HFGS: 4D Gaussian Splatting with Emphasis on Spatial and Temporal High-Frequency Components for Endoscopic Scene Reconstruction

Haoyu Zhao*[1]
haoyu.zhao@whu.edu.cn
Xingyue Zhao*[2]
zhaoxingyue@stu.xjtu.edu.cn
Lingting Zhu[3]
ltzhu99@connect.hku.hk
Weixi Zheng[1]
acnicotine@whu.edu.cn
Yongchao Xu[1]
yongchao.xu@whu.edu.cn

[1] School of Computer Science,
Wuhan University,
Hubei, P. R. China

[2] School of Software Engineering,
Xi'an Jiaotong University,
Shaanxi, P. R. China

[3] School of Computing and Data Science,
The University of Hong Kong,
Hong Kong, P.R. China

Abstract

Robot-assisted minimally invasive surgery benefits from enhancing dynamic scene reconstruction, as it improves surgical outcomes. While Neural Radiance Fields (NeRF) have been effective in scene reconstruction, their slow inference speeds and lengthy training durations limit their applicability. To overcome these limitations, 3D Gaussian Splatting (3D-GS) based methods have emerged as a recent trend, offering rapid inference capabilities and superior 3D quality. However, these methods still struggle with under-reconstruction in both static and dynamic scenes. In this paper, we propose **HFGS**, a novel approach for deformable endoscopic reconstruction that addresses these challenges from spatial and temporal frequency perspectives. Our approach incorporates deformation fields to better handle dynamic scenes and introduces Spatial High-Frequency Emphasis Reconstruction (SHF) to minimize discrepancies in spatial frequency spectra between the rendered image and its ground truth. Additionally, we introduce Temporal High-Frequency Emphasis Reconstruction (THF) to enhance dynamic awareness in neural rendering by leveraging flow priors, focusing optimization on motion-intensive parts. Extensive experiments on two widely used benchmarks demonstrate that **HFGS** achieves superior rendering quality. Source code is available at https://github.com/zhaohaoyu376/HFGS.

*Equal contributions.

1 Introduction

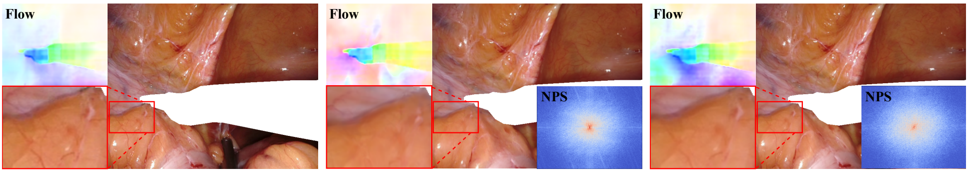
Endoscopic procedures are foundational to minimally invasive surgery, significantly reducing trauma and hastening patient recovery [6, 21, 25]. In Robotic-Assisted Minimally Invasive Surgery (RAMIS), the reconstruction of a 3D model of the surgical scene from stereo endoscopes is critical for surgical precision and efficiency. This technology enables surgeons to visualize observed tissues in 3D, which enhances their spatial awareness and navigation capabilities [15, 26]. Despite the many benefits of endoscopic reconstruction, several challenges remain, including limited field-of-view, obstructions, and dynamic tissue deformation [24, 28, 32, 39]. Previous studies [9, 13, 28, 32, 39, 42] have successfully employed depth maps for endoscopic reconstruction; however, these methods still face two significant issues: the lack of sufficient details in generated models and inadequate rendering of non-rigid deformations.

Recent advancements in endoscopic 3D reconstruction have been significantly enhanced by Neural Radiance Fields (NeRFs) [17]. EndoNeRF [28], a pioneering work, is the first to apply NeRF to endoscopic scenes for reconstructing deformable tissues using dual neural fields. Another approach, EndoSurf [39], utilizes the signed distance field (SDF) [27, 37] to regulate surface geometry. Although these methods produce satisfactory outcomes, they demand extensive computational resources. Rendering each image necessitates querying radiance fields at numerous points and rays, which significantly limits rendering speed and poses considerable challenges for practical applications, such as intraoperative use [4].

To address these issues, 3D Gaussian Splatting (3D-GS) [10] emerges as an effective alternative, providing rapid inference capability and enhanced quality of 3D representation. By optimizing anisotropic 3D Gaussians with a collection of scene images, 3D-GS effectively captures the spatial positioning, orientations, color properties, and alpha blending factors. 3D-GS reconstructs not only the geometry but also the visual texture of scenes with rapid rendering performances [10]. Although 3D-GS is extended to represent dynamic scenes [9, 13, 14, 35, 42], it often suffers from under-reconstruction [10] during the process of Gaussian densification [40], which affects both static and dynamic scenes. The under-reconstruction can be clearly observed with blur and artifacts in the rendered 2D images, the discrepancy of frequency spectrum of the render images and the corresponding ground truth, and the predicted optical flow results by [30], as illustrated in Fig. 1.

In this paper, we present an innovative method called **HFGS** for deformable endoscopic tissues reconstruction that addresses the under-reconstruction from both spatial and temporal frequency perspectives. Specifically, we propose a module called Spatial High-Frequency Emphasis Reconstruction (SHF), which minimizes the discrepancy in frequency spectra between the rendered image and the corresponding ground truth by focusing specifically on spatial high-frequency components of images. Additionally, we propose Temporal High-Frequency Emphasis Reconstruction (THF) module which enhances dynamic awareness in neural rendering by utilizing flow priors. This module targets motion areas identified through flow-based methods as temporal high-frequency components during optimization, thus improving the fidelity of moving tissues.

To summarize, our main contributions are three-fold: 1) We propose a Frequency Regularization Module to reduce spectral mismatches between rendered images and ground truth images, thereby improving in frequency space. 2) We introduce a novel module which offers dynamic awareness to existing regularization in neural rendering with the help of flow prior, providing special attention to the motion parts during optimization. 3) Experiments over multiple benchmarks show that **HFGS** achieves superior performances.



(a) GT (b) EndoGS [42] (c) HFGS

Figure 1: For the sample image from ENDONERF [28], (a-c) show the rendered image, the noise power spectrum (NPS) where blue indicates it is closer to GT, and optical flow predictions based on adjacent frame. Our **HFGS** not only achieves the best results, reconstructing the most detailed information and exhibiting the bluest NPS, but also renders images with optical flow that are closer to the GT.

2 Related Works

Neural Rendering for Dynamic Scenes. Neural Radiance Fields (NeRF) [17] marks a significant advance in high-quality neural rendering. Several efforts aim to adapt NeRF for dynamic scenes. For instance, some works [11, 12] integrate NeRF with time-conditioned latent codes to effectively represent dynamic scenes. Another group of works [19, 20, 22] incorporate an explicit deformation field that bends rays as they pass through various targets into a canonical space.

Some works [28, 32, 39] use NeRF to represent dynamic endoscopic scenes. A good representative is EndoNeRF [28] which follows the modeling of D-NeRF [22]. It trains two neural fields: one for tissue deformation and the other for canonical density and color. EndoNeRF can synthesize reasonable RGB images with post-processing filters. To tackle the lengthy training time requirement, LerPlane [32] constructs a 4D volume by introducing 1D time to the existing 3D spatial space. Although significantly accelerating the training process, they still cannot meet the practical need of rendering speed.

Reconstruction with 3D Gaussian Splatting. Additionally, 3D-GS [10] is notable for its pure explicit representation and differential point-based splatting method, enabling real-time rendering of novel views through a customized CUDA-based differentiable Gaussian rasterization pipeline. The application of 3D-GS in dynamic reconstruction is just beginning to unfold. D-3DGS [14] is proposed as the first attempt to adapt 3D-GS into a dynamic setup. Other works [29, 34, 36] model 3D Gaussian motions with a compact network or 4D primitives, resulting in highly efficient training and real-time rendering.

Some works [13, 42] make the first attempts to apply 3D-GS to represent dynamic endoscopic scenes. EndoGS [42] employs surface-aligned Gaussian Splatting [7] to reconstructing deformable endoscopic tissues. EndoGaussian [13] introduces Holistic Gaussian Initialization (HGI) and Spatio-temporal Gaussian Tracking (SGT) to initialize dense Gaussians and model surface dynamics, respectively. However, these approaches often suffers from under-reconstruction [10] during the process of Gaussian densification [40]. In contrast, our method also models 3D Gaussian motions with a deformation network for deformable endoscopic tissues reconstruction but addresses the under-reconstruction from both spatial and temporal frequency perspectives.

3 Preliminary

3.1 3D Gaussian Splatting

3D Gaussian Splatting (3D-GS) [10] explicitly represents scenes using point clouds, where each point is modeled as a 3D Gaussian defined by a covariance matrix Σ and a center point $\mathcal{X}$, the latter referred to as the mean. The value at point $\mathcal{X}$ is $G(X) = e^{-\frac{1}{2}\mathcal{X}^T\Sigma^{-1}\mathcal{X}}$. For differentiable optimization, the covariance matrix Σ is decomposed into a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$, such that $\Sigma = \mathbf{RSS}^T\mathbf{R}^T$.

In rendering novel views, differential splatting as introduced by [38] and [43], involves using a viewing transform W and the Jacobian matrix J of the affine approximation of the projective transformation to compute the transformed covariance matrix: $\Sigma' = JW\Sigma W^T J^T$. Each 3D Gaussian is characterized by several attributes: position $\mathcal{X} \in \mathbb{R}^3$, color defined by spherical harmonic (SH) coefficients $\mathcal{C} \in \mathbb{R}^k$ (where k is the number of SH functions), opacity $\alpha \in \mathbb{R}$, rotation factor $r \in \mathbb{R}^4$, and scaling factor $s \in \mathbb{R}^3$. The color and opacity at each pixel are computed from the Gaussian's representation $G(X) = e^{-\frac{1}{2}\mathcal{X}^T\Sigma^{-1}\mathcal{X}}$. The blending of N ordered points overlapping a pixel is given by the formula:

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

Here, c_i, α_i represent the density and color of this point computed by a 3D Gaussian G with covariance Σ multiplied by an optimizable per-point opacity and SH color coefficients.

3.2 Dynamic Gaussian Splatting with Deformation Fields

In our representation of a surgical scene as a 4-dimensional volume, the deformation of tissues is modeled over time. We adopt Gaussian deformation to represent the time-varying motions and shapes, based on the designs of [29]. Our primary objective is to accurately learn both the static parameters, $\{(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{sh}, \sigma)\}$ and dynamic parameters, $\{\Delta(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{sh}, \sigma)\}$ of the 3D Gaussians. For each 3D Gaussian, we compute the deformation using the the mean $\boldsymbol{\mu} = (x, y, z)$ and the time t. We encode the spatial and temporal information using six orthogonal feature planes [2, 5, 29, 31, 32]. Specifically, the multi-resolution HexPlane [2, 5] consists of three spatial planes XY, XZ, YZ and three spatial-temporal planes XT, YT, ZT. These planes encode features $F \in \mathbb{R}^{h\times N_1\times N_2}$, where h represents the hidden dimension and N_1, N_2 indicate the plane resolution. We utilize bilinear interpolation $\mathcal{B}$ to interpolate the four nearby queried voxel features. Voxel feature can be represented in the format of matrix element-wise multiplication with operation $\odot$:

$$f_{voxel}(\boldsymbol{\mu}, t) = \mathcal{B}(F_{XY}, x, y) \odot \mathcal{B}(F_{YZ}, y, z) \dots \mathcal{B}(F_{YT}, y, \tau) \odot \mathcal{B}(F_{ZT}, z, \tau). \tag{2}$$

We employ a single MLP to update the attributes of Gaussian. This MLP integrates all the information to decode various parameters such as the position, scaling factor, rotation factor, spherical harmonic coefficients, and opacity:

$$\Delta(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{sh}, \sigma) = \mathrm{MLP}(f_{voxel}(\boldsymbol{\mu}, t)). \tag{3}$$

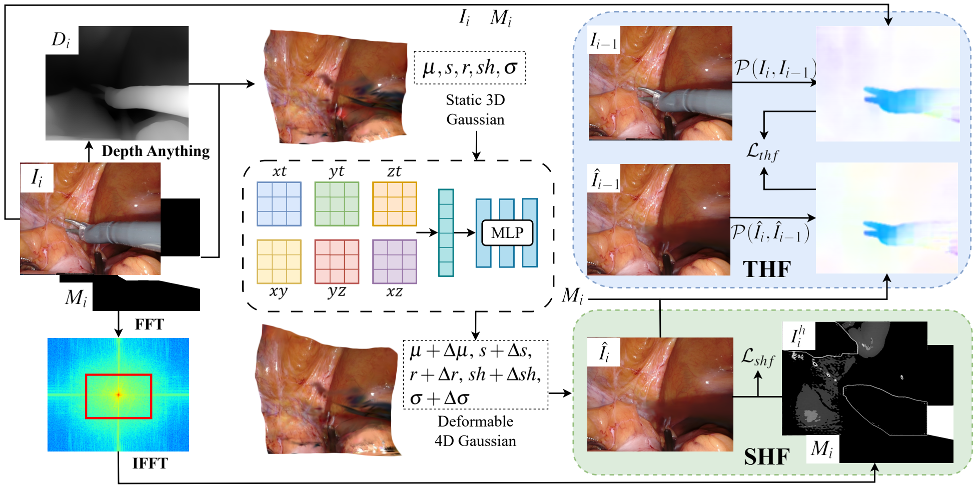


Figure 2: Pipeline of the proposed **HFGS**. We utilize monocular images, estimated depths from Depth-Anything [33] and tool masks for training [9]. A single MLP is used to derive the deformation associated with these 3D Gaussian, given the features queried via voxel planes. Then we address the under-reconstruction by emphasizing spatial and temporal high-frequency components.

Using the mean and time as inputs, we compute features for 3D Gaussians by querying multi-resolution voxel planes. A single MLP is then employed to derive the deformations of these Gaussians. Through differentiable rasterization, we generate rendered images and depth maps. The accuracy of these outputs is validated using ground truth images, depth maps, and tool masks, which serve as the basis for supervision.

4 Method

Given a single-viewpoint stereo video of a dynamic surgical scene, we aim to reconstruct 3D structures and textures of surgical scenes without occlusion of surgical instruments. We denote $\{(I_i, D_i, M_i)\}_{i=1}^{T}$ as a sequence of input stereo video frames, where T is the total number of frames, $I_i \in \mathbb{R}^{H\times W\times 3}$ and $D_i \in \mathbb{R}^{H\times W}$ is the i-th left RGB image and depth map with height H and width W. Mask M_i is utilized to specifically exclude pixels from surgical tools. Time of the i-th frame is i/T. We formulate our solution with a probabilistic model to learn statistics for depth from Depth-Anything [33].

In this paper, we present an innovative method called **HFGS** for deformable endoscopic tissues reconstruction that addresses the under-reconstruction by emphasizing spatial and temporal high-frequency components, as shown in Fig. 2. We first introduce Spatial High-Frequency Emphasis Reconstruction (SHF) in Section 4.1 which minimizes differences in the spatial frequency spectra of the rendered image and the corresponding ground truth by focusing specifically on spatial high-frequency components. We then introduce Temporal High-Frequency Emphasis Reconstruction (THF) in Section 4.2 which enhances dynamic awareness in neural rendering by utilizing a flow prior. This module targets motion areas identified through flow-based methods as temporal high-frequency components during optimization, thus improving the fidelity of moving tissues. Finally, we describe the optimization process in Section 4.3.

4.1 Spatial High-Frequency Emphasis Reconstruction

In naive pixel-wise L_1 loss implementations, the average gradient might be quite small, which can occur even in regions that are not well-reconstructed, which misleads the Gaussian densification [40]. As Gaussian densification is not applied to Gaussians with small gradients [10], these Gaussians cannot be densified through splitting into smaller Gaussians, leading to under-reconstruction. Spatial high-frequency components focus on object structures resembling identity [18, 41]. Thus, it is reasonable to guide the Gaussian densification by applying regularization in spatial frequency domain. For an ground truth image I_i, its frequency space signal $\mathcal{F}(I_i)$ can be obtained with Fast Fourier Transform (FFT), which is defined as follows:

$$\mathcal{F}(I_i)(u,v,c) = \sum_{h=1}^{H}\sum_{w=1}^{W} I_i(h,w,c)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)} = \mathcal{A}(I_i)e^{j\mathcal{P}(I_i)}, \tag{4}$$

where $j^2 = -1$, and $\mathcal{A}(I_i)$ and $\mathcal{P}(I_i)$ refer to the amplitude and phase spectra of I_i, respectively. We center the low-frequency components within the frequency spectrum, and then introduce a binary mask $\mathcal{B} \in \mathbb{R}^{H\times W}$, where all values are zero except in the central region. Following [41], the high-frequency components $\mathcal{A}_h(I_i)$ are given by:

$$\mathcal{A}_h(I_i) = (I-\mathcal{B}) \odot \mathcal{A}(I_i), I_i^h = \mathcal{F}^{-1}(\mathcal{A}_h(I_i)), \tag{5}$$

where $\odot$ denotes element-wise multiplication and $\mathcal{F}^{-1}$ means the Inverse Fast Fourier Transform (IFFT). We then obtain the spatial domain representation of the image I_i^h that contains only the spatial high-frequency components. The $\mathcal{L}_{shf}$ is defined as:

$$\mathcal{L}_{shf}(I_i,\hat{I}_i) = \sum_{x=1}^{W}\sum_{y=1}^{H} I_i^h(x,y)\cdot|I_i(x,y)-\hat{I}_i(x,y)|, \tag{6}$$

where $\hat{I}_i$ means the rendered image.

4.2 Temporal High-Frequency Emphasis Reconstruction

To extend 3D-GS [10] to dynamic scenes, Gaussian motions and shape changes are modeled using a Gaussian deformation field network, as discussed in [13, 29, 42]. However, these methods struggle to effectively render dynamic images in scenarios with rapid movement, such as in 3D dynamic endoscopic scene reconstruction [8]. This limitation stems from their insufficient use of the abundant motion data available from 2D observations. To address this, we propose a module called Temporal High-Frequency Emphasis Reconstruction (THF), which applies regularization in the temporal frequency domain of the deformation field network. This module enhances dynamic awareness in neural rendering by incorporating a flow prior. This flow prior is designed to prioritize regions exhibiting more significant movements in the current frame, thereby improving the rendering of dynamic scenes.

We feed both the rendered image $\hat{I}_i$ along with its adjacent frame $\hat{I}_{i-1}$, and the corresponding ground truth image I_i with its adjacent frame I_{i-1} into a pre-trained predictor $\mathcal{P}$ [30]. For the first frame, we treat its adjacent frame as the frame itself. This process is used to predict the optical flows $\hat{f}_i$ and f_i.

$$\hat{f}_i = \mathcal{P}(\hat{I}_i, \hat{I}_{i-1}), f_i = \mathcal{P}(I_i, I_{i-1}). \tag{7}$$

We define the loss $\mathcal{L}_{thf}$ as the sum of the Charbonnier loss [3] and the census loss [16], $\mathcal{L}_{thf} = \mathcal{L}_{char} + \mathcal{L}_{cen}$, which improves the quality of interpolation, making it more resilient to outliers and structural variations in the scene.

4.3 Optimization

In reconstructing from videos with tool occlusion, we face challenges similar to [28, 31, 32]. We address these challenges by using labeled tool masks to identify unseen pixels. We only optimize in the seen part by introducing the term $(1 - M_i)$, using the L_1 loss as follows:

$$\mathcal{L}_{L1}(I_i, \hat{I}_i) = \sum_{x=1}^{W} \sum_{y=1}^{H} |(1 - M_i(x,y)) \odot \hat{I}_i(x,y) - (1 - M_i(x,y)) \odot I_i(x,y)|. \tag{8}$$

Monocular reconstruction results in limited 3D information. We address this by integrating a depth-guided loss using estimated depth maps D_i with Huber loss $\mathcal{L}_D(i)$ following [42]. We also apply total variation (TV) loss $\mathcal{L}_{TV}$ as [29] to regularize the rendered images. To sum up, our final optimization target is:

$$\begin{aligned} \mathcal{L}(I_i, \hat{I}_i) = & \mathcal{L}_{L1}(I_i, \hat{I}_i) + \lambda_d \mathcal{L}_D(I_i, \hat{I}_i) + \lambda_s \mathcal{L}_S(I_i) + \lambda_{tv} \mathcal{L}_{TV}(I_i, \hat{I}_i) \\ & + \lambda_{shf} \mathcal{L}_{SHF}(I_i, \hat{I}_i) + \lambda_{thf} \mathcal{L}_{THF}(I_i, \hat{I}_i), \end{aligned} \tag{9}$$

where $\mathcal{L}_S$ is the surface-aligned item in EndoGS [42], which is modified from SuGaR [7] to encourage the surface alignment of the Gaussians. Hyperparameters λ_D, λ_{tv}, λ_{shf} and λ_{thf} control the regularization strength. We set λ_d to 0.5, λ_s to 0.2, λ_{tv} to 0.1, λ_{shf} to 1, and λ_{thf} to 10.

5 Experiments

5.1 Datasets

We conduct experiments on two public endoscope datasets, namely ENDONERF [28] and SCARED [1]. The ENDONERF dataset [28] includes **two** cases of in-vivo prostatectomy data providing single-viewpoint estimated depth maps and manually annotated tool masks. The SCARED dataset [1] offers ground truth RGBD images from four porcine cadaver abdominal anatomies, using a DaVinci endoscope and a projector. We preprocess SCARED dataset according to [13]. We evaluate our method by comparing it with recent surgical scene reconstruction methods [13, 28, 39, 42] using image quality metrics such as PSNR, SSIM, and LPIPS as outlined in EndoGS [42].

5.2 Implementation Details

In this work, we implement a two-stage training methodology as [29]. Initially, we focus on training the static field using 3D Gaussian models, followed by training the deformation field. The training involves 3,000 iterations for the static field and extends to 60,000 for the

Method	ENDONERF			SCARED			FPS
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
EndoNeRF [28]	34.20	0.935	0.156	23.52	0.754	0.400	0.2
ForPlane-9k [31]	33.63	0.918	0.100	22.68	0.745	0.431	1.7
ForPlane-32k [31]	36.65	0.947	0.056	23.50	0.762	0.348	1.7
EndoSurf [39]	34.99	0.955	0.113	23.94	0.779	0.384	0.04
EndoGS [42]	36.84	0.963	0.041	26.46	0.770	0.339	∼70
EndoGaussian [13]	37.99	0.966	0.043	26.39	0.792	0.530	∼100
HFGS	38.14	0.971	0.033	27.47	0.796	0.311	∼70

Table 1: Quantitative metrics of appearance (PSNR/SSIM/LPIPS) on ENDONERF [28] and SCARED [1]. The best and the second best results are denoted by pink and yellow.

deformation field. Initial point clouds are estimated using COLMAP [23]. All models are trained on an NVIDIA RTX 3090 GPU.

5.3 Comparison with State-of-the-art Methods

We conduct comparative experiments against various state-of-the-art (SOTA) methods for surgical scene reconstruction, including NerF-based methods such as EndoNeRF [28], ForPlane [31] (an updated version of LerPlane [32]) and EndoSurf [39], and 3D-GS-based methods such as EndoGS [42] and EndoGaussian [13].

Table. 1 presents a quantitative comparison on two public dataset. The FPS in Table. 1 represents the values collected by these methods on the ENDONERF [28] dataset, where all measurements are conducted using a single NVIDIA GeForce RTX 3090 GPU. We observe that while EndoNeRF [28], ForPlane [31] and EndoSurf [39] achieve high performance, however, they require hours of training and testing, making them time-consuming. In contrast, **HFGS** benefits from the rendering efficiency of Gaussian Splatting, enabling it to achieve real-time rendering speeds, and outperforms other SOTA methods in all evaluated metrics on both datasets.

Fig. 3 presents a qualitative comparison between **HFGS** and competitive methods. Notably, the visualizations show that our **HFGS** preserves a significant amount of details with accurate geometry features. Both quantitative and qualitative results strongly support the effectiveness of **HFGS** in achieving high-quality 3D reconstruction at real-time inference speeds. This highlights its potential for future real-time endoscopic applications.

Method	ENDONERF-pulling			ENDONERF-cutting		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Baseline	36.27	0.933	0.057	37.00	0.961	0.036
Ours w/o SHF	38.06	0.967	0.044	37.51	0.969	0.024
Ours w/o THF	37.93	0.965	0.044	37.67	0.968	0.023
Ours	38.44	0.968	0.043	37.83	0.969	0.022

Table 2: Ablation studies on the impact of each module in our method on ENDONERF [28].

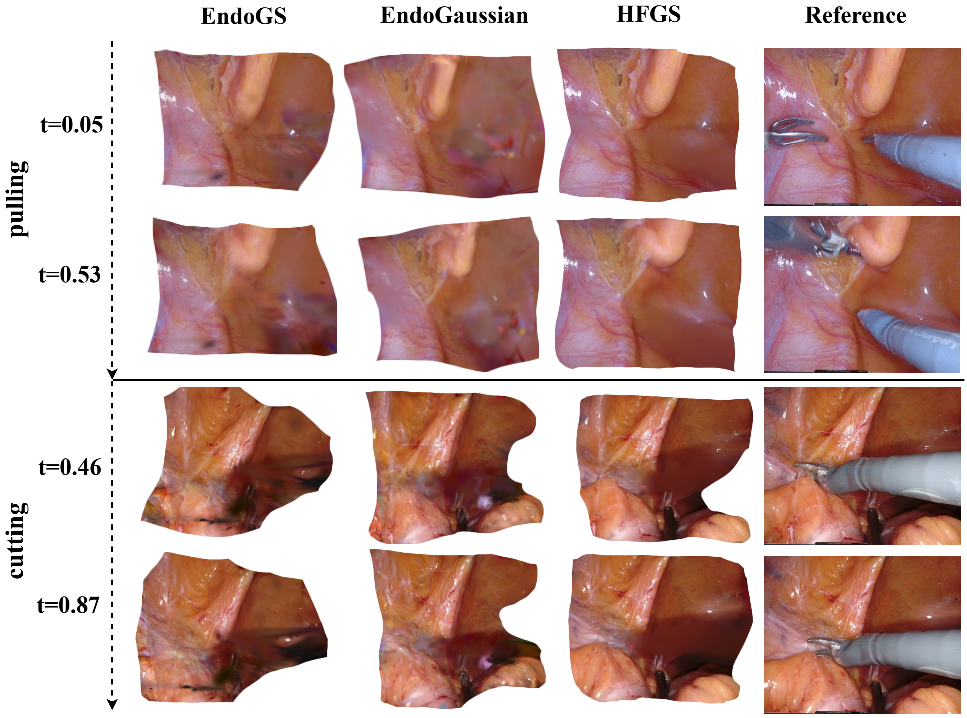


Figure 3: Illustration of reconstruction results of previous works and ours on scene "pulling soft tissues" and "cutting tissues twice" on ENDONERF [28].

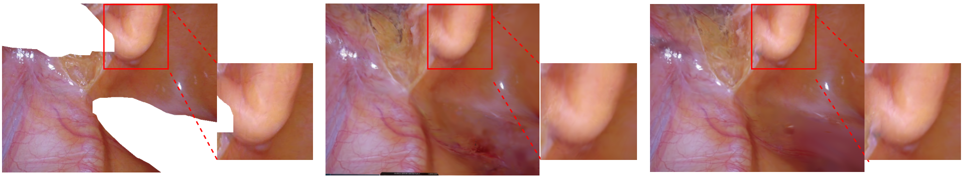
5.4 Ablation Studies

To evaluate the effectiveness of our proposed modules, including SHF and THF, we conduct ablation experiments using the ENDONERF [28] dataset. The corresponding results are shown in Table 2. In Fig. 4, we show the effectiveness of the THF. THF helps the model reconstruct more detailed information and addresses the under-reconstruction in static scenes. Baseline method struggles to effectively render dynamic images. This phenomenon can be mitigated during the optimization with THF as shown in Fig. 1. Results with THF are closer to the ground truth in the predicted optical flow results by [30], indicating more accurate rendering of dynamic scenes.

To sum up, SHF and THF all contribute to performance gains and address the under-reconstruction both in static and dynamic scenes. It is worth mentioning that our **HFGS** has a huge improvement based on the baseline **without any extra inference time**.

6 Conclusion

We introduce a method for deformable endoscopic tissue reconstruction that leverages spatial and temporal frequency analyses to improve under-reconstruction issues, enabling high-quality, real-time reconstruction from single-viewpoint videos. Our method includes two modules that enhance rendering in both static and dynamic scenes. Testing on two public



(a) Reference (b) w/o SHF (c) w/ SHF

Figure 4: Ablation on SHF. We show rendering frames w/ and w/o SHF on scene "pulling soft tissues" on ENDONERF [28].

datasets confirms significant performance gains over existing methods. However, 3D reconstruction from single-viewpoint videos still faces challenges for surgical use. Future work should focus on integrating multiple surgical cameras to enhance 3D tissue reconstruction accuracy and practicality in clinical environments.

References

[1] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

[2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 130–141, 2023.

[3] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proc. of International Conference on Image Processing*, volume 2, pages 168–172, 1994.

[4] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.

[5] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[6] Huxin Gao, Xiaoxiao Yang, Xiao Xiao, Xiaolong Zhu, Tao Zhang, Cheng Hou, Huicong Liu, Max Q-H Meng, Lining Sun, Xiuli Zuo, et al. Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection. *The International Journal of Robotics Research*, 43(3):281–304, 2024.

[7] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023.

[8] Zhiyang Guo, Wengang Zhou, Li Li, Min Wang, and Houqiang Li. Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. *arXiv preprint arXiv:2403.11447*, 2024.

[9] Yiming Huang, Beilei Cui, Long Bai, Ziqi Guo, Mengya Xu, and Hongliang Ren. Endo-4dgs: Distilling depth ranking for endoscopic monocular scene reconstruction with 4d gaussian splatting. *arXiv preprint arXiv:2401.16416*, 2024.

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

[11] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022.

[12] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[13] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

[14] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. 2024.

[15] Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and Jose Maria Martinez Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In *Computer-Assisted and Robotic Endoscopy*, pages 72–83, 2017.

[16] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 32, 2018.

[17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[18] A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[19] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5865–5874, 2021.

[20] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021.

[21] Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.

[22] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[23] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[24] Daniel J Scott, Juan C Cendan, Carla M Pugh, Rebecca M Minter, Gary L Dunnington, and Rosemary A Kozar. The changing face of surgical education: simulation as the new paradigm. *Journal of Surgical Research*, 147(2):189–193, 2008.

[25] Hongqiu Wang, Yueming Jin, and Lei Zhu. Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation. In *2023 IEEE International Conference on Robotics and Automation*, pages 2702–2709, 2023.

[26] Hongqiu Wang, Lei Zhu, Guang Yang, Yike Guo, Shichen Zhang, Bo Xu, and Yueming Jin. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 2024.

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

[28] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, pages 431–441, 2022.

[29] 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. *arXiv preprint arXiv:2310.08528*, 2023.

[30] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2023.

[31] Chen Yang, Kailing Wang, Yuehao Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Efficient deformable tissue reconstruction via orthogonal neural plane. *arXiv preprint arXiv:2312.15253*, 2023.

[32] Chen Yang, Kailing Wang, Yuehao Wang, Xiaokang Yang, and Wei Shen. Neural lerplane representations for fast 4d reconstruction of deformable tissues. In *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, pages 46–56, 2023.

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

[34] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023.

[35] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *Proc. of International Conference on Learning Representations*, 2024.

[36] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023.

[37] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Proc. of Advances in Neural Information Processing Systems*, 33:2492–2502, 2020.

[38] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics*, 38(6):1–14, 2019.

[39] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, pages 13–23, 2023.

[40] Jiahui Zhang, Fangneng Zhan, Muyu Xu, Shijian Lu, and Eric Xing. Fregs: 3d gaussian splatting with progressive frequency regularization. *arXiv preprint arXiv:2403.06908*, 2024.

[41] Haoyu Zhao, Wenhui Dong, Rui Yu, Zhou Zhao, Du Bo, and Yongchao Xu. Morestyle: Relax low-frequency constraint of fourier-based image reconstruction in generalizable medical image segmentation. *arXiv preprint arXiv:2403.11689*, 2024.

[42] Lingting Zhu, Zhao Wang, Zhenchao Jin, Guying Lin, and Lequan Yu. Deformable endoscopic tissues reconstruction with gaussian splatting. *arXiv preprint arXiv:2401.11535*, 2024.

[43] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 371–378, 2001.