1 Feature Loss (↓)
w/o F_{feat}	0.97	38.56	0.04	0.10
w/o L1 Feature Loss	0.97	38.43	0.04	0.18
w/o HexPlane	0.88	26.65	0.17	0.06
FE-4DGS (Ours)	**0.97**	**39.08**	**0.03**	**0.03**

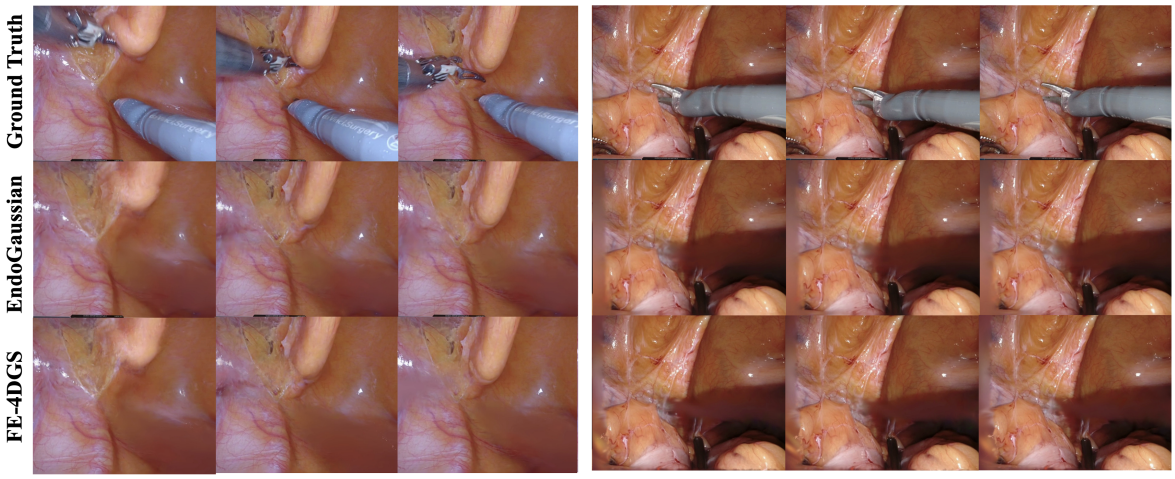

Figure 3: Comparison of qualitative renderings between EndoGaussian (Liu et al., 2024b), FE-4DGS, and ground truth on the cutting (right) and pulling (left) sets of the EndoNeRF dataset (Wang et al., 2022). Compared to previous methods, specular regions are reconstructed more finely in FE-4DGS.

Table 5: Comparison of baseline model sizes and FE-4DGS in megabytes (MB).

Model	FE-4DGS	EndoGaussian	EndoNeRF
Model Size (MB) (↓)	392.50	334.50	**13.00**

Table 6: FPS comparison across baselines on the same NVIDIA GPU.

GPU	FE-4DGS	EndoGaussian	EndoNeRF
RTX A6000 FPS (↑)	61.32	**140.36**	0.02
RTX 5090 FPS (↑)	287.95	**399.32**	0.36

392.5 MB. The tradeoff in speed is incurred because FE-4DGS performs real time rendering of RGB, depth, and semantic features in one pass, while previous works only render RGB and depth Liu et al. (2024b). Despite the compromise in rendering speed, 287.95 FPS remains clinically substantial for typical endoscopic displays. Model size and FPS comparison can be found in Table 5 and 6. We also note that FPS can be a hardware dependent metric and should not be overanalyzed.

Results on Segmentation: Tables 2 and 3 report the binary and multi-label segmentation performance on the EndoVis18 dataset. For multi-label segmentation, Table 3 shows the averaged results. We show the detailed per-class performance in Table 10 of the appendix. In the binary setting, FE-4DGS with F_{feat} achieves the best IoU, DSC, and recall scores, while maintaining competitive precision compared to state-of-the-art models such as SAM 2 (Ravi et al., 2024). For multi-label segmentation, FE-4DGS outperforms SAM (ViT-B/H) and MedSAM across almost all metrics, though SAM 2 shows superior performance in some classes. This discrepancy suggests that while FE-4DGS captures high-level semantic features effectively, further refinement is needed to encode more class-specific details. Yet FE-4DGS still delivers superior performance in binary segmentation and maintains strong performance in multi-label segmentation, outperforming its teacher model (i.e., ViT-H). This showcases the advantage of directly incorporating se-

Table 7: Reconstruction results on varying the dimensionality of the distilled semantic features.

Dim	PSNR (↑)	SSIM (↑)	LPIPS (↓)	L1 feature Loss (↓)
16	38.52	0.96	0.04	0.07
32	38.81	0.97	0.03	0.05
64	38.98	0.97	0.03	0.04
128	**39.08**	**0.97**	**0.03**	**0.03**
256	38.99	0.97	0.03	0.03

mantic features during training as opposed to simple post-hoc application of 2D segmentation models. Qualitative results are provided in Figure 2.

4.4. Ablation Study

Component Analysis: Table 4 presents an ablation study highlighting the contribution of each key component in FE-4DGS to the overall rendering quality. Removing the semantic deformation decoder F_{feat} results in degradation across all metrics, most notably in PSNR and L_1 feature loss, indicating that the absence of fine-grained semantic updates hinders the model's ability to capture detailed semantic variations. Thus, leveraging physical deformation cues encoded in extracted features, h_μ, h_R, h_S, h_o, enables coherent, physically grounded updates for semantic features, leading to improved reconstructed feature maps and better downstream segmentation performance. Similarly, omitting the per-pixel L_1 feature loss significantly degrades the alignment between the rendered and ground truth semantic features from the foundation model (SAM), as reflected by the increased L_1 error. This misalignment confirms the importance of explicit semantic supervision. Lastly, excluding the HexPlane (Cao and Johnson, 2023) leads to a drastic decline in reconstruction quality, underscoring its importance in providing deformation representations.

Feature Dimension Sweep: Table 7 varies the distilled feature dimensionality from 16 to 256. We observe a monotonic improvement in reconstruction and perceptual quality up to 128 (best PSNR/SSIM and lowest LPIPS), alongside the lowest per-pixel L_1 feature error at 128. Increasing capacity further to 256 yields marginal regression in PSNR/SSIM and a higher feature L_1, suggesting mild overparameterization under the same training budget. These results indicate that 128 provides the best accuracy–capacity trade-off for reliable semantic field distillation and downstream tasks.

Table 8: Ablation study on various net widths of HexPlane.

Net Width	PSNR (↑)	SSIM (↑)	LPIPS (↓)	L1 feature Loss (↓)
32	38.63	0.96	0.03	0.04
64	**39.08**	**0.97**	**0.03**	**0.03**

HexPlane Encoder Capacity: Table 8 compares HexPlane encoder widths of 32 and 64 while holding all other settings fixed. Doubling width from 32 to 64 consistently improves PSNR/SSIM and reduces LPIPS and the feature L_1 loss. In FE-4DGS, the encoder must represent both standard Gaussian attributes and the additional semantic feature property; the larger width reduces underfitting in this joint representation space without incurring prohibitive runtime or memory costs. We therefore adopt width = 64 as the default.

5. Conclusion

We introduced FeatureEndo-4DGS (FE-4DGS), the first system to achieve real-time surgical scene reconstruction and multi-label semantic segmentation of dynamic endoscopic videos. By distilling features from 2D segmentation foundation models into the 4D rendering process, FE-4DGS consistently improves reconstruction quality over existing baselines while maintaining comparable FPS and model size. On binary and multi-label benchmarks, it matches or surpasses state-of-the-art 2D models, performing segmentation directly during rendering.

Our main limitation is the lack of high-quality endoscopic datasets with both reconstruction and segmentation annotations. For example, EndoVis18 (Allan et al., 2020) provides masks but suffers from low image quality that hinders rendering. We hope this work motivates the collection of richer multimodal data—including text, images, and audio—that can be seamlessly integrated into our pipeline. Beyond reconstruction and segmentation, FE-4DGS also opens paths toward language-guided editing and promptable segmentation, as explored in Feature 3DGS (Zhou et al., 2024) and NeRF- (Kobayashi et al., 2022).

Acknowledgments

We thank Haohong Lin for the valuable discussions.

References

Mansoor Ali, Rafael Martinez Garcia Pena, Gilberto Ochoa Ruiz, and Sharib Ali. A comprehensive survey on recent deep learning-based methods applied to surgical data, 2023. URL https://arxiv.org/abs/2209.01435.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. 2018 robotic scene segmentation challenge, 2020. URL https://arxiv.org/abs/2001.11190.

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, Zhanshi Zhu, Huoling Luo, Fucang Jia, Xiran Zhang, Xiaohong Li, Lalith Sharan, Tom Kurmann, Sebastian Schmid, Raphael Sznitman, Dimitris Psychogyios, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. Stereo correspondence and reconstruction of endoscopic data challenge, 2021. URL https://arxiv.org/abs/2101.01133.

Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes, 2023. URL https://arxiv.org/abs/2301.09632.

Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103310. URL https://www.sciencedirect.com/science/article/pii/S1361841524002354.

Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting, 2024. URL https://arxiv.org/abs/2401.03890.

Jialei Chen, Xin Zhang, Mobarakol Islam, Francisco Vasconcelos, Danail Stoyanov, Daniel S. Elson, and Baoru Huang. Surgicalgs: Dynamic 3d gaussian splatting for accurate robotic-assisted surgical scene reconstruction, 2024b. URL https://arxiv.org/abs/2410.09292.

Long Chen, Wen Tang, Nigel W. John, Tao Ruan Wan, and Jian Jun Zhang. 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, 2018. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2018.02.006. URL https://www.sciencedirect.com/science/article/pii/S0169260717301694.

Marie-Claire J. Doornbos, Jette J. Peek, Alexander P.W.M. Maat, Jelle P. Ruurda, Pieter De Backer, Bart M.W. Cornelissen, Edris A.F. Mahtab, Amir H. Sadeghi, and Jolanda Kluin. Augmented reality implementation in minimally invasive surgery for future application in pulmonary surgery: A systematic review. *Surg. Innov.*, 31(6):646–658, 2024. doi: 10.1177/15533506241290412. URL https://doi.org/10.1177/15533506241290412.

Wei Gao and Russ Tedrake. Surfelwarp: Efficient non-volumetric single view dynamic reconstruction, 2019. URL https://arxiv.org/abs/1904.13073.

Hanxue Gu, Haoyu Dong, Jichen Yang, and Maciej A. Mazurowski. How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with segment anything model. *Machine Learning for Biomedical Imaging*, 3(May 2025):88–120, May 2025. ISSN 2766-905X. doi: 10.59275/j.melba.2025-86a6. URL http://dx.doi.org/10.59275/j.melba.2025-86a6.

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

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation, 2022. URL https://arxiv.org/abs/2205.15585.

Huiqian Li, Dingwen Zhang, Jieru Yao, Longfei Han, Zhongyu Li, and Junwei Han. Asps: Augmented segment anything model for polyp segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 118–128, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72114-4.

Yang Li, Florian Richter, Jingpei Lu, Emily K. Funk, Ryan K. Orosco, Jianke Zhu, and Michael C. Yip. Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics. *IEEE Robotics and Automation Letters*, 5(2):2294–2301, April 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2970659. URL http://dx.doi.org/10.1109/LRA.2020.2970659.

Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, 2021. URL https://arxiv.org/abs/2011.02910.

Hengyu Liu, Yifan Liu, Chenxin Li, Wuyang Li, and Yixuan Yuan. Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 660–670, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-72384-1.

Xingtong Liu, Maia Stiber, Jindan Huang, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. Reconstructing sinus anatomy from endoscopic video – towards a radiation-free approach for quantitative longitudinal assessment, 2020. URL https://arxiv.org/abs/2003.08502.

Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. Endogaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction, 2024b. URL https://arxiv.org/abs/2401.12561.

Yonghao Long, Zhaoshuo Li, Chi Hang Yee, Chi Fai Ng, Russell H. Taylor, Mathias Unberath, and Qi Dou. E-dssr: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception, 2021. URL https://arxiv.org/abs/2107.00229.

Jingpei Lu, Ambareesh Jayakumari, Florian Richter, Yang Li, and Michael C. Yip. Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction, 2021. URL https://arxiv.org/abs/2003.03472.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections, 2021. URL https://arxiv.org/abs/2008.02268.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. URL https://arxiv.org/abs/1905.00889.

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, 2020. URL https://arxiv.org/abs/2003.08934.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields, 2021. URL https://arxiv.org/abs/2011.12100.

Veronica Penza, Elena De Momi, Nima Enayati, Thibaud Chupin, Jesús Ortiz, and Leonardo S. Mattos. Envisors: Enhanced vision system for robotic surgery. a user-defined safety volume tracking to minimize the risk of intraoperative bleeding. *Frontiers in Robotics and AI*, 4, 2017. ISSN 2296-9144. doi: 10.3389/frobt.2017.00015. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2017.00015.

AKM Shahariar Azad Rabby and Chengcui Zhang. Beyondpixels: A comprehensive review of the evolution of neural radiance fields, 2024. URL https://arxiv.org/abs/2306.03000.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields, 2022. URL https://arxiv.org/abs/2212.09802.

Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters*, 3(1):155–162, January 2018. ISSN 2377-3774. doi: 10.1109/lra.2017.2735487. URL http://dx.doi.org/10.1109/LRA.2017.2735487.

Aliza Subedi, Smriti Regmi, Nisha Regmi, Bhumi Bhusal, Ulas Bagci, and Debesh Jha. Classification of endoscopy and video capsule images using cnn-transformer model, 2024. URL https://arxiv.org/abs/2408.10733.

Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting, 2024. URL https://arxiv.org/abs/2403.15124.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery, 2022. URL https://arxiv.org/abs/2206.15255.

Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43996-4.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering, 2024. URL https://arxiv.org/abs/2310.08528.

Weixing Xie, Junfeng Yao, Xianpeng Cao, Qiqin Lin, Zerui Tang, Xiao Dong, and Xiaohu Guo. Surgicalgaussian: Deformable 3d gaussians for high-fidelity surgical scene reconstruction, 2024. URL https://arxiv.org/abs/2407.05023.

Chen Yang, Kailing Wang, Yuehao Wang, Xiaokang Yang, and Wei Shen. Neural lerplane representations for fast 4d reconstruction of deformable tissues, 2023. URL https://arxiv.org/abs/2305.19906.

Jialin Yu, Huogen Wang, and Ming Chen. Colonoscopy polyp detection with massive endoscopic images, 2022. URL https://arxiv.org/abs/2202.08730.

Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos, 2023. URL https://arxiv.org/abs/2307.11307.

Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation, 2021. URL https://arxiv.org/abs/2103.15875.

Haoyin Zhou and Jagadeesan Jayender. Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 331–340, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-87202-1.

Haoyin Zhou and Jagadeesan Jayender. Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 12904:331–340, 2021b. URL https://api.semanticscholar.org/CorpusID:237621783.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields, 2024. URL https://arxiv.org/abs/2312.03203.

Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 324–342, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72658-3.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2, 2024. URL https://arxiv.org/abs/2408.00874.

Appendix A. Feature-Spatiotemporal (FST) Deformation Module

This supplementary material provides additional implementation and dataset details not included in the main paper.

The FST deformation module is detailed in Figure 4, which provides a visual supplement of Section 3.2. The implementation details for FST are as follows: for the HexPlane encoder, we follow the same configuration as given in EndoGaussian Liu et al. (2024b). That is, grid dimension of size 2, input coordinate dimension of size 4, output coordinate dimension of size 32, and resolutions of [64,64,64,100]. For the HexPlane decoder MLP, we set the model width W to 64 and the depth D to 8. Each feature extractor follows the same structure consisting of 2 linear layers that downsample the embeddings to W//2 before passing into deformation heads. The semantic feature MLP is structured similarly, with 2 linear layers where the second layer projects the embedding back into the feature size of 128 corresponding to the number of semantic channels we use in practice.

Appendix B. Dataset Details

EndoNeRF (Wang et al., 2022) comprises 2 in-vivo prostatectomy cases captured with stereo cameras from a single viewpoint. The dataset features non-rigid deformations and tool occlusions. We also use the Stereo Correspondence and Reconstruction of Endoscopic Data (SCARED) (Allan et al., 2021) dataset from the 2022 Endoscopic Vision Challenge, which consists of RGB-D scans of porcine cadaver abdominal anatomy collected using a da Vinci Xi endoscope and a projector.

Within the EndoNeRF dataset, the pulling video contains 63 frames. We select every eighth frame for testing, resulting in 58 training frames and 5 test frames. The cutting video comprises 156 frames and is split in a similar manner, yielding 136 training frames and 20 test frames.

For segmentation, we employ the Robotic Scene Segmentation Sub-Challenge from the 2018 Endoscopic Vision Challenge (EndoVis18) (Allan et al., 2020). This dataset contains 4 sequences, each with 149 frames depicting diverse surgical scenes with annotations for various anatomical and tool components. In total, there are 10 labels in this multi-label segmentation dataset: background tissue, instrument shaft, instrument clasper, instrument wrist, kidney parenchyma, covered kidney, thread, clamps, suturing needles, suction instrument, and small intestine. The original images are sized at 1280×1024 pixels and are downsampled to 640×512 pixels for training. Although the original EndoVis18 videos were captured at 60 Hz, they were subsampled to 2 Hz to

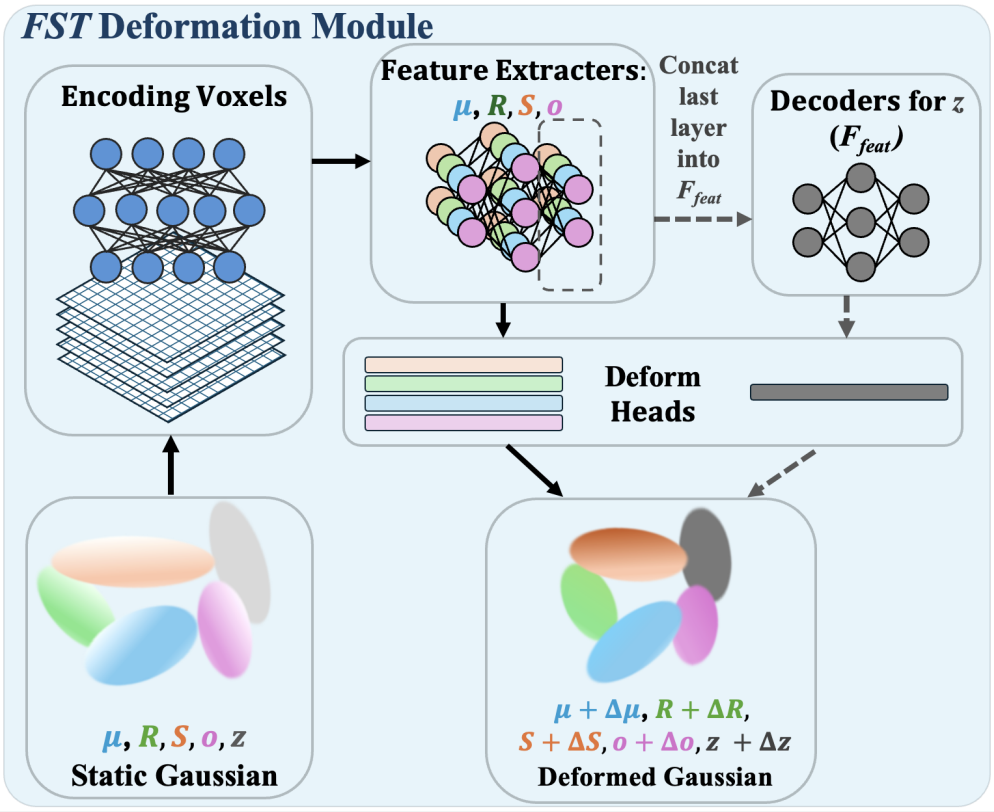

Figure 4: Overview of the FST feature deformation module found in FE-4DGS, it includes the HexPlanes (Cao and Johnson, 2023) and requires decoder architectures for updating position, rotation, scale, opacity, and semantic features of Gaussians.

reduce labeling efforts. Furthermore, sequences with minimal movement were removed, resulting in the final 149-frame video. These postprocessing steps present significant challenges for high-fidelity surgical scene reconstruction. To prepare these images for FE-4DGS, we first obtain stereo depth masks using the lite stereo transformer (Li et al., 2021). Additionally, we prepare a camera poses and intrinsics file in the LLFF (Local Light Field Fusion) format (Mildenhall et al., 2019). As a simplification, we assume a single viewpoint and set all camera poses to the identity to avoid interference from ill-calibrated poses.

To demonstrate the utility of FE-4DGS for segmentation, we select 60 frames from sequence 1, which depict a continuous surgical scene of the kidney. In these frames, a clip applier tool is used, and parts of the small intestine appear in the foreground. These frames were chosen for training because the scene contains multiple key segmentation targets, including the kidney, small intestine, clamps, and all components of grasping instruments (i.e., instrument head, instrument wrist, and instrument clasper). These targets vary in size, segmentation complexity, and instance count per frame, presenting a challenging multi-label segmentation task suitable for evaluating our model. Due to substantial tissue and tool movements and a camera repositioning midway through the scene, we split these 60 frames into 2 segments for rendering. The first segment, comprising 33 frames, is processed through our FE-4DGS pipeline, while the latter 27 frames are rendered as a separate scene. Aside from the mid-scene movements, the camera remains fairly stable, with smooth transitions and consistent overlap in scene coverage, ensuring a well-constrained reconstruction. During segmentation, we aggregate the feature maps from these 2 segments to benchmark against standalone SAM models.

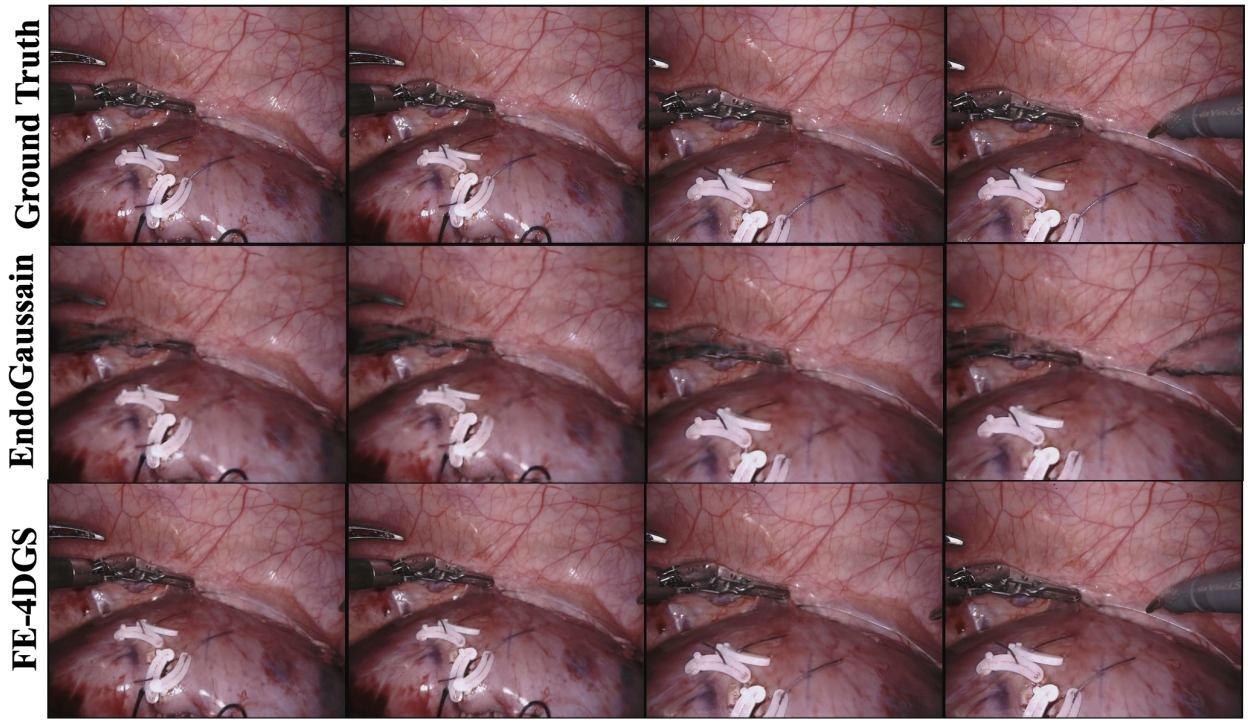

Figure 5: Comparison of qualitative renderings between EndoGaussian (Liu et al., 2024b), FE-4DGS, and ground truth on the EndoVis18 dataset (Allan et al., 2020). We can see that across all scenes, FE-4DGS and EndoGaussian experience similar performances, although FE-4DGS is able to capture semantic features which can later be used for segmentation.

Appendix C. Additional Details of Integrating Semantic Features into FE-4DGS

This supplementary material provides additional details on integrating semantic features into FE-4DGS that were not included in the main paper.

For segmentation experiments, we use SAM's pretrained ViT-H encoder to generate a ground truth feature map for each frame in our video snippet. As detailed in SAM (Kirillov et al., 2023), the image encoder is pretrained via MAE and employs 14×14 windowed attention along with 4 equally spaced global attention blocks. The resulting 256 feature maps have dimensions of 64×64, which are then resized to 51×64 to match the aspect ratio of the training image. This resizing is accomplished by simply cropping the feature map height, as done in Feature 3DGS (Zhou et al., 2024). Since only the 51×64 portion of the 64×64 feature map contains meaningful semantic information and the rest is padding, we crop the side corresponding to the longer dimension of the original image.

For segmentation, we use the same configuration as with the EndoNeRF cutting dataset. We note that no hyperparameter tuning has been performed for the EndoVis dataset, and performance metrics may be further improved with additional tuning. After running FE-4DGS, the generated feature maps are directly fed into the pretrained SAM decoder to produce multi-label segmentations. Prior to this step, all feature maps are restored to their original aspect ratio of 64×64 via bilinear interpolation. For the segmentation prompt, bounding boxes computed from the ground truth mask for each class are used. If a class appears in multiple masks within an image, bounding boxes are computed for each instance.

Appendix D. Fine-tuning SAM Details

We use the SAM model with the ViT-B encoder and standard pretrained weights. Instead of fully fine-tuning all layers, we adopt an "adapter" training strategy (Gu et al., 2025), fine-tuning only the adapter layer parameters while keeping the rest of the SAM backbone frozen. This significantly reduces VRAM consumption during training, allowing us to work within our limited computational resources. For training, we use sequences 2, 3, and 4 from EndoVis18, each containing 149 frames, for a total of 447 images. Sequence 1, which also contains 149 images, is reserved for validation and is excluded from training because it is later used for zero-shot segmentation in Tables 2 and 3.

The preprocessing steps are straightforward: images are normalized and resized. The loss function is defined as the sum of the dice loss and the cross entropy. We employ an Adam optimizer with a base learning rate of 1×10^{-3} and a StepLR scheduler with a step size of 10 and $\gamma = 0.5$. The batch size is set to 4, and validation is performed every other epoch. Hyperparameters are not tuned; training continues until the validation loss does not decrease for more than 20 epochs. In total, we train for 66 epochs, achieving the best validation score of 0.955 at epoch 46.

Subsequently, we extract embeddings from this fine-tuned model to serve as the ground truth feature embeddings for our FE-4DGS pipeline. Zero-shot segmentation is performed using the semantic feature maps generated by FE-4DGS on 60 selected image frames from EndoVis18 for binary segmentation. For this experiment, we follow the same 7:1 train-test split as in EndoNeRF, which enables us to obtain performance metrics on the test set. Segmentation performance is measured via micro-averaging across all classes on the train image frames, with each label weighted by its mask size, and we observe a marked improvement when using these fine-tuned features compared to the baseline.

Table 9 compares segmentation and rendering performance using the original versus the fine-tuned SAM model (trained on EndoVis18 sets 2, 3, and 4). Fine-tuning significantly improves segmentation metrics—IoU and DSC—demonstrating that adapting SAM to the domain-specific characteristics of endoscopic scenes yields more accurate semantic cues. These enhanced semantic features, in turn, facilitate better guidance during rendering, as reflected by SSIM and PSNR scores. Overall, the results affirm that task-specific refinement of the segmentation backbone is important in boosting both segmentation fidelity and the subsequent rendering quality.

Table 9: Fine-tuning SAM (EndoVis18 (Allan et al., 2020)).

Finetuned	IoU (↑)	DSC (↑)	SSIM (↑)	PSNR (↑)
X	0.60	0.73	0.67	20.43
✓	**0.69**	**0.81**	**0.67**	**20.77**

Appendix E. Additional Results

Table 10: Class-wise multi-label segmentation results on the EndoVis18 dataset.

Model	Class	IoU (↑)	DSC (↑)	Recall (↑)	Precision (↑)
FE-4DGS (ViT-H)	kidney	0.87	0.93	0.96	0.90
	small intestine	0.83	0.90	0.91	0.90
	instrument shaft	0.34	0.38	0.36	0.67
	instrument clasper	0.26	0.36	0.32	0.66
	clamps	0.27	0.40	0.31	0.70
	instrument wrist	0.69	0.81	0.91	0.75
SAM (ViT-H)	kidney	0.85	0.91	0.92	0.92
	small intestine	0.85	0.92	0.93	0.90
	instrument shaft	0.32	0.37	0.34	0.66
	instrument clasper	0.26	0.36	0.31	0.65
	clamps	0.28	0.40	0.32	0.69
	instrument wrist	0.76	0.85	0.87	0.85
SAM (ViT-B)	kidney	0.80	0.89	0.93	0.86
	small intestine	0.85	0.92	0.93	0.90
	instrument shaft	0.32	0.37	0.34	0.60
	instrument clasper	0.26	0.36	0.32	0.63
	clamps	0.28	0.40	0.32	0.69
	instrument wrist	0.76	0.85	0.86	0.85
MedSAM (ViT-B)	kidney	0.83	0.90	0.90	0.91
	small intestine	0.84	0.91	0.94	0.88
	instrument shaft	0.32	0.36	0.34	0.74
	instrument clasper	0.25	0.35	0.30	0.74
	clamps	0.24	0.36	0.28	0.73
	instrument wrist	0.71	0.81	0.79	0.89
SAM 2 (Hiera-T)	kidney	0.85	0.92	0.90	0.94
	small intestine	0.83	0.88	0.87	0.91
	instrument shaft	0.84	0.90	0.93	0.90
	instrument clasper	0.36	0.48	0.59	0.58
	instrument wrist	0.40	0.46	0.49	0.45
	clamps	0.66	0.79	0.88	0.72
SAM 2 (Hiera-H)	kidney	0.87	0.93	0.92	0.94
	small intestine	0.90	0.94	0.95	0.94
	instrument shaft	0.83	0.90	0.92	0.90
	instrument clasper	0.69	0.67	0.84	0.69
	instrument wrist	0.38	0.45	0.50	0.43
	clamps	0.66	0.79	0.87	0.72

E.1. Class-wise Multi-label Segmentation Results

We provide additional results on each individual class present in the EndoVis18 dataset in Table 10.

E.2. Teacher Model Choice

Table 12 evaluates various segmentation teacher models for semantic supervision. We find that all mod-

Table 11: Hyperparameter Settings for pulling and cutting sets from EndoNeRF (Wang et al., 2022) and the SCARED dataset (Allan et al., 2021).

Hyperparameter	**EndoNeRF Pulling**	**EndoNeRF Cutting**	**SCARED**
Initial Points	90,000	90,000	30,000
Grid LR (Initial / Final)	0.0032 / 0.0000032	0.0016 / 0.0000016	0.0016 / 0.000016
Deformation LR (Initial / Final)	0.00016 / 1.6e-7	0.0004 / 4e-7	0.00008 / 0.0000008
Position LR (Initial / Final)	0.00016 / 0.0000016	0.00016 / 0.0000016	0.00016 / 0.0000016
Iterations (Coarse / Fine)	1000 / 6000	1000 / 6000	1000 / 3000
Percent Dense	0.01	0.01	0.01
Opacity Reset Interval	6000	6000	3000
Prune Interval	6000	6000	3000
Position LR Max Steps	7000	7000	3000
Deformation LR Delay Multiplier	0.01	0.01	0.01
Grid Dimensions	2	2	2
Input Coordinate Dim	4	4	4
Output Coordinate Dim	64	64	32
Multiresolution Levels	[1,2,4,8]	[1, 2, 4, 8]	[1, 2, 4, 8]

Table 12: Reconstruction performances on various segmentation teacher models.

Teacher	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**L1-feature Loss (↓)**
SAM ViT-B	38.97	0.97	0.03	0.03
SAM ViT-H	**39.08**	**0.97**	**0.03**	**0.03**
MedSAM-B	39.04	0.97	0.03	0.03

els achieve comparable performance, with SAM ViT-H showing a slightly higher PSNR than the others. Therefore, we primarily use SAM ViT-H as the teacher model.

Appendix F. Additional Visualizations

F.1. Additional Rendering Visualizations

We provide additional visualizations of rendering on the EndoVis18 dataset in Figure 5.

Appendix G. Training Hyperparameters

In Table 11, we present the hyperparameter configurations used in our experiments for the EndoNeRF and SCARED datasets. Most other hyperparameters remain fixed at the default values provided in the code.

We observed stability issues during the fine stage on the EndoNeRF dataset, likely due to our model's enhanced rendering performance during the coarse stage, which is attributed to our revamped rasterizer. An excessively high PSNR during the coarse phase—reaching up to 50—indicated overfitting during initialization with a single image, thereby hindering subsequent deformation optimization. Although the first 100–300 iterations of the fine stage showed promising convergence, the model occasionally experienced an abrupt drop in PSNR, sometimes as low as 5. To maintain stable training dynamics and ensure generalizable deformation capabilities, it was necessary to adjust hyperparameters, particularly by reducing the learning rates and fine-tuning gradient propagation strategies. Furthermore, the coarse initialization stage must be capped to conclude at a lower PSNR to ensure stability during the fine deformation stage.