EndoSparse: Real-Time Sparse View Synthesis of Endoscopic Scenes using Gaussian Splatting

Chenxin Li[1], Brandon Y. Feng[2⋆], Yifan Liu[1], Hengyu Liu[1], Cheng Wang[1], Weihao Yu[1], Yixuan Yuan[1⋆]

[1] The Chinese University of Hong Kong, [2] Massachusetts Institute of Technology

Abstract. 3D reconstruction of biological tissues from a collection of endoscopic images is a key to unlock various important downstream surgical applications with 3D capabilities. Existing methods employ various advanced neural rendering techniques for photorealistic view synthesis, but they often struggle to recover accurate 3D representations when only sparse observations are available, which is usually the case in real-world clinical scenarios. To tackle this sparsity challenge, we propose a framework leveraging the prior knowledge from multiple foundation models during the reconstruction process, dubbed as *EndoSparse*. Experimental results indicate that our proposed strategy significantly improves the geometric and appearance quality under challenging sparse-view conditions, including using only three views. In rigorous benchmarking experiments against state-of-the-art methods, *EndoSparse* achieves superior results in terms of accurate geometry, realistic appearance, and rendering efficiency, confirming the robustness to sparse-view limitations in endoscopic reconstruction. *EndoSparse* signifies a steady step towards the practical deployment of neural 3D reconstruction in real-world clinical scenarios. Project page: `https://endo-sparse.github.io/`.

Keywords: Sparse View Synthesis · Gaussian Splatting · Endoscopy.

1 Introduction

Reconstructing 3D surgical scenes from endoscope videos [14] can create immersive virtual surgical environments, benefiting robot-assisted surgery and augmented/virtual reality surgical training [37] for medical professionals [30,37]. The ongoing development of real-time photorealistic reconstruction broadens the scope of applications to include intraoperative usage, enabling surgeons to navigate and precisely control surgical instruments while maintaining a comprehensive view of the surgical scene [22,12]. This advancement could further minimize the need for invasive follow-up procedures.

Previous investigations into the 3D reconstruction of surgical scenes have focused on depth estimation methodologies [4], the integration of point clouds in a SLAM pipeline [28], and the design of spatial warping fields [23]. Recently,

[⋆] Equal advising

advances in neural rendering, spearheaded by Neural Radiance Fields (NeRFs) [24,2,9], kick-started the trend of representing the surgical 3D scene as a radiance field [30,37,34]. Seminal papers, including EndoNeRF [30] and its follow-up works [30,37,34], encapsulate deformable 3D scenes as a canonical neural radiance field with a temporally varying deformable field. Although they achieve convincing reconstruction of pliable tissues, these methods incur a heavy rendering cost since the NeRF approach requires querying such neural radiance fields multiple times for a single pixel, limiting the applicable usage in intraoperative applications [30,34].

As a promising alternative, the recently introduced 3D Gaussian Splatting (3D-GS) [8] exhibits pleasing properties to overcome the inefficiency of NeRF-based methods without sacrificing visual quality. Through using a collection of 3D Gaussians as explicit representations with attributes of geometric shape and color appearance and an efficient splatting-based rasterization, 3D-GS can achieve real-time image rendering and such success has enabled endoscopic 3D reconstruction in real-time from a dense collection of camera viewpoints by a holistic framework using 3D-GS and a deformable modeling [22,10,21,40].

However, despite the enormous progress in applying state-of-the-art neural 3D reconstruction pipelines to endoscopic surgical scenes, a common assumption for these methods is the access to a dense collection of training views. However, this assumption is often unrealistic in clinical settings, as real-world captures are often accompanied by equipment instability and variable noise and lighting conditions, necessitating eliminating a significant number of low-quality views [41,22,13]. As a result, the geometric and visual quality of existing neural rendering methods like 3D-GS would both significantly degrade with the decreased available views [41]. To alleviate such performance deterioration in clinical practice, this paper presents the **first investigation** into the medical scene reconstruction under **sparse-view** settings.

Our insight is inspired by the impressive results delivered from Visual Foundation Models (VFMs) [27,35] that using the prior knowledge extracted on large-scale pre-training to facilitate learning for downstream tasks [38,36,19]. While relevant efforts of using foundation models have been revealed effective for 2D medical image segmentation [18,17,39,15], 3D volume segmentation [7,29,16,6], and depth estimation [3,25], VFMs have yet to empower more computational extensive medical tasks like 3D medical scene reconstruction. In this paper, we introduce *EndoSparse*, a framework enabling efficient reconstruction and rendering of endoscopic scenes from sparse observations. *EndoSparse* enhances 3D-GS scene reconstruction by distilling [11] geometric and appearance priors from pre-trained foundation models. Specifically, the optimization of 3D-GS is designed to obey the data distribution with large-scale pre-trained generative models. Given the images produced by to-be-optimized 3D representations, we enforce the rendered RGB images to maximize the score distilled from an image diffusion model (Stable Diffusion [27]), and that the rendered depth maps to be consistent with the prediction obtained via Depth-Anything [35]. Our framework significantly improves the geometric and visual accuracy of the reconstructed 3D scene

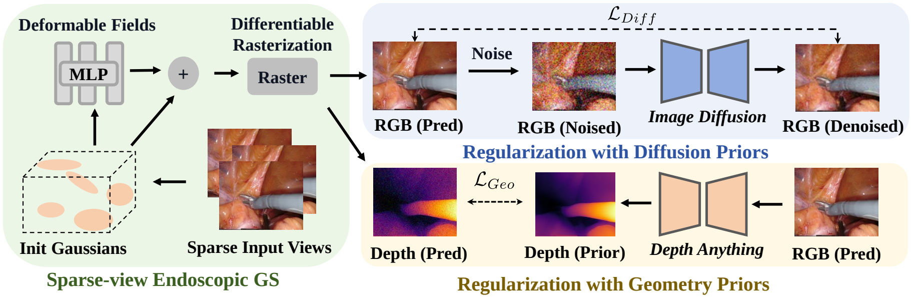


Fig. 1. ***EndoSparse* Overview.** Within a 3D-GS scene reconstruction framework, we incorporate vision foundation models as effective regularizers of the 3D scene. We incorporate geometric prior knowledge from Depth-Anything [35] and image appearance priors from Stable Diffusion [27], which provide valuable guidance signals for optimization at viewpoints without camera coverage.

despite dealing with the challenging condition of only having access to spare observations (ranging from 3 to 12 views).

In short, our contributions are outlined as: (**i**) We present state-of-the-art results on surgical scene reconstruction from a sparse set of endoscopic views, achieving and significantly enhancing the practical usage potential of neural reconstruction methods. (**ii**) We demonstrate an effective strategy to instill prior knowledge from a pre-trained 2D generative model to improve and regularize the visual reconstruction quality under sparse observations. (**iii**) We introduce an effective strategy to distill geometric prior knowledge from a visual foundation model that drastically improves the geometric reconstruction quality under sparse observations.

2 Method

As shown in Fig. 1, *EndoSparse* aims to perform accurate and efficient endoscopic scene reconstruction with a collection of sparse observations. The Gaussians are each initialized with attributes related with color, position, and shape (Sec. 2.1). To bolster the appearance quality for the representation constructed on insufficient perspectives, a diffusion prior is leveraged to effectively regularize the synthesized results to be plausible (Sec. 2.2). To further facilitate accurate geometry, we exploit priors distilled from a foundation model with depth estimation abilities (Sec. 2.3). Overall, the proposed *EndoSparse* is robust against degraded reconstruction quality due to only having sparse observations (Sec. 2.4).

2.1 Deformable Endoscopic Reconstruction with 3D-GS

3D Gaussian Splatting. 3D Gaussian Splatting (3D-GS) [8] provides an explicit representation of a 3D scene, utilizing an array of 3D Gaussians, each endowed with specific attributes: a positional vector $\boldsymbol{\mu} \in \mathbb{R}^3$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, which can be further deconstructed into a scaling factor $\boldsymbol{s} \in \mathbb{R}^3$ and a rotation quaternion $\boldsymbol{r} \in \mathbb{R}^4$, both of which cater to the requirements of differentiable optimization. Additionally, the opacity logit $\boldsymbol{o} \in \mathbb{R}$ and Spherical Harmonic (SH) coefficients $\boldsymbol{c} \in \mathbb{R}^k$ (where k represents numbers of SH functions) can be utilized to represent colors and view-dependent appearances respectively

$$G(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{1}$$

Accordingly, 3D-GS arranges all the Gaussians contributing to a pixel in a specific order, and subsequently blends the ordered Gaussians overlapping the pixels utilizing: $\hat{C} = \sum_{i=1}^{n} \ c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$, where c_i, α_i denotes the color and density computed by a Gaussian G with covariance Σ, which is then multiplied by an optimizable SH color coefficients and opacity that are unique to each point.

Deformable Scene Reconstruction. Building upon 3D-GS, [31] introduces a deformation module that incorporates a 4D encoding voxel F_ν and a compact MLP F_θ to learn a deformation field of Gaussians, thereby facilitating the effective modeling the dynamical components in a scene. Specially, given a 4D input including the Gaussian center $\boldsymbol{\mu}$ and query time τ, the 4D encoding voxel F_ν retrieves the latent feature of inputs, $F_\nu(\boldsymbol{\mu}, \tau)$. Then, the MLP F_θ computes the time-varying deformation in position, rotation, and scaling as $\{\Delta\boldsymbol{\mu}, \Delta\boldsymbol{r}, \Delta\boldsymbol{s}\} = F_\theta \circ F_\nu(\boldsymbol{\mu}, \tau)$. Consequently, the representation of Gaussians could be depicted as a dynamic fashion: $\{\boldsymbol{\mu}+\Delta\boldsymbol{\mu}, \boldsymbol{r}+\Delta\boldsymbol{r}, \boldsymbol{s}+\Delta\boldsymbol{s}, \boldsymbol{o}, \boldsymbol{c}\}$.

Despite the fact that the introduced pipeline can achieve satisfactory rendering quality when there is an abundance of training views, its performance significantly *declines* as *the number of available viewpoints decreases.* In what follows, we propose the strategies to leverage the priors from foundational models to recuperate the compromised performance under the sparsity challenge.

2.2 Instilling Diffusion Prior for Plausible Appearance

In essence, during training, we introduce random noise to the rendered image from the novel viewpoints, and let the diffusion model predict the original image without noise, and we use that predicted clean image as the pseudo-ground-truth view to derive loss for our current scene representation. As the diffusion model is trained on a huge amount of visual content, it inherently possesses a general image prior and is capable of providing plausible guidance gradients even for regions with missing details [33].

Specifically, random noise is gradually added at levels $t \in \{1, \ldots, T\}$ to the rendered images $\hat{\mathbf{C}}$ to obtain noisy samples $\tilde{\mathbf{C}}_t$ as

$$\tilde{\mathbf{C}}_t = \sqrt{\bar{\alpha}_t}\hat{\mathbf{C}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \bar{\alpha}_t := \prod_{s=1}^{t} 1-\beta_s$, and $\{\beta_1, \ldots, \beta_T\}$ is the variance schedule of a process with T steps. In the reverse denoising diffusion process, the conditional denoising model $\epsilon_\phi(\cdot)$ parameterized with learned parameters ϕ gradually removes noise from $\tilde{\mathbf{C}}_t$ to obtain $\tilde{\mathbf{C}}_{t-1}$. The guidance signals can be obtained by noising $\hat{\mathbf{C}}$ with sampled noise ϵ at a random timestep t, computing the noise estimate $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\phi \left(\hat{\mathbf{C}}_t, \tilde{\mathbf{C}}, t\right)$ and minimizing the following quantity of Score Distillation Sampling (SDS) as introduced in [26]:

$$SDS(\hat{\boldsymbol{C}}, \tilde{\boldsymbol{C}}) = \mathbb{E}_{\hat{\mathbf{C}}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,I), t \sim \mathcal{U}(T)} \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi \left(\hat{\mathbf{C}}_t, \tilde{\mathbf{C}}, t\right) \|_2^2. \tag{3}$$

2.3 Distilling Geometric Prior for Accurate Geometry

Under conditions of sparse training views, the scarcity of observational data inhibits the ability to coherently learn geometry, subsequently heightening the propensity for overfitting on training views and yielding less than desirable extrapolation to novel views.

Geometry Coherence in Monocular Depth. Using the foundational depth estimation model, DepthAnything [35], which is trained using a substantial dataset comprising 1.5 million paired image-depth observations and 62 million unlabeled images, we can generate monocular depth maps for all rendered images. To reconcile the scale ambiguity inherent between the actual scene scale and the estimated monocular depth, we employ a relaxed relative loss, i.e., Pearson correlation, to measure the distributional similarity between the rendered depth maps $\hat{\boldsymbol{D}}$ and the estimated ones $\tilde{\boldsymbol{D}}$:

$$\mathrm{Corr}(\hat{\boldsymbol{D}}, \tilde{\boldsymbol{D}}) = \frac{\mathrm{Cov}(\hat{\boldsymbol{D}}, \tilde{\boldsymbol{D}})}{\sqrt{\mathrm{Var}(\hat{\boldsymbol{D}})\,\mathrm{Var}(\tilde{\boldsymbol{D}})}} \tag{4}$$

This soft constraint allows for the alignment [41,20] of depth structure without being hindered by the inconsistencies in absolute depth values.

Differentiable Depth Rasterization. To facilitate the backpropagation from the depth prior to guide the training of the Gaussian, we employ a differentiable depth rasterizer. This allows for the comparison and evaluation of the discrepancy between the rendered depth $\hat{\boldsymbol{D}}$ and the estimated depth $\tilde{\boldsymbol{D}}$ by DepthAnything. Specifically, we leverage the alpha-blending rendering technique used in 3D-GS for depth rasterization, where the z-buffer from the sequentially arranged Gaussians contributing to a pixel is accumulated to generate the depth value:

$$\hat{\boldsymbol{D}} = \sum_{i=1}^{n} d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{5}$$

where d_i signifies the z-buffer corresponding to the i-th Gaussians. The incorporation of a fully differentiable implementation facilitates the depth correlation loss, thereby enhancing the congruence between rendered and estimated depths.

2.4 Overall Optimization

The overall training objective can be derived by combining all the above terms. Meanwhile, reconstructing from videos wherein tool occlusion is present poses a significant challenge. In line with previous studies [30,34], we utilize labeled tool occlusion masks M (where 1 denotes tool pixels and 0 for tissue pixels) to denote the unseen pixels in the final training loss function:

$$\mathcal{L} = \lambda_1 \underbrace{\|\overline{M} \odot (\hat{\mathbf{C}} - \mathbf{C})\|_1}_{\mathcal{L}_{\text{RGB}}} + \lambda_2 \underbrace{\|\overline{M} \odot \text{SDS}(\hat{\mathbf{C}}, \tilde{\boldsymbol{C}})\|_1}_{\mathcal{L}_{\text{Diff}}} + \lambda_3 \underbrace{\|\overline{M} \odot (1 - \text{Corr}(\hat{\mathbf{D}}, \tilde{\mathbf{D}}))\|_1}_{\mathcal{L}_{\text{Geo}}} \tag{6}$$

where $\overline{M} = 1 - M$ is applied since the loss functions is merely calculated on the tissue pixels. Note that the geometrical loss $\mathcal{L}_{Geo}$ is implemented by applying 1 - Pearson similarity ($Corr$ in Eq. 4). λ_1, λ_2, λ_3 is the trade-off coefficients. Finally, the parameters of Gaussians $P_G = \{\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{o}, \boldsymbol{c}\}$, MLP θ and encoding fields ν is updated jointly with the gradients $\nabla_{P_G,\theta,\nu} \mathcal{L}$, with regard to the total objective in Eq. 6.

3 Experiments

3.1 Experiment Settings

Datasets and Evaluation. Empirical evaluations are conducted on two public repositories, specifically, EndoNeRF-D [30] and SCARED [1]. EndoNeRF-D [30] incorporates two instances of in-vivo prostatectomy data, collected from stereo cameras positioned at a singular vantage point. This dataset encapsulates intricate scenarios hallmarked by non-rigid deformation and instrument occlusion. The SCARED compilation [1] comprises RGBD visuals of five porcine cadaver abdominal anatomical structures, procured using a DaVinci endoscope and a projector. The efficacy of our methodology is assessed utilizing inference speed, quantified as frames per second (FPS), geometrical quality in terms of total variations (TV) and SSIM of depth maps, and the standard visual quality metrics for rendered images as PSNR, SSIM, and LPIPS.

Implementation Details. Following [31,22], we adopt a two-stage training methodology in to model the static and deformation fields. In the first stage, we train the 3D-GS model only for static modeling while in the second stage, we train the 3D-GS with the deformable field jointly. We set the coefficients of photometric loss term, diffusion prior and geometry prior, i.e., λ_1, λ_2, λ_3, as 1, 0.001, 0.01 respectively by grid search. we utilize an Adam optimizer with an inaugural learning rate of 1.6×10^{-3}. Following [31,22,32], we adopt a warm-up strategy, which initially optimize Canonical Gaussians without involving deformation fields for 1k iterations, and then train the whole framework for an additional 3k iterations. All experiments are executed on a RTX 4090 GPU.

3.2 Comparison with State-of-the-arts

EndoSparse is evaluated in comparison to the existing state-of-the-art reconstruction methods, namely, EndoNeRF [30], EndoSurf [37], LerPlane [34], EndoGS [5] and EndoGaussian [22]. As shown in Tab. 1, *EndoSparse* excels over the

Table 1. Quantitative comparisons on two datasets, with three training views.

Dataset	Method	Efficiency	Geometrical Quality			Visual Quality		
		FPS ↑	TV ↓	δ_1 ↑	SSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓
EndoNeRF-D	EndoNeRF [30]	0.06	96.06	0.920	0.851	25.01	0.762	0.246
	EndoSurf [37]	0.05	91.56	0.931	0.858	25.34	0.783	0.225
	LerPlane-9K [34]	0.96	93.20	0.913	0.849	23.93	0.755	0.249
	LerPlane-32K [34]	0.91	83.63	0.957	0.856	25.83	0.789	0.201
	EndoGS [5]	112.5	90.64	0.942	0.863	24.83	0.774	0.241
	EndoGaussian [22]	186.4	92.58	0.938	0.859	25.37	0.792	0.222
	EndoSparse (Ours)	195.2	74.61	0.976	0.899	26.55	0.826	0.193
SCARED	EndoNeRF [30]	0.03	130.2	0.748	0.256	18.73	0.675	0.356
	EndoSurf [37]	0.02	121.1	0.782	0.290	19.64	0.693	0.318
	EndoGaussian [22]	179.5	116.4	0.765	0.273	19.40	0.681	0.331
	EndoSparse (Ours)	183.1	105.7	0.806	0.309	20.95	0.718	0.294

state-of-the-art methods based on the NeRF representation [30,37,34] for endoscopic scene reconstruction in terms of rendering efficiency, geometric precision and visual quality. Furthermore, *EndoSparse* surpasses EndoGS and EndoGaussian in all aspects, indicating that our method effectively recovers a accurate representation of scenes from sparse views thanks to our designed strategy to incorporate priors from vision foundation models. Fig. 2 further showcases the qualitative results of our method and prior state-of-the-arts. Compared with other techniques, the rendered images (in 1st Row) by our proposed *EndoSparse* preserves greater details and proffers superior visual renditions of the deformable tissues. Besides, we provide the visualization of rendered depth maps (normalized and applied colormap) and we can see that our method demonstrates better geometrical precision compared to the reference ones.

3.3 Ablation Studies

Efficacy of Key Components. Figure 3**(a)** presents a detailed depiction of the ablations, focusing on the key components of the proposed *EndoSparse* model, specifically the diffusion prior and the geometry prior. It is noteworthy that when these two aforementioned priors are applied in their respective capacities, the performance of the model experiences a noticeable uptick. This enhancement not only validates the overall effectiveness of our designs but also underscores the capability and promising potential of vision foundational models. Our ablation study results confirm their pivotal roles in the overall performance and effectiveness of the model under the challenging condition of sparse observations.
Ablations on Quantity of Training Views. As shown in Figure 3**(b)**, we perform a series of ablations studies on the number of training views, ranging from a minimum of 3 views to a maximum of 12 views. As expected, we can see a consistent trend showing that an increase in the number of views correlates with an improvement in the visual and geometrical quality of the output. In

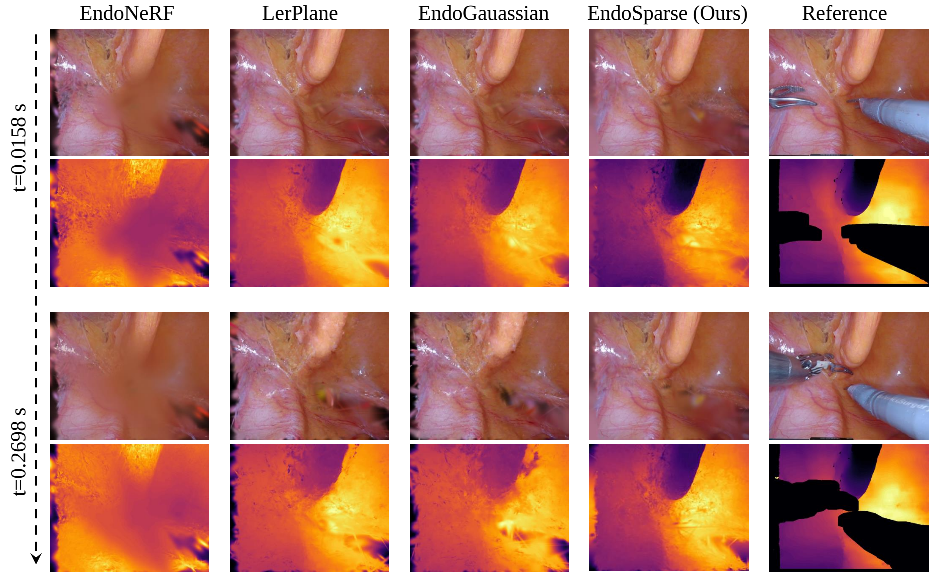
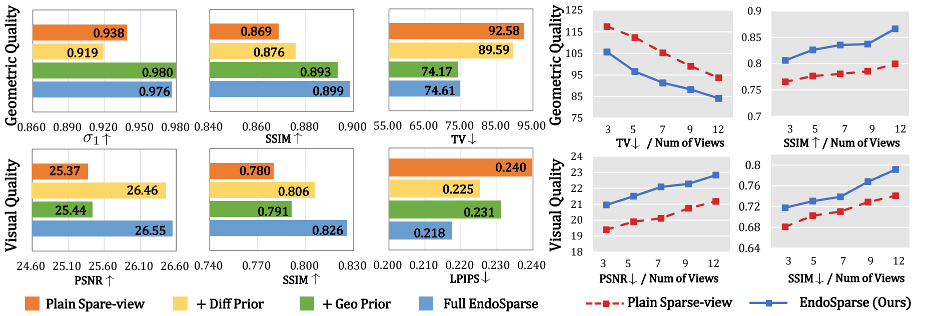


Fig. 2. Qualitative results of rendered images and depth maps on EndoNeRF-D.



(a) Ablations on Efficacy of Key Components in EndoNeRF-D (b) Ablations on Number of Views in SCARED

Fig. 3. Ablation analysis on EndoNeRF-D and SCARED datasets, with the results in terms of geometrical quality (top) and visual quality (bottom).

addition, it is important to note that our proposed *EndoSparse* consistently outperforms the baseline across all settings. This suggests that *EndoSparse* is not only capable of generalizing to a larger number of views, but also able to deliver superior performance in terms of quality and accuracy.

4 Conclusion

This paper introduces an efficient and robust framework 3D reconstruction of endoscopic scenes, achieving real-time and photorealistic reconstruction using sparse observations. Specifically, we utilize vision foundation models as effective regularizers for the optimization of 3D representation. We incorporate geometric prior knowledge from Depth-Anything [35] and image appearance priors from Stable Diffusion [27]. Collectively, *EndoSparse* delivers superior results in terms of accuracy, rendering efficiency, and sparse-view robustness in the reconstruction of endoscopic scenes. With *EndoSparse*, we make steady strides towards the real-world deployment of neural 3D reconstruction in practical clinical scenarios.

Acknowledgments. This work was supported by Hong Kong Innovation and Technology Commission Innovation and Technology Fund ITS/229/22 and Research Grants Council (RGC) General Research Fund 14204321, 11211221.

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

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. ICCV (2021)
3. Beilei, C., Mobarakol, I., Long, B., Hongliang, R.: Surgical-dino: Adapter learning of foundation model for depth estimation in endoscopic surgery. arXiv (2024)
4. Brandao, P., Psychogyios, D., Mazomenos, E., Stoyanov, D., Janatka, M.: Hapnet: hierarchically aggregated pyramid network for real-time stereo matching. CMBBE: Imaging & Visualization **9**(3), 219–224 (2021)
5. Chen, Y., Wang, H.: Endogaussians: Single view dynamic gaussian splatting for deformable endoscopic tissues reconstruction. arXiv:2401.13352 (2024)
6. Ding, Z., Dong, Q., Xu, H., Li, C., Ding, X., Huang, Y.: Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In: ICONIP. pp. 133–144. Springer (2022)
7. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv:2306.13465 (2023)
8. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
9. Li, C., Feng, B.Y., Fan, Z., Pan, P., Wang, Z.: Steganerf: Embedding invisible information within neural radiance fields. In: ICCV. pp. 441–453 (2023)
10. Li, C., Feng, B.Y., Liu, Y., Liu, H., Wang, C., Yu, W., Yuan, Y.: Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting. arXiv:2407.01029 (2024)

11. Li, C., Lin, M., Ding, Z., Lin, N., Zhuang, Y., Huang, Y., Ding, X., Cao, L.: Knowledge condensation distillation. In: ECCV. pp. 19–35. Springer (2022)
12. Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., Yu, Y.: Domain generalization on medical imaging classification using episodic training with task augmentation. Computers in biology and medicine **141**, 105144 (2022)
13. Li, C., Liu, H., Fan, Z., Li, W., Liu, Y., Pan, P., Yuan, Y.: Gaussianstego: A generalizable stenography pipeline for generative 3d gaussians splatting. arXiv:2407.01301 (2024)
14. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. arXiv:2403.11050 (2024)
15. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. arXiv:2406.02918 (2024)
16. Li, C., Ma, W., Sun, L., Ding, X., Huang, Y., Wang, G., Yu, Y.: Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. Neural Computing and Applications pp. 1–14 (2022)
17. Li, C., Zhang, Y., Li, J., Huang, Y., Ding, X.: Unsupervised anomaly segmentation using image-semantic cycle translation. arXiv:2103.09094 (2021)
18. Li, C., Zhang, Y., Liang, Z., Ma, W., Huang, Y., Ding, X.: Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 61–65. IEEE (2021)
19. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR. pp. 5291–5300 (2022)
20. Liang, Z., Rong, Y., Li, C., Zhang, Y., Huang, Y., Xu, T., Ding, X., Huang, J.: Unsupervised large-scale social network alignment via cross network embedding. In: CIKM. pp. 1008–1017 (2021)
21. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. arXiv:2406.16073 (2024)
22. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv:2401.12561 (2024)
23. 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: MICCAI. pp. 415–425. Springer (2021)
24. 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)
25. Pan, P., Fan, Z., Feng, B.Y., Wang, P., Li, C., Wang, Z.: Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images. arXiv:2306.07598 (2023)
26. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv:2209.14988 (2022)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
28. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robotics and Automation Letters **3**(1), 155–162 (2017)
29. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. Computers in biology and medicine **140**, 105067 (2022)

30. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: MICCAI. pp. 431–441 (2022)
31. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv (2023)
32. Xu, H., Li, C., Zhang, L., Ding, Z., Lu, T., Hu, H.: Immunotherapy efficacy prediction through a feature re-calibrated 2.5 d neural network. Computer Methods and Programs in Biomedicine **249**, 108135 (2024)
33. Xu, H., Zhang, Y., Sun, L., Li, C., Huang, Y., Ding, X.: Afsc: Adaptive fourier space compression for anomaly detection. arXiv:2204.07963 (2022)
34. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. arXiv:2305.19906 (2023)
35. 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)
36. Yang, Q., Li, W., Li, B., Yuan, Y.: Mrm: Masked relation modeling for medical image pre-training with genetics. In: Proc. ICCV. pp. 21452–21462 (2023)
37. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: MICCAI. pp. 13–23. Springer (2023)
38. Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H.: Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: CVPR. pp. 15211–15222 (2023)
39. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: MICCAI. pp. 150–160. Springer International Publishing (2021)
40. Zhu, L., Wang, Z., Jin, Z., Lin, G., Yu, L.: Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv:2401.11535 (2024)
41. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. arXiv:2312.00451 (2023)