B169 (Fri-AM)

EndoFlow-SLAM: Real-Time Endoscopic SLAM with Flow-Constrained Gaussian Splatting

Taoyu Wu[1], Yiyi Miao[1], Zhuoxiao Li [1,2], Haocheng Zhao[1] , KangDang[1] , Jionglong Su[1] , Limin Yu [1], and Haoang Li [2]

[1]Xi'an Jiaotong Liverpool University [2]The Hong Kong University of Science and Technology (Guangzhou)

Abstract: Efficient three-dimensional reconstruction and real-time vi-sualization are critical in surgical scenarios such as endoscopy. In recentyears, 3D Gaussian Splatting (3DGS) has demonstrated remarkable performance in efficient 3D reconstruction and rendering. Most 3DGS-basedSimultaneous Localization and Mapping (SLAM) methods only rely onthe appearance constraints for optimizing both 3DGS and camera poses. However, in endoscopic scenarios, the challenges include photometric in-consistencies caused by non-Lambertian surfaces and dynamic motionfrom breathing affects the performance of SLAM systems. To addressthese issues, we additionally introduce optical flow loss as a geomet-ric constraint, which effectively constrains both the 3D structure of thescene and the camera motion. Furthermore, we propose a depth reg-ularisation strategy to mitigate the problem of photometric inconsis-tencies and ensure the validity of 3DGS depth rendering in endoscopicscenes. In addition, to improve scene representation in the SLAM sys-tem, we improve the 3DGS refinement strategy by focusing on view-points corresponding to Keyframes with suboptimal rendering qualityframes, achieving better rendering results. Extensive experiments onthe C3VD static dataset and the StereoMIS dynamic dataset demonstrate that our method outperforms existing state-of-the-art methods innovel view synthesis and pose estimation, exhibiting high performancein both static and slightly dynamic surgical scenes.

Background

1、Importance in Surgery

Accurate camera pose estimation and tissue reconstruction are crucial for minimally invasive surgeries. They improve spatial awareness between organs and instruments, aiding surgical navigation and postoperative evaluation.

2、Challenges in Endoscopy

Limited fields of view, dynamic tissue deformations, and complex lighting conditions reduce accuracy. Traditional visual SLAM often assumes rigid environments and relies on sparse reconstructions, which are unsuitable for endoscopic scenarios with weak textures.

3、Limitations of Dense SLAM

Dense SLAM enables real-time reconstructions but typically depends on RGB-D data. However, in endoscopic environments, limited instrument mobility and access constraints hinder effective data acquisition, leading to incomplete and occluded scene representations.

4、Advances with NeRF and 3DGS

Recent SLAM systems integrate Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS). NeRF offers high-precision view synthesis but is computationally heavy, while 3DGS provides faster rendering with realistic images, achieving effective tracking and dense reconstruction.

Methodology

Preliminary: 3D Gaussian Splatting

3DGS [10] is a differentiable rendering framework that models a 3D scene using a collection of Gaussian primitives. The i-th Gaussian is characterized by a center μ_i, an opacity o_i, and a covariance matrix Σ_i and can be expressed as:

$$G_i(\mathbf{X}) = o_i \cdot \exp\left\{-\frac{1}{2}(\mathbf{X}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{X}-\mu_i)\right\}, \quad (1)$$

where $\mathbf{X}$ represents an arbitrary point in 3D space. In the rendering process, these Gaussians are projected onto the 2D image plane along camera rays, and their contributions are composited using an alpha-blending scheme.

Non-keyframe Optimization

Existing 3DGS-based SLAM systems [25,16,19] typically employ photometric ($\mathcal{L}_{\text{rgb}}$) and depth ($\mathcal{L}_{\text{depth}}$) error minimization for camera pose estimation. However, this paradigm faces scale ambiguity challenges in endoscopic scenarios due to monocular depth estimation limitations [28,15]. To address this limitation, we propose a modified optimization framework incorporating scale-invariant loss $\mathcal{L}_{\text{scale}}$ [21] and depth gradient regularization $\mathcal{L}_{\text{depth}}^{\text{reg}}$. The depth regularization term is defined as the weighted gradient difference between the estimated and ground truth depth maps: $\mathcal{L}_{\text{depth}}^{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(w_h \cdot |\nabla_h d_i| + w_v \cdot |\nabla_v d_i|)$, where d_i denotes the depth value at pixel i, ∇_h and ∇_v represent horizontal and vertical gradients respectively, w_h and w_v are corresponding weight factors, and N represents the total number of image pixels. The estimated camera pose $\hat{\mathcal{T}}_{t+1}$ at time $t+1$ can be optimized by minimizing the photometric, scale-invariant and depth regularization losses:

$$\hat{\mathcal{T}}_{t+1} = \underset{\mathcal{T}_{t+1}}{\operatorname{argmin}} \lambda_1 \mathcal{L}_{\text{rgb}} + \lambda_2 \mathcal{L}_{\text{depth}}^{\text{reg}} + \lambda_3 \mathcal{L}_{\text{scale}}, \quad (2)$$

Keyframe Optimization with Flow Constraints

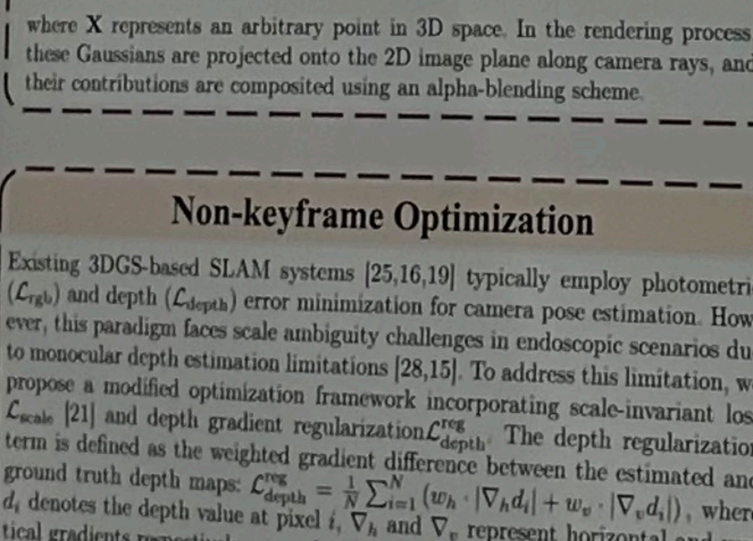


Fig. 2. GaussianFlow estimation. At time t, each pixel x_t results from K overlapping Gaussians. At time $t+1$, each K Gaussian will have a corresponding Gaussian flow (Black arrow). By accumulating these Gaussian flows, we obtain the overall Gaussian flow. Our goal is to minimize the difference between GS fLow and Optical flow by optimizing both the camera pose $\hat{\mathcal{T}}_{t+1}$ and 3DGS primitive $\hat{\mathcal{G}}$.

Keyframe Optimization and Global Refinement

Keyframe-Oriented Local Bundle Adjustment

Keyframe-Oriented Local Bundle Adjustment. To further improve the expressiveness of the scene, we also introduce flow loss as a geometric constraint in Mapping. After the camera tracking module, we perform 3D Gaussian map representation optimization only when frame $\mathbf{I}_{t+1}$ is designated as a Keyframe. We follow the Keyframe management strategy from [19]. We perform BA on all frames within the Keyframe window. By minimizing the loss function $\mathcal{L}_{\text{rgb}}$, $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{flow}}$ in the objective function (3), we simultaneously optimize the estimated camera pose $\hat{\mathcal{T}}_{t+1}$ and 3DGS primitive $\hat{\mathcal{G}}$.

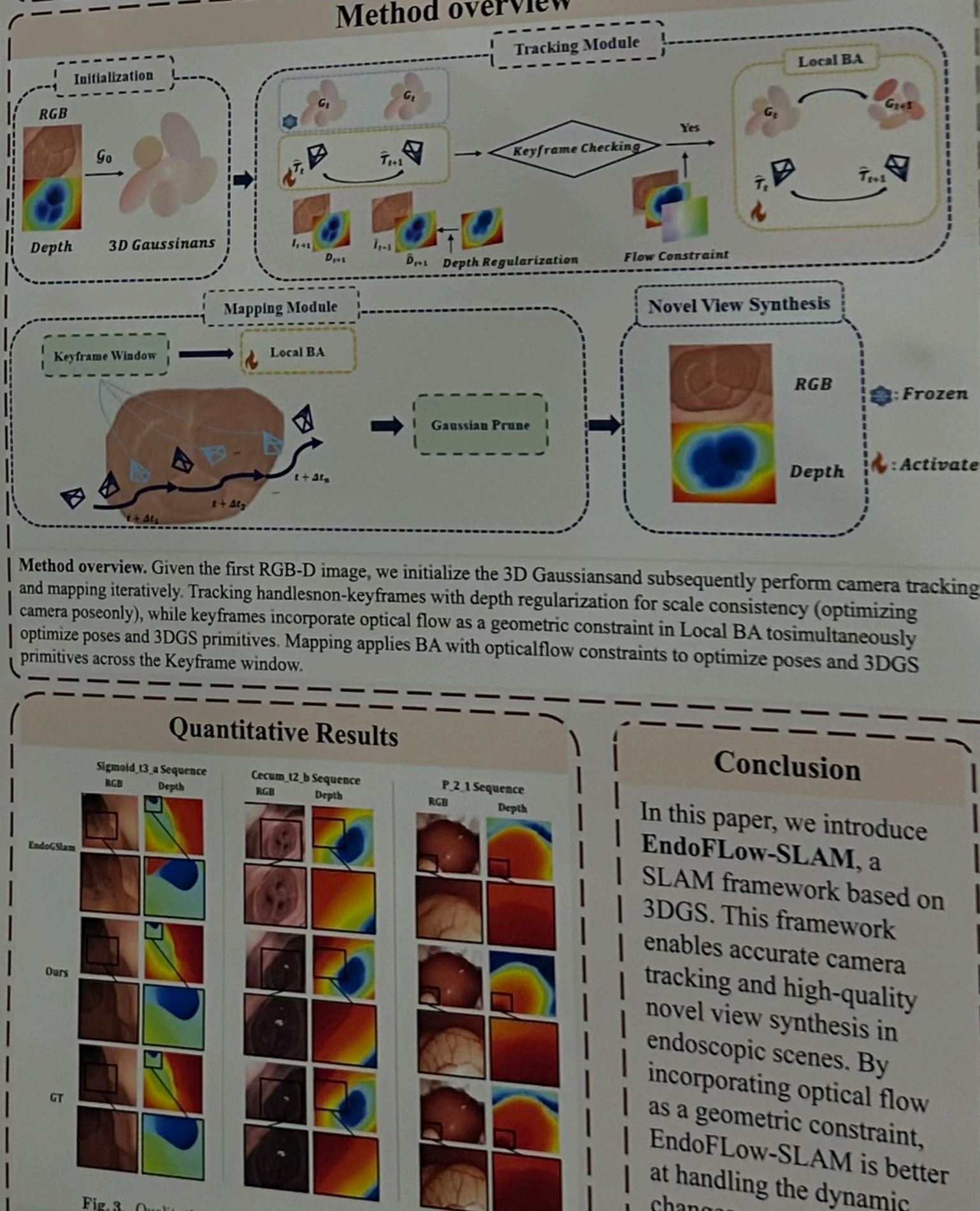
Global Refinement

For each selected viewpoint, the optimization of the Gaussian primitive is performed using the following loss function:

$$\mathcal{L}_{\text{refine}} = (1-\lambda_{\text{ssim}})\mathcal{L}_{\text{rgb}} + \lambda_{\text{ssim}}(1-\text{SSIM}) + |\nabla\hat{D} - \nabla D|_2^2, \quad (4)$$

where λ_{ssim} controls the relative weight of the Structural Similarity Index (SSIM). The SSIM term ensures structural preservation, maintaining the integrity of the scene's overall structure while minimizing pixel-level discrepancies.

Method overview



Method overview. Given the first RGB-D image, we initialize the 3D Gaussiansand subsequently perform camera tracking and mapping iteratively. Tracking handlesnon-keyframes with depth regularization for scale consistency (optimizing camera poseonly), while keyframes incorporate optical flow as a geometric constraint in Local BA tosimultaneously optimize poses and 3DGS primitives. Mapping applies BA with opticalflow constraints to optimize poses and 3DGS primitives across the Keyframe window.

Quantitative Results

Fig. 3. Qualitative results on C3VD and StereoMIS Dataset.

Qualitative Results

Table 1. Quantitative results on the C3VD dataset.

Methods	PSNR↑	SSIM↑	LPIPS↓	RMSE (mm)↓	ATE (mm)↓
NICE-SLAM [29]	22.07	0.73	0.33	1.88	0.48
Endo-Depth [22]	18.13	0.64	0.33	5.10	1.25
EndoGSLAM-H [25]	22.16	0.77	**0.22**	2.17	0.34
Ours	**25.18**	**0.82**	0.27	**1.54**	**0.23**

Table 2. Quantitative results on the StereoMIS dataset.

Methods	PSNR↑	SSIM↑	LPIPS↓	ATE (mm) ↓
NICESLAM [29]	13.07	0.49	0.61	38.24
ESLAM [8]	18.70	0.54	0.57	16.73
EndoGSLAM-H [25]	16.67	0.52	0.45	18.82
Ours	**21.96**	**0.59**	**0.27**	**15.63**

Conclusion

In this paper, we introduce **EndoFLow-SLAM**, a SLAM framework based on 3DGS. This framework enables accurate camera tracking and high-quality novel view synthesis in endoscopic scenes. By incorporating optical flow as a geometric constraint, EndoFLow-SLAM is better at handling the dynamic changes caused by breathing in real-world endoscopic environments. Extensive experimental results demonstrate that, compared to traditional SLAM methods and those based on 3DGS, EndoFLow-SLAM achieves superior tracking and rendering performance. Future work will focus on the construction and optimization of object-level 3DGS to better adapt the intense d