Free-SurGS: SfM-Free 3D Gaussian Splatting for Surgical Scene Reconstruction

Jiaxin Guo[1], Jiangliu Wang[1], Di Kang[2], Wenzhen Dong[1], Wenting Wang[1], and Yun-hui Liu[1,3](✉)

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Tencent AI Lab, Shenzhen, China
[3] Hong Kong Center for Logistics Robotics, Hong Kong SAR, China

Abstract. Reconstructing surgical scenes plays a vital role in computer-assisted surgery, holding a promise to enhance surgeons' visibility. Recent advancements in 3D Gaussian Splatting (3DGS) have shown great potential for real-time novel view synthesis of general scenes, which relies on accurate poses and point clouds generated by Structure-from-Motion (SfM) for initialization. However, 3DGS with SfM fails to recover accurate camera poses and geometry in surgical scenes due to the challenges of minimal textures and photometric inconsistencies. To tackle this problem, in this paper, we propose the first SfM-free 3DGS-based method for surgical scene reconstruction by jointly optimizing the camera poses and scene representation. Based on the video continuity, the key of our method is to exploit the immediate optical flow priors to guide the projection flow derived from 3D Gaussians. Unlike most previous methods relying on photometric loss only, we formulate the pose estimation problem as minimizing the flow loss between the projection flow and optical flow. A consistency check is further introduced to filter the flow outliers by detecting the rigid and reliable points that satisfy the epipolar geometry. During 3DGS optimization, we randomly sample frames to optimize the scene representations to grow the 3D Gaussians progressively. Experiments on the SCARED dataset demonstrate our superior performance over existing methods in novel view synthesis and pose estimation with high efficiency. Code is available at https://github.com/wrld/Free-SurGS.

Keywords: Novel View Synthesis · 3D Reconstruction · 3D Gaussian Splatting · Endoscopic Surgery.

1 Introduction

Reconstructing surgical scenes is crucial for revealing internal anatomical structures during minimal invasive surgery (MIS), and enables many downstream applications such as augmented reality, virtual reality, surgical planning, and surgical simulation [3, 12, 17]. While neural radiance fields (NeRF) [1] methods demonstrate success for novel view synthesis from multiple photos or videos, their applicability is limited for computational efficiency in training and inference. Recently, 3D Gaussian Splatting (3DGS) [9], which introduces anisotropic

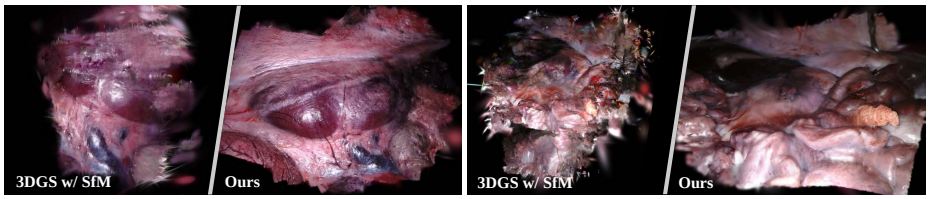


Fig. 1. 3DGS [9] meets a major limitation in its reliance on SfM. We propose Free-SurGS to eliminate this need and demonstrate better performance.

3D Gaussians to build explicit scene representations, emerges as a powerful rendering technique for its rendering efficiency and the ability to produce high-fidelity images. 3DGS showcases significant potential in advancing novel view synthesis, offering a promising pathway to establish real-time, interactive surgical simulations.

Despite the advances, 3DGS encounters a major limitation in its reliance on the camera poses and sparse point clouds from Structure-from-Motion (SfM) [6], which inevitably influences its application in surgical videos. This pre-processing stage is too time-consuming to run for long sequence endoscopic videos, limiting their employment in inter-operative applications. Furthermore, SfM is prone to fail on the appearance of surgical scenes that contain minimal surface textures and photometric inconsistencies like non-Lambertian surfaces, reflective surfaces, and illumination fluctuation. This creates difficulties in detecting features for correspondence search, leading to pose estimation failure and point clouds from incomplete views. As shown in Fig. 1, taking the inaccurate poses and point clouds for initialization, the 3D Gaussians show floaters and artifacts in the rendered images and reconstruct incorrect geometry. To address this issue, some SfM-free studies [2, 4, 7, 11, 19] are proposed to reduce or eliminate the reliance on SfM by estimating the camera poses along with optimizing the scene representations. However, most approaches optimize the camera poses by minimizing the photometric loss between the rendered image and input frame, leading to inaccurate pose estimation due to the homogeneity of textures and photometric inconsistencies.

In this paper, we address the challenges and present Free-SurGS for fast surgical scene reconstruction and real-time rendering from monocular inputs, realizing joint optimization for both 3D Gaussians and camera poses. However, the challenges of the appearance in surgical scenes motivate us to exploit the optical flow priors based on video continuity to guide the projection flow derived from the 3D Gaussians. Our contribution is summarized as threefold: **1)** We present the first SfM-free 3DGS-based approach for fast surgical scene reconstruction and real-time rendering from monocular inputs only. **2)** Unlike previous methods relying on photometric loss only, we formulate the pose estimation problem as matching the projection flow derived from 3D Gaussians with optical flow. A consistency check is further proposed to detect the rigid and reliable points that are consistent with the epipolar geometry. **3)** The extensive experimental

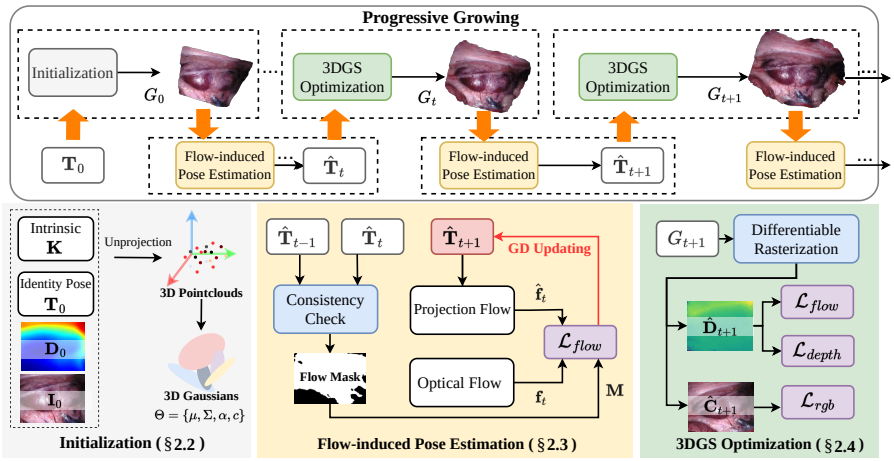


Fig. 2. Overview of our proposed Free-SurGS. Given endoscopic monocular images as input, we jointly estimate the camera poses and optimize the 3D Gaussians iteratively by progressive growing.

results on the SCARED datasets demonstrate that our method outperforms the existing methods in both novel view synthesis and pose estimation, achieving photo-realistic surgical scene rendering with real-time inference speed.

2 Methodology

In this paper, we model the surgical scene as 3D Gaussians to render photo-realistic images from free viewpoints. Given a sequence of monocular images $\{\mathbf{I}_0, \ldots, \mathbf{I}_{N-1}\}$ shot by a moving endoscope, our goal is to better reconstruct the complete surgical scene via a joint optimization of the camera poses and the 3D representation (i.e. 3DGS).

Given the input image sequence, we utilize off-the-shelf methods to obtain the monocular depth $\{\mathbf{D}_t\}_{t=0}^{N-1}$ from Depth-Anything [20] and optical flow between $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ as $\{\mathbf{O}_{t\to t+1}\}_{t=0}^{N-1}$ from RAFT [18] as pseudo-GT. As shown in Fig. 2, we first initialize the 3D Gaussians G_0 from the frame $\mathbf{I}_0$ utilizing the point clouds from monocular depth $\mathbf{D}_0$ and the identity camera pose $\mathbf{T}_0$ (Sec. 2.2). Based on the continuity of surgical video, the 3D Gaussian is updated from every input image consequently following a progressive growing process. We formulate the pose estimation problem as guiding the projection flow of 3D Gaussians with the robust correspondences from $\mathbf{O}_{t\to t+1}$ under a consistency check, to compensate for the limitation of photometric loss (Sec. 2.3). During 3DGS optimization, we randomly sample frames with estimated poses to optimize the scene representation (Sec. 2.4).

2.1 Preliminary: 3D Gaussian Splatting

3DGS [9] introduces the 3D Gaussians as differential volumetric representations of radiance fields, allowing high-quality real-time novel view synthesis. The set of 3D Gaussians is initialized from the calibrated camera poses and sparse point clouds generated from SfM. Each Gaussian is defined by position $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$: $G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})}$. The covariance can be decomposed from a scaling matrix $\mathbf{S}$ and rotation matrix $\mathbf{R}$: $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$. To render a novel view, the 3D Gaussians are projected to 2D camera view $\mathbf{T}$: $\boldsymbol{\Sigma}' = \mathbf{JT}\boldsymbol{\Sigma}\mathbf{T}^T\mathbf{J}^T$, where $\mathbf{J}$ is the Jacobian of the affine approximation of the projective transformation. To render the color, 3DGS further optimizes opacity and SH coefficients, following the point-based differential rendering by rasterizing anisotropic splats with α-blending. The color and depth are rasterized following:

$$\hat{\mathbf{C}} = \sum_i^N \mathbf{c}_i\alpha_i\prod_j^{i-1}(1-\alpha_j), \quad \hat{\mathbf{D}} = \sum_i^N d_i\alpha_i\prod_j^{i-1}(1-\alpha_j), \tag{1}$$

where $\mathbf{c}_i$ and α_i denote the color and opacity of the Gaussian, d_i is the z-axis of the points by projecting the center of 3D Gaussians $\boldsymbol{\mu}$ to the camera coordinate. In summary, the parameters to optimize for the Gaussians include: $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \mathbf{c}\}$. To realize SfM-free scene reconstruction, we need to both recover the camera poses $\mathbf{T}$ and optimize the Gaussian parameters Θ.

2.2 Initialization from Monocular Depth

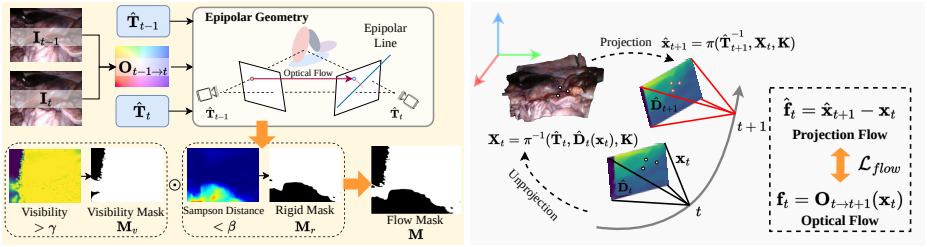
Given first frame $\mathbf{I}_0$ and the known intrinsic $\mathbf{K}$, we generate the pointcloud $\mathbf{P}$ by unprojecting the monocular depth $\mathbf{D}_0$ by the initial identity camera pose $\mathbf{T}_0$: $\mathbf{P} = \pi^{-1}(\mathbf{T}_0, \mathbf{D}_0, \mathbf{K})$, where π^{-1} is the pixel-to-world projection. The center of Gaussians $\boldsymbol{\mu}$ is initialized by $\mathbf{P}$. The color of each point $\mathbf{c}$ is initialized with the SH coefficient from the first frame. Other parameters are initialized following the implementation in 3DGS [9]. After initialization, we optimize the 3D Gaussians G_0 by minimizing the losses introduced in Sec. 2.4.

2.3 Flow-induced Pose Estimation

In this step, we fix the parameters of 3D Gaussians (i.e. assume the current GS is pseudo-GT) and update the camera pose by matching the projection flow from 3D Gaussians with the robust correspondences from filtered optical flow.

Pose Estimation via Pointcloud Transformation. We formulate the camera pose estimation problem into predicting the transformation of 3D Gaussians following [4, 8]. Given the position of Gaussian center $\boldsymbol{\mu}$, we can project it to 2D camera view $\mathbf{T}$ by $\boldsymbol{\mu}_{2D} = \mathbf{K}\frac{\mathbf{T}\boldsymbol{\mu}}{(\mathbf{T}\boldsymbol{\mu})_z}$. Therefore, the camera pose estimation is equivalent to estimating the transformation of 3D Gaussians.

To update the camera pose by gradient descent, we first transform the 3D Gaussians G with the camera pose $\mathbf{T}$. We take the camera poses as the optimizable variables and represent the rotations in quaternion $\mathbf{q}$ and translation



(a) Visibility & Consistency Check (b) Flow Loss

Fig. 3. Illustration of our proposed flow-induced pose estimation. (a) The consistency check is introduced to filter out the outliers in the optical flow map $\mathbf{O}_{t-1\to t}$ to obtain reliable and robust correspondences. (b) We formulate the pose estimation problem as matching the projection flow with the optical flow, to compensate for the limitations of photometric loss.

vector $\mathbf{t}$. At timestep $t+1$, its camera pose $\hat{\mathbf{T}}_{t+1}$ is initialized from the previous camera poses $\hat{\mathbf{T}}_t$ and $\hat{\mathbf{T}}_{t-1}$ based on the constant velocity assumption: $\hat{\mathbf{q}}_{t+1} = \hat{\mathbf{q}}_t + (\hat{\mathbf{q}}_t - \hat{\mathbf{q}}_{t-1}) + \Delta\mathbf{q}$, $\hat{\mathbf{t}}_{t+1} = \hat{\mathbf{t}}_t + (\hat{\mathbf{t}}_t - \hat{\mathbf{t}}_{t-1}) + \Delta\mathbf{t}$.

Previous methods [2, 11] mostly adopt the photometric loss $\mathcal{L}_{rgb}$ to match the rendered color $\hat{\mathbf{C}}_{t+1}$ and ground truth color $\mathbf{I}_{t+1}$ for pose estimation with gradient optimization:

$$\mathcal{L}_{rgb} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}. \tag{2}$$

However, the application of photometric loss in optimizing camera poses within surgical scenes encounters limitations. First, the homogeneity and sparse texturing of the surgical surfaces lead to ambiguities in feature matching. Second, photometric inconsistencies across different views are quite common due to varied lighting conditions, the existence of reflective surfaces of the surgical instruments and tissues, and the presence of non-Lambertian surfaces. Consequently, using only the photometric loss for pose estimation is prone to converge to some local minima, thus leading to inaccurate reconstruction in the following step.

Projection Flow. As shown in Fig. 3(b), we introduce a projection flow to compute the per-pixel movement by projecting the 3D Gaussians from camera view $\hat{\mathbf{T}}_t$ to $\hat{\mathbf{T}}_{t+1}$. Specifically, we first unproject $\mathbf{x}_t$ (i.e. each pixel of $\mathbf{I}_t$) to 3D points $\mathbf{X}_t$ with rendered depth $\mathbf{D}_t$ and $\hat{\mathbf{T}}_t$. Next, the correspondences $\hat{\mathbf{x}}_{t+1}$ can be obtained by projecting $\mathbf{X}_t$ to camera view $\hat{\mathbf{T}}_{t+1}$. The projection flow $\hat{\mathbf{f}}_t$ can be computed by:

$$\begin{aligned} \hat{\mathbf{f}}_t = \hat{\mathbf{x}}_{t+1} - \mathbf{x}_t = \pi(\hat{\mathbf{T}}_{t+1}^{-1}, \mathbf{X}_t, \mathbf{K}) - \mathbf{x}_t, \\ \text{where} \quad \mathbf{X}_t = \pi^{-1}(\hat{\mathbf{T}}_t, \hat{\mathbf{D}}_t(\mathbf{x}_t), \mathbf{K}). \end{aligned} \tag{3}$$

By computing the transformation of 3D Gaussians from one camera view to the next, the projection flow is less dependent on texture variations, making it more reliable in surgical scenes.

Visibility & Consistency Check. First, we employ a visibility check to filter the optical flow from the visibility map to exclude not yet constructed regions. During the first epoch to learn the scene representation, the 3D Gaussians are partially reconstructed, resulting in empty regions in the rendered view. We compute the visibility map $\mathbf{M}_v$ of 3DGS in the rendered view under $\hat{\mathbf{T}}_{t+1}$, by accumulating the opacity of Gaussians under camera view $\hat{\mathbf{T}}_{t+1}$: $\mathbf{M}_v = \sum_i^N \alpha_i \prod_j^{i-1}(1-\alpha_j) > \gamma$, where γ is the threshold for visibility.

Second, a consistency check is introduced to remove the outliers to maintain rigid and reliable points in the optical flow. In dynamic surgical environments characterized by transient objects and photometric inconsistencies, it is essential to identify and preserve correspondences that are both rigid and reliable for accurate matching. Utilizing the optical flow $\mathbf{O}_{t-1\to t}$, we assess the epipolar geometry informed by the estimated camera poses $\hat{\mathbf{T}}_{t-1}$ and $\hat{\mathbf{T}}_t$. This assessment ensures that correctly matched points align with their respective epipolar lines for robust matching. Therefore, we can find the rigid and reliable points that better satisfy the epipolar geometry in t to further filter out outliers in $\mathbf{O}_{t\to t+1}$ based on the continuity of endoscopic video. As shown in Fig. 3(a), we compute the Sampson distance [5] to measure the geometric error between a point in one image and its corresponding epipolar line in the other image. We take a threshold β to obtain a rigid mask $\mathbf{M}_r$ for time t, ensuring that only robust correspondences are utilized for subsequent pose estimation tasks from t to $t+1$. Finally, we obtain the flow mask from the consistency check: $\mathbf{M} = \mathbf{M}_v \odot \mathbf{M}_r$.

Flow Loss. To guide the pose estimation from dense correspondence in optical flow $\mathbf{O}_{t\to t+1}$, the flow loss is defined by minimizing the L_2 loss between the optical flow and projection flow with flow mask $\mathbf{M}$:

$$\mathcal{L}_{flow} = \| \mathbf{M} \odot (\hat{\mathbf{f}}_t - \mathbf{f}_t) \|_2^2, \quad \text{where } \mathbf{f}_t = \mathbf{O}_{t\to t+1}(\mathbf{x}_t). \tag{4}$$

The flow loss compensates for the photometric loss to tackle the challenging surgical scene and enhance the pose estimation accuracy:

$$\hat{\mathbf{T}}_{t+1} = \underset{\mathbf{T}_{t+1}}{\operatorname{argmin}} \ \lambda_1 \mathcal{L}_{rgb} + \lambda_2 \mathcal{L}_{flow}, \tag{5}$$

where λ_1 and λ_2 denote the weight for $\mathcal{L}_{rgb}$ and $\mathcal{L}_{flow}$. By addressing both the geometric consistency through $\mathcal{L}_{flow}$ and the photometric similarity through $\mathcal{L}_{rgb}$, our free-GS ensures a more robust alignment of the camera poses, even in the presence of textural homogeneity or photometric anomalies.

2.4 3D Gaussians Optimization

After estimating the camera pose $\hat{\mathbf{T}}_{t+1}$, we optimize the parameters Θ of 3D Gaussians G. Here, we keep the camera pose fixed and optimize the scene representation by minimizing the photometric loss, depth loss, and flow loss:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \ \lambda_1 \mathcal{L}_{rgb} + \lambda_2 \mathcal{L}_{flow} + \lambda_3 \mathcal{L}_{depth}, \tag{6}$$

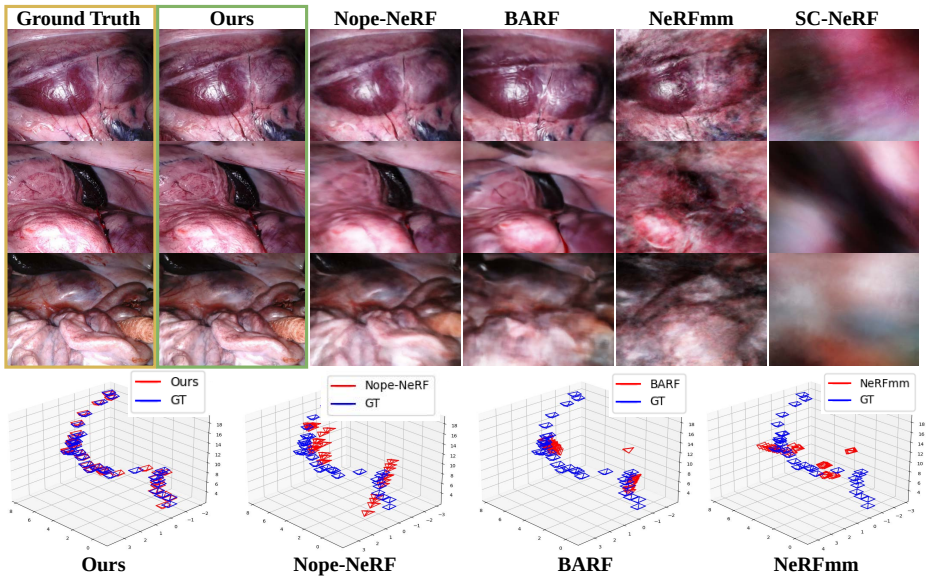


Fig. 4. Qualitative results of novel view synthesis and pose estimation.

where λ_3 is the weight for depth loss, $\mathcal{L}_{depth}$ denotes a scale-invariant loss [16] between rendered depth $\hat{\mathbf{D}}$ and monocular depth $\mathbf{D}$ generated from Depth-Anything [20]. Since the projection flow is derived from rendered depth for reprojection, the flow loss directly contributes to a more precise estimation of depth. By optimizing the 3D Gaussians for both photometric consistency and flow dynamics, the geometry of the 3D Gaussians is not only consistent with the observed image data but also adheres to the expected motion patterns across frames. Finally, we add or prune the 3D Gaussians with adaptive density control, resulting in a progressive growing process for reconstruction.

3 Experiments

3.1 Implementation Details

Experimental Setup. All experiments are implemented using Pytorch [14] on NVIDIA RTX 3090 GPU. We set the same parameters for all the surgical scenes. The optimizer and hyper-parameters of 3D Gaussians follow the original implementation of 3DGS [9]. We use Adam optimizer [10] for pose estimation with a learning rate of 4×10^{-3}. During the progressive growing, we set 30 iterations for both pose estimation and 3DGS optimization.

Datasets. We evaluate our approach on the SCARED Dataset [13], which is a real-world dataset with challenging endoscopic scenes containing reflective surfaces, illumination fluctuations, and weak textures. The image resolution for training and evaluation is 640×480 on the SCARED Dataset. We test 9 scenes

Table 1. Quantitative comparison results on the SCARED Dataset [13].

Methods	Novel View Synthesis			Pose Estimation			Efficiency		
	PSNR ↑	SSIM ↑	LPIPS ↓	RPE_t ↓	RPE_r ↓	ATE ↓	Train ↓	FPS ↑	GPU ↓
SC-NeRF [7]	9.943	0.344	0.654	6.436	6.802	13.67	5.0 h	0.074	**3.2 G**
NeRFmm [19]	16.55	0.361	0.540	5.681	9.108	12.74	9.5 h	0.27	6.0 G
BARF [11]	16.25	0.511	0.658	5.005	6.515	9.832	7.2 h	0.12	8.5 G
Nope-NeRF [2]	21.42	0.620	0.523	5.632	5.685	12.30	50.0 h	0.34	8.0 G
Ours	**24.35**	**0.741**	**0.270**	**3.299**	**1.966**	**5.854**	**1.0 h**	**60.0**	3.8 G

Table 2. Ablation study of flow-induced pose estimation. "Con." refers to the consistency check to maintain rigid and reliable points.

$\mathcal{L}_{rgb}$	$\mathcal{L}_{flow}$	Con.	PSNR ↑	SSIM ↑	LPIPS ↓	RPE_t ↓	RPE_r ↓	ATE ↓
✓			20.57	0.603	0.438	8.574	4.151	10.08
	✓		22.75	0.652	0.382	4.133	2.769	7.410
✓	✓		23.53	0.688	0.291	3.512	2.438	6.435
✓	✓	✓	**24.35**	**0.741**	**0.270**	**3.299**	**1.966**	**5.854**

from the SCARED Dataset with 50-150 frames for each scene with one-eighth of the images for test following [2]. The SCARED Dataset also provides ground truth camera poses of every frame for evaluation.

Evaluation Metrics. We evaluate the performance of novel view synthesis via PSNR, SSIM [21], and LPIPS [15]. To compare the accuracy of estimated camera poses, we evaluate Absolute Trajectory Error (ATE), and Relative Pose Error (RPE), including rotation RPE_r and translation RPE_t following [2]. Note that the unit for RPE_t and ATE is millimeter (mm), and the unit for RPE_r is degree.

3.2 Quantitative and qualitative results

We compare our method with existing state-of-the-art SfM-free methods: Nope-NeRF [2], BARF [11], NeRFmm [19] and SC-NeRF [7]. Quantitative results in Tab. 1 demonstrate that our method outperforms all the baselines. Only based on photometric loss, BARF [11], NeRFmm [19], and SC-NeRF [7] fail to recover the correct camera pose, suffering from the challenging surgical scenes. With constraints from depth distortion, Nope-NeRF [2] improves the performance compared to other baselines but still fails to handle large endoscopic movement (See Fig. 4). Thanks to the flow matching and the consistency check, our Free-SurGS could estimate accurate camera poses for scene reconstruction and render photo-realistic images with 3DGS. The efficiency comparison in Tab. 1 also demonstrates our faster training, higher inference speed, and lower memory of parameters, satisfying real-world surgical applications.

We conduct ablation studies to validate the effectiveness of the proposed modules in Tab. 2. The flow loss $\mathcal{L}_{flow}$ compensates for the limitation of photometric loss and improves the accuracy of pose estimation. The consistency check could further enhance the robustness of large movement and semi-static scenes. With more accurate poses as input, the performance of 3DGS is further improved to reconstruct the surgical scene.

4 Conclusion

In this paper, we propose Free-SurGS as the first SfM-free 3DGS-based method to realize multi-view surgical scene reconstruction. To handle the challenging surgical scene with minimal textures and photometric inconsistencies, we use the optical flow priors to guide the projection flow derived from 3D Gaussians for robust pose estimation. Extensive experiments on the SCARED dataset show that our method outperforms the previous methods in both novel view synthesis and pose estimation, achieving fast reconstruction and real-time rendering with less training time. Our method shows potential to provide a highly realistic and interactive environment that could advance preoperative planning and training practices. However, our method is limited in handling dynamic scenes with severe tissue deformations, which we will address in the future work.

Acknowledgments. This work is supported in part by Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part by the Research Grants Council of Hong Kong under Grant T42-409/18-R, Grant 14218322, and Grant 14207320, in part by the Hong Kong Centre for Logistics Robotics, in part by the Multi-Scale Medical Robotics Centre, InnoHK, and in part by the VC Fund 4930745 of the CUHK T Stone Robotics Institute.

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

References

1. B. Mildenhall *et al.*: Nerf: Representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
2. Bian, W., Wang, Z., Li, K., Bian, J.W., Prisacariu, V.A.: Nope-nerf: Optimising neural radiance field with no pose prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4160–4169 (2023)
3. C. Long *et al.*: Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. Comput. Methods Programs. Biomed. **158**, 135–146 (2018)
4. Fu, Y., Liu, S., Kulkarni, A., Kautz, J., Efros, A.A., Wang, X.: Colmap-free 3d gaussian splatting. arXiv preprint arXiv:2312.07504 (2023)
5. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
6. J. Schonberger *et al.*: Structure-from-motion revisited. In: ICCV. pp. 4104–4113 (2016)
7. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5846–5854 (2021)
8. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. arXiv (2023)
9. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)

10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)
12. Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 3–13. Springer (2020)
13. M. Allan *et al.*: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
15. R. Zhang *et al.*: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
16. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
17. T. Rui *et al.*: Augmented reality technology for preoperative planning and intraoperative navigation during hepatobiliary surgery: A review of current methods. HBPD INT **17**(2), 101–112 (2018)
18. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
19. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
20. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
21. Z. Wang *et al.*: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)