Endo-4DGS: Endoscopic Monocular Scene Reconstruction with 4D Gaussian Splatting

Yiming Huang[1,4 ⋆][0000−0002−9754−8608], Beilei Cui[1,4 ⋆][0009−0009−7900−8032], Long Bai[1,4 ⋆][0000−0002−9762−6821], Ziqi Guo[1][0009−0004−5604−1888], Mengya Xu[1][0000−0002−4338−7079], Mobarakol Islam[3][0000−0002−7162−2822], and Hongliang Ren[1,2,4][0000−0002−6488−1551] ⋆⋆

[1] Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China
[2] Shun Hing Institute of Advanced Engineering, CUHK, Hong Kong SAR, China
[3] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, UK
[4] Shenzhen Research Institute, CUHK, Shenzhen, China
{yhuangdl, beileicui, b.long}@link.cuhk.edu.hk, hren@cuhk.edu.hk

Abstract. In the realm of robot-assisted minimally invasive surgery, dynamic scene reconstruction can significantly enhance downstream tasks and improve surgical outcomes. Neural Radiance Fields (NeRF)-based methods have recently risen to prominence for their exceptional ability to reconstruct scenes but are hampered by slow inference speed, prolonged training, and inconsistent depth estimation. Some previous work utilizes ground truth depth for optimization but it is hard to acquire in the surgical domain. To overcome these obstacles, we present Endo-4DGS, a real-time endoscopic dynamic reconstruction approach that utilizes 3D Gaussian Splatting (GS) for 3D representation. Specifically, we propose lightweight MLPs to capture temporal dynamics with Gaussian deformation fields. To obtain a satisfactory Gaussian Initialization, we exploit a powerful depth estimation foundation model, Depth-Anything, to generate pseudo-depth maps as a geometry prior. We additionally propose confidence-guided learning to tackle the ill-pose problems in monocular depth estimation and enhance the depth-guided reconstruction with surface normal constraints and depth regularization. Our approach has been validated on two surgical datasets, where it can effectively render in real-time, compute efficiently, and reconstruct with remarkable accuracy. Our code is available at https://github.com/lastbasket/Endo-4DGS.

Keywords: 3D Reconstruction · Neural Rendering · Robotic Surgery.

1 Introduction

Endoscopic procedures have become a cornerstone in minimally invasive surgery, offering patients with reduced trauma and quicker recovery times [9, 17, 29]. In

⋆ Co-first authors.
⋆⋆ Corresponding author.

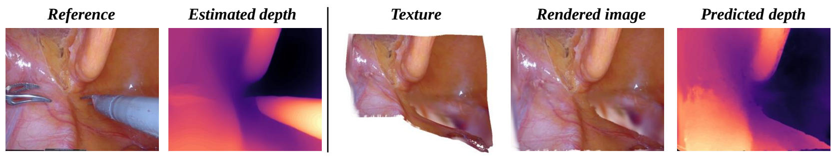

Fig. 1. Ground truth reference, estimated depth from Depth-Anything; 3D textures, rendered image, and predicted depth of our proposed method.

this case, accurate and dynamic 3D reconstruction of the endoscopic scene is critical to enhancing the surgeon's spatial understanding and navigation, facilitating more precise and efficient interventions [14]. However, the complex and constrained nature of endoscopic scenes poses significant challenges for traditional 3D reconstruction techniques due to factors such as limited field-of-view, occlusions, and dynamic tissue deformation [22, 24, 27].

Recent advancements in endoscopic 3D reconstruction have been boosted by the capabilities of Deep Neural Networks (DNNs) [19] and Neural Radiance Fields (NeRFs) [15]. Some studies have achieved strong performance in depth estimation and reconstruction under endoscopy, particularly through stereo reconstruction [1, 13], structure from motion [2], depth and pose estimation [16, 18] or extensive visual pre-training [7]. EndoNeRF [22] is the first to leverage NeRF [15] in endoscopic scenes by dual neural fields approach to model tissue deformation and canonical density. EndoSurf [27] further employs signed distance functions to model tissue surfaces, imposing explicit self-consistency constraints on the neural field. To tackle the lengthy training time requirement, LerPlane [24] constructs a 4D volume by introducing 1D time to the existing 3D spatial space. This extension allows for the formulation of both static fields and dynamic fields by utilizing the spatial-temporal planes, respectively, which leads to a substantial decrease in computational resources. However, reconstructing high-dimensional deformable scenes in real-time remains a challenge.

NeRF-based methods have revolutionized 3D scene reconstruction but face challenges such as slow rendering speeds and suboptimal localization accuracy [4]. Addressing these issues, 3D Gaussian Splatting (GS) has emerged as an effective alternative, offering fast inference and superior 3D representation [11]. By optimizing anisotropic 3D Gaussians using a set of scene images, 3D GS successfully captures the spatial positioning, orientations, color properties, and alpha blending factors, reconstructing both the geometry and visual texture of the scene. Concurrent works [12, 28] also demonstrate fast rendering performance by using 4D Gaussians for scene reconstruction.

To tackle the deformable tissue reconstruction challenges in endoscopic scenes, we further incorporate the temporal dimension as the fourth axis to model dynamic environments [23]. Moreover, current solutions for depth prior-assisted reconstruction depend on multi-view information and the static scene assumption [6, 21], which are not always feasible in the surgical scenario. Meanwhile, the predictions of existing monocular depth estimation methods [25] also suffer from ill-posed problems. The predicted depth results in uncertain measurements even with little changes in the environment, e.g. small deformation on the tissues. Therefore, reconstruction using depth prior supervision remains a challenge in deformable surgery scenarios. To overcome these hurdles, we leverage a foundation model pre-trained on the large-scale dataset, e.g. Depth-Anything [25]. By applying Depth-Anything, we project the pre-trained depth into 3D for more robust 4D Gaussian initialization. To address the challenges posed by inaccurate issues in estimating depth using a monocular camera, we introduce a confidence-guided learning approach that effectively reduces the influence of noisy or uncertain measurements in the pre-trained depth estimation. We additionally implement surface normal constraints and depth regularization to strengthen the pseudo-depth's accuracy and geometry constraint. Fig. 1 showcases our 3D textures, the rendered images, and the depth predictions for endoscopic views. Specifically, our contributions in this paper are threefold:

- We present Endo-4DGS, an innovative technique that adapts Gaussian Splatting for endoscopic scene reconstruction. Utilizing pseudo-depth generated by Depth-Anything, Endo-4DGS achieves remarkable reconstruction outcomes without needing ground truth depth data.
- We propose confidence-guided learning to tackle the ill-pose monocular depth adaption problems, and further employ depth regularization and surface normal constraints against the depth prior adaption challenge in the deformable surgical reconstruction task.
- Our extensive validation on two real surgical datasets shows that Endo-4DGS attains high-quality reconstruction, excels in real-time performance, reduces training expenditures, and demands less GPU memory, which sets the stage for advancements in robot-assisted surgery.

2 Methodology

In this section, we introduce the representation and rendering formula of 4D Gaussians [23] in Sec. 2.1 and demonstrate our motivation and detailed implementation of the depth prior-based reconstruction in Sec. 2.2.

2.1 Preliminaries

3D GS [11] utilizes 3D differentiable Gaussians as the unstructured representation, allowing for a differentiable volumetric representation that can be rapidly rasterized and projected onto a 2D surface for swift rendering. With covariance matrix $\mathbf{\Sigma}$ and mean μ the 3D GS at position x is described as $G(x) = e^{-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)}$, where the covariance $\mathbf{\Sigma}$ can be further decomposed into $\mathbf{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$ with the scaling $\mathbf{S}$ and rotation $\mathbf{R}$. Introducing by [26], with the viewing transform $\mathbf{W}$ and the Jacobian of the affine approximation of the projective transformation $\mathbf{J}$, covariance in the camera plane can be described as

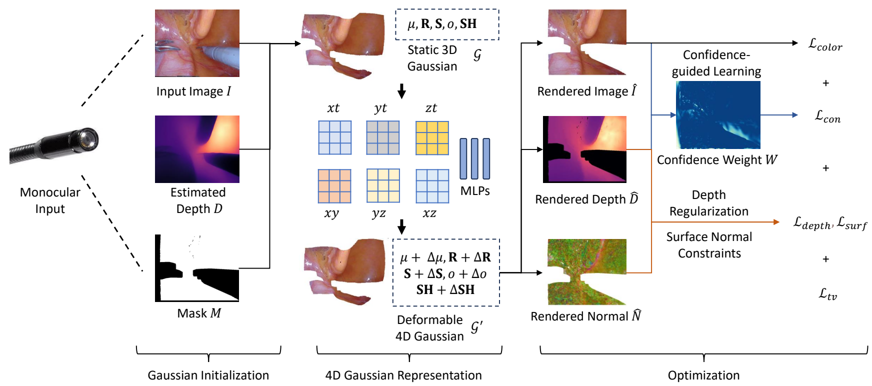

Fig. 2. Illustration of our proposed Endo-4DGS framework. We utilize monocular images, estimated depths from Depth-Anything, and surgical tool masks for training. 3D Guassian is represented as $\mathcal{G}$ with position mean μ, rotation $\mathbf{R}$, scaling $\mathbf{S}$ opacity o, and spherical harmonics $\mathbf{SH}$. 4D Gaussian is described as $\mathcal{G}' = \mathcal{G} + \Delta\mathcal{G}$. $\mathcal{L}_{color}, \mathcal{L}_{con}, \mathcal{L}_{depth}, \mathcal{L}_{surf}, \mathcal{L}_{tv}$ are the color loss, confidence loss, depth regularization loss, surface normal loss and total-variational loss, respectively.

$\mathbf{\Sigma}' = \mathbf{J}\mathbf{W}\mathbf{\Sigma}\mathbf{W}^T\mathbf{J}^T$. The final rendering equation is:

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

where $\hat{C}$ is the predicted pixel color from N points. c_i, α_i are the color defined by the spherical harmonics coefficients and the density calculated by multiplying the 2D covariance $\mathbf{\Sigma}'$ with the learned opacity o_i.

2.2 Proposed Methodology

4D Gaussian Splatting for Deformable Scene Representation. Inspired by [23], we represent the deformable surgical scene with the 4D Gaussian $\mathcal{G}' = \Delta\mathcal{G} + \mathcal{G}$ which includes a static 3D Gaussian $\mathcal{G}$ and its deformation $\Delta\mathcal{G} = \mathcal{F}(\mathcal{G}, t)$, where $\mathcal{F}$ is the deformation network and t is the time. The spatial-temporal encoder $\mathcal{H}$ is defined with multi-resolution Hexplanes [3] $R_l(i,j)$ and a tiny MLP ϕ_d, $\mathcal{H}(\mathcal{G},t) = \{R_l(i,j), \phi_d | (i,j) \in \{(x,y),(x,z),(y,z),(x,t),(y,t),(z,t)\}, l \in \{1,2\}\}$, and the spatial-temporal feature is encoded as $f_d = \mathcal{H}(\mathcal{G},t)$.

A multi-head Gaussian deformation decoder $\mathcal{D} = \{\phi_\mu, \phi_r, \phi_s, \phi_o, \phi_{\mathbf{SH}}\}$ is designed for decoding the deformation of position, rotation, scaling, opacity and spherical harmonics $\mathbf{SH}$ with five tiny MLPs. The final representation of 4D Gaussian can be expressed as:

$$\begin{aligned}\mathcal{G}' &= \{\mu + \phi_\mu(f_d), r + \phi_r(f_d), s + \phi_s(f_d), o + \phi_o(f_d), \mathbf{SH} + \phi_{\mathbf{SH}}(f_d)\} \\ &= \{\mu + \Delta\mu, \mathbf{R} + \Delta\mathbf{R}, \mathbf{S} + \Delta\mathbf{S}, o + \Delta o, \mathbf{SH} + \Delta\mathbf{SH}\}\end{aligned} \tag{2}$$

Gaussians Initialization with Depth Prior. Retrieving accurate point clouds in surgical scenes is challenging since there is only monocular visual information from the consumer-level endoscopes. Therefore, we propose to use the pre-trained depth to implement the point cloud initialization for the 4D Gaussian. With the pre-trained depth estimation model and the input image I, we estimate an inverse depth map D_{inv}. Then a scaling β is applied to recover the depth map $D = \frac{\beta}{D_{inv}}$ in the camera coordinate. Given the camera intrinsic matrix K_1, and the extrinsic matrix K_2, we project the point cloud $P \in \mathbb{R}^{N\times 3}$ with size N from the given image I as follows:

$$P = K_2^{-1} K_1^{-1}[(I \odot M), D], \tag{3}$$

where M is the mask for the input image, $\odot$ is the element-wise multiplication, and $[\cdot]$ indicates concatenation. With the point cloud from depth prior, we initialize $\mu, \mathbf{R}$, making the training process faster for convergence and more robust in terms of geometry.

Confidence Guided Learning. Monocular reconstruction with estimated depth is an ill-pose problem since there is no access to the ground truth geometry information. Inspired by [6, 20], we formulate our solution with a probabilistic model to learn statistics for depth from Depth-Anything, which is defined as:

$$\hat{D} = \frac{\sum_{i\in N} d_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_i)}{\sum_{i\in N} W_i},\ W_i = \alpha_i \prod_{j=1}^{i-1}(1-\alpha_i) \tag{4}$$

where d_i is the depth of the center of the Gaussian obtained by projecting to the z-axis of the camera coordinate. $W_i \in (0, 1)$ is defined as the confidence weight for the corresponding point, which is closer to 1 with higher confidence. Following the above definition, the confidence guidance loss can be expressed as:

$$\mathcal{L}_{con} = \mathbb{E}[\frac{1}{2W^2}||\hat{D}_{norm} - D_{norm}||_2^2 + \log(W)] + \mathbb{E}[\frac{1}{2W^2}||\hat{C} - C||_2^2 + \log(W)], \tag{5}$$

where $\mathbb{E}(\cdot)$ is the expectation, D_{norm} and $\hat{D}_{norm}$ are the depth prior and rendered depth normalized to $(0, 1)$. While we penalize the depth and color with less confidence, we also add the $\log(\cdot)$ as a regularization term. The confidence weight, therefore, maximizes the error where the rendered depth is different from the depth prior while reducing the influence of the uncertain value of the pre-trained depth estimation.

Surface Normal Constraints and Depth Regularization. To utilize the pre-trained depth map more effectively as the pseudo-ground truth, we propose to utilize depth regularization loss and surface normal loss. Following [5], we approximate the surface normal $\hat{n}_i \in \hat{\mathbf{N}}$ with the shortest axis:

$$\hat{\mathbf{n}}_\mathbf{i} = \mathbf{R}_i[r:],\ r = \text{argmin}([s_1, s_2, s_3]), \tag{6}$$

where r is the index of the shortest scaling in $\mathbf{S}_i = diag(s_1, s_2, s_3)$ selected by argmin$(\cdot)$. Then we calculate the gradient of the depth prior $\nabla D = (G^W, G^H)$, and formulate the pseudo surface normal as:

$$\mathbf{n}_i = [\frac{G_i^W}{\sqrt{(G_i^W)^2 + (G_i^H)^2 + 1}}, \frac{G_i^H}{\sqrt{(G_i^W)^2 + (G_i^H)^2 + 1}}, \frac{1}{\sqrt{(G_i^W)^2 + (G_i^H)^2 + 1}}], \tag{7}$$

where G^W, G^H are the gradients along the width and height of the depth map. The surface normal constraints is described as $\mathcal{L}_{surf} = \|\mathbf{N} - \hat{\mathbf{N}}\|_1$. We also regularize the predicted depth from 4D Gaussian with a normalized depth loss and gradient loss. The depth regularization term $\mathcal{L}_{depth}$ is expressed as:

$$\mathcal{L}_{depth} = \lambda_{norm}\|D_{norm} - \hat{D}_{norm}\|_1 + \lambda_{grad}(1 - P_{corr}(\|\nabla D\|_2, \|\nabla \hat{D}\|_2)), \tag{8}$$

where $P_{corr}(\cdot)$ is the Pearson Correlation Coefficient, $\lambda_{norm}, \lambda_{grad}$ are the weights for the normalized depth loss and gradient loss.

With the $\mathcal{L}_{color}$ color loss and a grid-based total-variational loss $\mathcal{L}_{tv}$ [3,8,11], our final loss for optimizing can be represented as:

$$\mathcal{L} = \mathcal{L}_{color} + \mathcal{L}_{tv} + \mathcal{L}_{depth} + \lambda_{surf}\mathcal{L}_{surf} + \lambda_{con}\mathcal{L}_{con}, \tag{9}$$

where $\lambda_{surf}, \lambda_{con}$ are the weights for the surface constraints and confidence loss. Following [23], we emit $\mathcal{L}_{tv}$ for the training of the static 3D Gaussians.

3 Experiments

3.1 Dataset

We evaluate the performance based on two publicly available datasets, StereoMIS [10] and EndoNeRF [22]. The StereoMIS dataset [10] is a stereo video dataset captured by the da Vinci Xi surgical system, consisting of 11 surgical sequences by in-vivo porcine subjects, where we extract the 800 to 1000 frames from the first scene. The EndoNeRF dataset [22] includes two samples of prostatectomy via stereo cameras and provides estimated depth maps based on stereo-matching techniques, they also include challenging scenes with tool occlusion and non-rigid deformation. The training and validation splitting follows the 7:1 strategy in [27]. We use PSNR, SSIM, and LPIPS to evaluate the 3D scene reconstruction performance. We also report the results of training time, inference speed, and GPU memory usage on one single RTX4090 GPU.

3.2 Implementation Details

All experiments are conducted on the RTX4090 GPU with the Python PyTorch framework. We adopt the Adam optimizer with an initial learning rate of 1.6×10^{-3}. We employ the Depth-Anything-Small model for pseudo-depth map generation by considering accuracy, memory usage, and computation efficiency. We adopt the depth scale $\beta = 1000$ and $\lambda_{norm} = 0.01, \lambda_{grad} = 0.001, \lambda_{surf} = 0.001, \lambda_{con} = 0.0001$ for regularization. An encoding voxel size of $[64, 64, 64, 75]$ is applied, where the four dimensions are length, width, height, and time.

3.3 Results

Table 1. Comparison experiments on the EndoNeRF dataset [22] against EndoNeRF [22], EndoSurf [27], and LerPlane [24]. The best results are in bold.

Models	EndoNeRF-Cutting			EndoNeRF-Pulling			Training Time ↓	FPS ↑	GPU Usage ↑
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓			
EndoNeRF [22]	35.84	0.942	0.057	35.43	0.939	0.064	6 hours	0.2	4 GB
EndoSurf [27]	34.89	0.952	0.107	34.91	0.955	0.120	7 hours	0.04	17 GB
LerPlane-32k [24]	34.66	0.923	0.071	31.77	0.910	0.071	8 mins	1.5	20 GB
Endo-4DGS	**36.56**	**0.955**	**0.032**	**37.85**	**0.959**	**0.043**	**4 mins**	**100**	**4GB**

Table 2. Comparison experiments on the StereoMIS [10], against EndoNeRF [22], EndoSurf [27], and LerPlane [24]. The best results are in bold.

Models	PSNR ↑	SSIM ↑	LPIPS ↓	Training Time ↓	FPS ↑	GPU Usage ↓
EndoNeRF [22]	21.49	0.622	0.360	5 hours	0.2	4 GB
EndoSurf [27]	29.87	0.809	0.303	8 hours	0.04	14 GB
LerPlane-32k [24]	30.80	0.826	0.174	7 mins	1.7	19 GB
Endo-4DGS	**32.69**	**0.850**	**0.148**	**7 mins**	**100**	**4 GB**

We conducted a comprehensive comparison of our proposed method with state-of-the-art approaches for surgical scene reconstruction. Specifically, we reproduce EndoNeRF [22], EndoSurf [27], and LerPlane [24] with the original implementation. The evaluation results on the EndoNeRF [22] and StereoMIS [10] datasets are presented in Table 1 and Table 2. Upon analysis, we observed that while EndoNeRF [22] and EndoSurf [27] achieved relatively high performance, they required hours of training, making them time-consuming. On the other hand, LerPlane [24] significantly reduced the training time to approximately 8 minutes but incurred a slight degradation in rendering performance. It is important to note that all of these state-of-the-art methods suffered from very low frames per second (FPS), which limited their practical application in real-time surgical scene reconstruction tasks. In contrast, our proposed method not only outperformed all evaluated metrics on both datasets but also achieved a real-time inference speed of 100 FPS, where the training was accomplished with only 4 minutes and 4GB of GPU memory. The significant improvement in inference speed makes our method highly suitable for real-time endoscopic applications.

We have provided qualitative results for EndoNeRF datasets [22] in Fig. 3. Notably, the visualizations demonstrate that our proposed method preserved a substantial amount of visible details with accurate geometry features. The aforementioned quantitative and qualitative results strongly support the effectiveness of our method in achieving high-quality 3D reconstruction scenes at real-time

Table 3. Ablation experiments of the proposed method on EndoNeRF dataset [22]. To observe the performance changes, we remove (i) the depth regularization, (ii) the surface constraints, and (iii) the confidence guidance. The best results are in bold.

Depth Regularization	Surface Constraints	Confidence Guidance	EndoNeRF-Cutting			EndoNeRF-Pulling		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
✗	✗	✗	35.14	0.938	0.046	35.39	0.937	0.082
✓	✗	✗	36.00	0.949	0.040	35.68	0.942	0.072
✗	✓	✗	35.22	0.940	0.057	35.97	0.945	0.066
✗	✗	✓	35.54	0.941	0.048	35.68	0.942	0.066
✓	✓	✗	36.24	0.951	0.038	36.35	0.945	0.062
✓	✗	✓	36.22	0.951	0.036	36.94	0.952	0.053
✗	✓	✓	36.08	0.946	0.036	36.15	0.943	0.064
✓	✓	✓	**36.56**	**0.955**	**0.032**	**37.85**	**0.959**	**0.043**

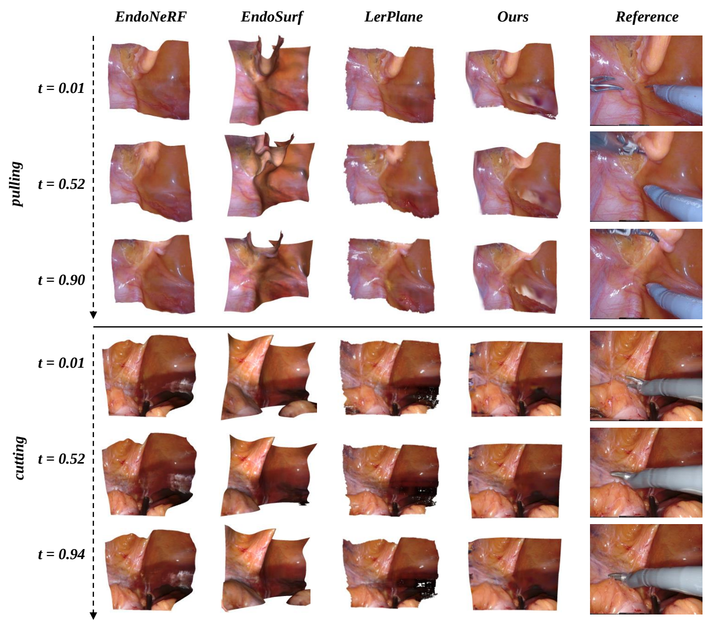

Fig. 3. Qualitative comparison on the EndoNeRF dataset [22] against EndoNeRF [22], EndoSurf [27], and LerPlane [24].

inference speeds. This highlights its potential for future real-time endoscopic applications. We provide more visualizations on StereoMIS in the supplementary.

To further analyze the contributions of our designs, we conducted an ablation study on the EndoNeRF dataset [22] by removing (i) depth regularization, (ii) surface normal constraints, (iii) confidence-guided learning. The experimental results in Table 3 unequivocally demonstrate that the absence of any of the components leads to a substantial degradation in performance. These results highlight the crucial role played by each component in enhancing the quality, accuracy, and overall performance of our method.

4 Conclusion

In this paper, we propose Endo-4DGS, a real-time, high-fidelity reconstruction method of deformable tissues. Different from previous works, lightweight MLPs are implemented to capture temporal dynamics with Gaussian deformation fields. We further propose to estimate the depth map by a foundation model Depth-Anything for Gaussian Initialization. The framework is additionally enhanced with confidence-guided strategy, surface normal constraints, and depth regularization to better utilize the depth prior constraint. Extensive experiments demonstrate the superior performance and fast inference speed of our proposed method against other state-of-the-art methods. These results underline the vast potential of Endo-4DGS to improve a variety of surgical applications, allowing for better decision-making and safety during operations.

Acknowledgements. This work was supported by Hong Kong RGC CRF C4026-21G, RIF R4020-22, GRF 14211420, 14216020 & 14203323); Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGDX20210823103535014 (202108233000303).

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

References

1. Bae, G., Budvytis, I., Yeung, C.K., Cipolla, R.: Deep multi-view stereo for dense 3d reconstruction from monocular endoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 774–783. Springer (2020)
2. Barbed, O.L., Montiel, J.M., Fua, P., Murillo, A.C.: Tracking adaptation to improve superpoint for 3d reconstruction in endoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 583–593. Springer (2023)
3. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
4. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)
5. Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., Chen, X.: Gaussianpro: 3d gaussian splatting with progressive propagation. arXiv preprint arXiv: (2024)
6. Chung, J., Oh, J., Lee, K.M.: Depth-regularized optimization for 3d gaussian splatting in few-shot images. arXiv preprint arXiv:2311.13398 (2023)
7. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-dino: Adapter learning of foundation model for depth estimation in endoscopic surgery. arXiv preprint arXiv:2401.06013 (2024)
8. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)

9. Gao, H., Yang, X., Xiao, X., Zhu, X., Zhang, T., Hou, C., Liu, H., Meng, M.Q.H., Sun, L., Zuo, X., et al.: Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection. The International Journal of Robotics Research p. 02783649231209338 (2023)
10. Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., Sznitman, R.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery pp. 1185—-1192 (2023)
11. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
12. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2401.12561 (2024)
13. Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q.: E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 415–425. Springer (2021)
14. Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., Montiel, J.M.M.: Orbslam-based endoscope tracking and 3d reconstruction. In: Computer-Assisted and Robotic Endoscopy: Third International Workshop, CARE 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 3. pp. 72–83. Springer (2017)
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
16. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical image analysis **71**, 102058 (2021)
17. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
18. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical image analysis **77**, 102338 (2022)
19. Stucker, C., Schindler, K.: Resdepth: Learned residual stereo reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 184–185 (2020)
20. Wang, F., Chen, Z., Wang, G., Song, Y., Liu, H.: Masked space-time hash encoding for efficient dynamic scene reconstruction. Advances in Neural Information Processing Systems **36** (2024)
21. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
22. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022)

23. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Xinggang, W.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
24. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. arXiv preprint arXiv:2305.19906 (2023)
25. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv:2401.10891 (2024)
26. Yifan, W., Serena, F., Wu, S., Öztireli, C., Sorkine-Hornung, O.: Differentiable surface splatting for point-based geometry processing. ACM Transactions on Graphics (TOG) **38**(6), 1–14 (2019)
27. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2023)
28. Zhu, L., Wang, Z., Jin, Z., Lin, G., Yu, L.: Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv preprint arXiv:2401.11535 (2024)
29. Zia, A., Bhattacharyya, K., Liu, X., Berniker, M., Wang, Z., Nespolo, R., Kondo, S., Kasai, S., Hirasawa, K., Liu, B., et al.: Surgical tool classification and localization: results and methods from the miccai 2022 surgtoolloc challenge. arXiv preprint arXiv:2305.07152 (2023)