

Track2Map: Online Deformable SLAM with Motion-Aware Pose Optimization in Robotic Surgery

Tianyi Song^{1*}(✉), Sierra Bonilla^{1*}, Xinwei Ju¹, Evangelos Mazomenos¹, Danail Stoyanov¹, Adam Schmidt², Omid Mohareri², Sophia Bano¹, and Francisco Vasconcelos¹

¹ Department of Computer Science and UCL Hawkes Institute, University College London, London WC1E 6BT, UK

² Intuitive Surgical, Inc., Sunnyvale, CA 94086, USA
{tianyi.song.24,sierra.bonilla.21}@ucl.ac.uk

Abstract. Gaussian splatting is the current state-of-the-art for dense, deformable 3D anatomy reconstruction in robot-assisted minimally invasive surgery (RAMIS); however, most pipelines are offline and depend on accurate camera trajectory priors (often from robotic kinematics), limiting applicability when priors are missing or noisy. To address these limitations, we propose **Track2Map**, an online 3D Gaussian Splatting pipeline that jointly optimizes camera trajectory and 3D deformable scene representation directly from surgical video. Track2Map is therefore capable of robust 3D reconstructions when camera trajectory priors are either absent or noisy, and due to its online nature it effectively works as a Simultaneous Localisation and Mapping (SLAM) method. To stabilize optimization in the presence of tissue motion and ambiguous visual cues, we introduce a track-anchored deformation initialization using dense 2D point tracks. Track statistics are further utilized to disentangle camera motion from scene deformation by detecting static camera periods and reducing drift during incremental mapping. Experiments on StereoMIS show improved reconstruction quality and camera trajectory against competing SLAM methods, as well as compared to non-SLAM methods that utilize camera trajectory priors. The code is available at <https://track2map.github.io/>.

Keywords: Robot-assisted surgery · Stereo endoscopy · Online 3D reconstruction · Camera pose optimization

1 Introduction

Online 3D reconstruction of deformable tissue is a key prerequisite for developing a range of computer guidance techniques in RAMIS, including characterisation of instrument-tissue interactions, autonomous manipulation, and stable

* These authors contributed equally.

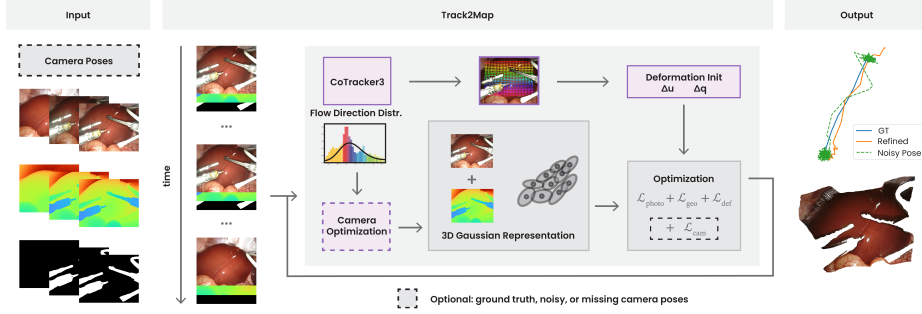


Fig. 1. Given stereo RGB frames, depth maps, and *optional* (clean/noisy) pose extrinsics/intrinsics and tool masks, Track2Map jointly performs 2D tracking and online deformable 3D reconstruction. [12] provides 2D correspondences that are lifted to 3D and used to estimate/refine camera motion via motion-gated pose optimization and initialize sparse anchor-based deformation. A Gaussian representation like [8] is updated online. The system outputs refined poses and metric reconstruction; 2D/3D tracks are available as by-products of the correspondence module and 3D lifting.

augmented reality overlays. However, achieving metric, stable reconstruction in RAMIS remains difficult because the scene is non-rigid, frequently occluded by instruments, and captured with limited viewpoint variation.

3D reconstruction methods have progressed from classical SfM [19] to neural radiance fields [15] and 3D Gaussian Splatting (3DGS) [14] representations tailored to endoscopy [4,25,21]. However, extending these representations to deformable surgical scenes introduces a fundamental coupling problem: camera pose, non-rigid deformation, and appearance/geometry estimation must be estimated jointly, but supervision and priors are often unreliable. Most surgical neural/3DGS methods either assume static or known camera motion, or depend on relatively clean pose inputs to prevent drift in deformable sequences [8,10]. In RAMIS, camera pose is often calculated from robot joint kinematics, which can be inaccurate due to calibration errors and measurement noise, and may also not always be available in a clinical setting. Therefore, camera motion availability and quality varies widely across available surgical datasets; some provide accurate trajectories [7], others provide noisy priors [3], and many tracking benchmarks provide none [5,18]. This motivates deformable reconstruction systems that can exploit pose priors when available, remain stable when priors are corrupted, and can initialize and operate when pose is absent.

Joint estimation of camera motion and deformable structure has a long history in NRSfM and non-rigid SLAM, including locally NRSfM [22], sequential physical-prior NRSfM [2], online self-calibrating NRSfM [1], and deformable SLAM [28]. These works establish the broader paradigm of coupling camera motion with non-rigid scene structure, while our setting targets online dense 3DGS reconstruction from surgical stereo video with optional or noisy pose priors. In surgical neural/3DGS mapping, recent systems have begun to jointly optimize

camera motion and deformable scene structure [24,27]. However, a key failure mode persists in RAMIS: endoscope motion is often intermittent, so long segments have a nearly static camera while tools deform tissue. In these intervals, image motion is dominated by local deformation and instrument motion, and indiscriminate pose optimization can incorrectly attribute tissue dynamics to camera motion, causing trajectory drift and destabilizing reconstruction. Our work addresses these issues by introducing two key additions. First, rather than completely eliminating pose priors, we use them as initial estimates whenever available, making our system flexible enough to operate both as a SLAM system and as an online 3D reconstruction pipeline that can handle noisy camera priors. Second, we use dense 2D point tracking to disentangle camera motion from tissue deformation. This allows us to detect when the camera view is static and dynamically introduce camera pose constraints that improve 3D reconstruction quality and reduce camera trajectory drift. Modern methods for dense 2D point tracking build on deep optical flow [23] and sequence models [16,12]. Surgical extensions improve robustness and efficiency [13,20,18]; in this work, we use CoTracker3 [12] as our 2D tracking initialization for Gaussian Splatting after empirically verifying its state-of-the-art performance on the STIR2024 benchmark [18].

Experiments on StereoMIS [7] demonstrate Track2Map’s improved camera tracking and scene reconstruction over baselines including [16,12], and robustness under pose corruption and absence. Our core contributions are as follows:

- An online stereo 3D Gaussian reconstruction pipeline in which dense 2D correspondences are lifted to 3D anchors to initialize and constrain deformation.
- A pose update strategy that prevents drift during camera-static periods and enables joint optimization of pose and deformation when motion is detected.
- A unified system that operates with *clean*, *noisy*, or *absent* pose priors without retuning, including a vision-only initialization via robust 3D alignment of tracked anchor points when priors are unavailable.
- Improved reconstruction quality and robust pose recovery compared to prior deformable 3DGS pipelines.

2 Method

Fig. 1 summarizes Track2Map’s online inference loop. The central technical challenge is disentangling camera motion from non-rigid tissue motion under limited viewpoint change: inverse rendering on essentially one viewpoint (two short-baseline stereo images) provides weak constraints on pose when the scene itself also deforms, and unconstrained pose updates can easily absorb tissue deformation or tool motion, leading to significant errors. Track2Map therefore injects two procedure-driven priors into the optimization: (i) stereo depth provides metric 3D support for lifting correspondences, and (ii) endoscope motion is typically intermittent. Within RAMIS, the endoscope arm is often stationary while surgical tools operate on the scene. Therefore, pose should be updated conservatively

and only when the observed motion is consistent with camera motion rather than local deformation or tool interaction.

Following [8], we represent the scene as 3D Gaussians [14] coupled with a sparse set of deformation control points. We use the tracked correspondences from [12] as the primary coupling signal between tracking and mapping in three stages: (1) compute a robust per-frame motion score from track displacements and use this score to gate pose updates to occur only when global motion is detected (Sec. 2.1); (2) lift the same correspondences to 3D and use their displacements to initialize a sparse anchor deformation prior, interpolated to the full set of Gaussians (Sec. 2.2); and (3) jointly refine scene parameters (and pose when enabled) using photometric and geometric rendering losses, deformation regularization, and 2D reprojection consistency of anchors (Sec. 2.3).

2.1 Motion-Aware Camera Update

At frame t , the tracker [12] works on a sliding window and provides 2D correspondences $\{(\mathbf{q}_k^{t-1}, \mathbf{q}_k^t)\}_{k \in \mathcal{K}_t}$ with flow $\mathbf{f}_k^t = \mathbf{q}_k^t - \mathbf{q}_k^{t-1}$. We use these tracks to decide when pose refinement is warranted. When the camera moves, tracked points exhibit a coherent global displacement field; when the camera is stationary, observed motion is dominated by heterogeneous tissue deformation and tool interaction. So we gate pose refinement using a simple statistic of the optical-flow direction distribution. We compute the flow direction, $\theta_k^t = \text{atan2}(f_{k,y}^t, f_{k,x}^t)$. Empirically, camera motion produces a narrow, unimodal distribution of $\{\theta_k^t\}$ (low dispersion), whereas camera-static periods produce a broad distribution (high dispersion) due to local tissue deformation, tool motion, and physiological effects (distribution visualizations are provided in the supplementary material). We quantify dispersion using the (circular) standard deviation of flow directions. Let

$$\bar{c}_t = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} \cos \theta_k^t, \quad \bar{s}_t = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} \sin \theta_k^t, \quad R_t = \sqrt{\bar{c}_t^2 + \bar{s}_t^2} \quad (1)$$

We then compute the circular standard deviation $\sigma_{\theta,t} = \sqrt{-2 \ln(R_t)}$. Low $\sigma_{\theta,t}$ indicates camera-dominant motion, while high $\sigma_{\theta,t}$ indicates scene motion. The motion gate is $g_t = \mathbb{I}[\sigma_{\theta,t} \leq \tau]$ where τ is set empirically.

When no pose prior is available, we initialize the relative motion $\Delta \mathbf{T}_t$ by aligning 3D anchor positions across frames where anchors are initialized once by sparsely subsampling gaussian centers similarly to [8]. Each tracked 2D point $\mathbf{q}_k^t$ is lifted to 3D using stereo depth and intrinsics, yielding $\mathbf{P}_k^{t-1}$ and $\mathbf{P}_k^t$. We estimate $\Delta \mathbf{T}_t \in SE(3)$ via robust weighted alignment:

$$\Delta \mathbf{T}_t = \arg \min_{\Delta \mathbf{T} \in SE(3)} \sum_{k \in \mathcal{K}_t} \omega_k \rho(\|\mathbf{P}_k^t - \Delta \mathbf{T} \mathbf{P}_k^{t-1}\|_2^2) \quad (2)$$

where ω_k is the tracker confidence and $\rho(\cdot)$ is a robust kernel. We set $\mathbf{T}_t^{\text{init}} = \Delta \mathbf{T}_t \mathbf{T}_{t-1}$. If $g_t=0$, we freeze the pose to avoid absorbing non-rigid motion into

the camera trajectory. If $g_t=1$, we refine pose with a small 6D increment $\delta\boldsymbol{\xi} \in \mathbb{R}^6$ (3D translation + 3D rotation in $\mathfrak{se}(3)$), optimized with the scene (Sec. 2.3) where $\exp(\cdot^\wedge)$ projects the vector on to a valid transformation matrix:

$$\mathbf{T}_t = \exp(\delta\boldsymbol{\xi}^\wedge) \mathbf{T}_t^{\text{init}} \quad (3)$$

2.2 Tracker-Guided Deformation Initialization

We represent tissue deformation using the same sparse anchor set $\mathcal{K}$ used for pose estimation (Eq. 2). Unlike prior work [8], which treats control-point offsets as free variables optimized per frame, we explicitly drive the deformation field using tracked anchor motion. For each anchor $k \in \mathcal{K}$, its 3D displacement between consecutive frames is:

$$\Delta\mathbf{P}_k = \mathbf{P}_k^t - \mathbf{P}_k^{t-1}, \quad (4)$$

where $\mathbf{P}_k^t$ is obtained by lifting the tracked 2D correspondence $\mathbf{q}_k^t$ using stereo depth and the current camera pose $\mathbf{T}_t$. The initial deformation offset $\Delta\mathbf{x}_i^{(0)}$ for each Gaussian center $\mathbf{x}_i$ is then interpolated over its N nearest anchors:

$$\Delta\mathbf{x}_i^{(0)} = \sum_{k \in \mathcal{N}_i} w_{ik} \Delta\mathbf{P}_k, \quad w_{ik} \propto \exp(-\alpha \|\mathbf{x}_i - \mathbf{P}_k^{t-1}\|_2) \quad (5)$$

2.3 Joint Online Optimization

Given the deformation initialization from Sec. 2.2 and the pose initialization from Sec. 2.1, we jointly refine the 3D scene and camera pose. The scene loss is:

$$\mathcal{L}_{\text{scene}} = w_{\text{photo}} \mathcal{L}_{\text{photo}} + w_{\text{geo}} \mathcal{L}_{\text{geo}} + w_{\text{def}} \mathcal{L}_{\text{def}} \quad (6)$$

where $\mathcal{L}_{\text{photo}}$ and $\mathcal{L}_{\text{geo}}$ are masked photometric and geometric rendering losses, and $\mathcal{L}_{\text{def}}$ regularizes the deformation field, as in [8]. The photometric and geometric losses are defined as the average distance between the rendered $(\hat{I}, \hat{D})$ and input (I, D) color and depth maps over valid pixels Ω_t and the deformation loss is defined by $\mathcal{L}_{\text{def}} = \sum_k \lambda_k \mathcal{L}_k$, incorporating local rigidity, isometry, and rotation consistency losses as in [8]. When the camera is static ($g_t = 0$), only $\mathcal{L}_{\text{scene}}$ is minimized. If $g_t = 1$, we additionally optimize an incremental pose update $\delta\boldsymbol{\xi} \in SE(3)$ via:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{scene}} + w_{\text{trk}} \mathcal{L}_{\text{trk}} + w_{\text{reg}} \|\delta\boldsymbol{\xi}\|^2, \quad (7)$$

where the $\mathcal{L}_{\text{trk}}$ loss enforces 2D reprojection consistency of the tracked anchors under the estimated camera motion:

$$\mathcal{L}_{\text{trk}} = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} \|\mathbf{q}_k^t - \text{proj}(\mathbf{T}_t \mathbf{P}_k^{t-1})\|_1 \quad (8)$$

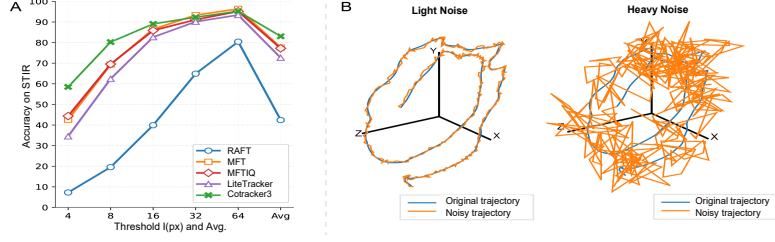


Fig. 2. Illustrating method and experimental design choices. A) 2D tracker selection showing SOTA tracking accuracy across pixel thresholds and average on STIR [17], attributed to [12]. B) Pose noise perturbation strategy applied to sequence P2_1 from [7]. Light noise: $\sigma_t = 6 \times 10^{-4}$, $\sigma_r = 0.6^\circ$. Heavy noise: $\sigma_t = 6 \times 10^{-3}$, $\sigma_r = 0.6^\circ$.

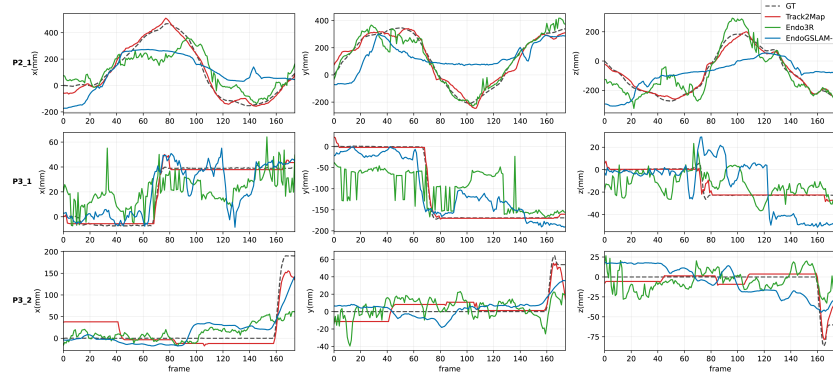


Fig. 3. Comparison of estimation of pose on three StereoMIS [7] sequences compared to a direct pose estimation network [6] and SLAM method [25].

3 Experimental Setup and Results

3.1 Experimental setup

We benchmark 2D point tracker accuracy on STIRC2024 [18] to select CoTracker3 as the correspondence module used by Track2Map (Fig. 2A). In Tab. 1 we evaluate the scene reconstruction quality of Track2Map. Without pose priors, we compare it against SLAM baselines, and with pose priors (clean and noisy) we compare it against [8]. The noise magnitudes injected to pose priors in this experiment can be visualised in Fig. 2B. In Tab. 2 and Fig. 3 we evaluate camera pose estimation against a direct pose regression model [6] and a SLAM baseline [25]. We perform an ablation in Tab. 3 to quantify the effect of pose optimisation, static camera gating, and deformation in our method.

We use public CoTracker3 online weights [12] with an 8-frame sliding window, stereo depth from [26], a conservative motion-gating threshold of 0.38 (Sec. 2.1), and PyTorch on an NVIDIA Tesla V100. Key loss weights are set as: $w_{\text{depth}}=10$, $w_{\text{color}}=5$, $w_{\text{pose_prior}}=0.25$.

Table 1. 3D reconstruction performance on StereoMIS grouped by pose initialization regime. Online-endo-track [8] is evaluated only when poses are available. Track2Map (Monocular) uses left RGB with Depth Anything V2 [29]; Track2Map (Stereo) is the main setting.

Pose Init	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
None	Endo-2DTAM [9]	15.15	0.35	0.55
	EndoGSLAM-H [25]	16.67	0.52	0.45
	ESLAM [11]	18.70	0.54	0.57
	EndoSD-SLAM [24]	19.56	0.65	0.31
	EndoFlow-SLAM [27]	21.96	0.59	0.27
	Track2Map (Monocular)	26.70	0.70	0.31
	Track2Map (Stereo)	27.58	0.745	0.274
Noisy	Online-endo-track(light noise) [8]	27.31	0.727	0.349
	Online-endo-track(heavy noise) [8]	24.78	0.661	0.452
	Track2Map (Stereo)(light noise)	27.68	0.747	0.276
	Track2Map (Stereo)(heavy noise)	27.56	0.742	0.302
Clean	Online-endo-track [8]	27.58	0.744	0.235
	Track2Map (Stereo)	27.83	0.750	0.201

Table 2. Pose estimation on StereoMIS. ATE and consecutive-motion RPE are reported.

Method	ATE(m) $\downarrow$	RPE _r (deg) $\downarrow$	RPE _t (m) $\downarrow$
EndoGSLAM-H [25]	0.1077	2.4840	0.0131
Endo3R [6]	0.0897	3.4796	0.0354
Track2Map	0.0285	0.2903	0.0061

3.2 Results & Discussion

Tab. 1 summarizes reconstruction on StereoMIS under absent, noisy, and clean pose initialization. In the no-pose regime, Track2Map substantially improves PSNR and SSIM over prior endoscopic SLAM and neural mapping baselines, and achieves the strongest overall reconstruction when run in stereo. This improvement is likely driven by two factors: stereo input grounds the reconstruction in a consistent metric scale, and the coupling of online mapping with motion-aware pose updates prevents pose drift during camera-static intervals, stabilizing the incremental mapping process under deformable motion. These components allow the system to recover coherent geometry without relying on robot kinematics or precomputed trajectories. When noisy pose priors are provided, Track2Map has nearly the same reconstruction results as compared to the clean setting. In contrast, methods that depend more directly on the accuracy of pose initialization, such as [8], exhibit larger performance drops under heavy corruption. The qualitative comparison in Fig. 4 further illustrates this robustness under heavy pose noise: Track2Map maintains more coherent anatomy and fewer visible geometric

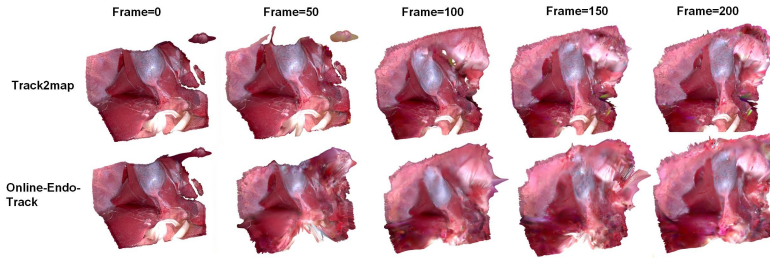


Fig. 4. Comparison of reconstruction under heavy noise poses.

Table 3. Reconstruction ablations on StereoMIS under noisy pose priors.

Method Variant	Heavy noise			Light noise		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o Pose Opt.	25.42	0.676	0.359	27.54	0.744	0.244
w/o Gating	24.41	0.632	0.410	26.43	0.675	0.263
w/o Deformation	20.64	0.602	0.451	23.25	0.638	0.330
Full Track2Map	27.56	0.742	0.302	27.68	0.747	0.276

distortions than the pose-prior baseline, consistent with the quantitative gains in Tab. 1. For pose estimation, Track2Map achieves lower ATE and RPE than both Endo3R [6] and EndoGSLAM [25] (Tab. 2), with trajectories visually improved in Fig. 3.

Tab. 3 isolates each component under noisy pose priors. Removing deformation causes the largest drop, confirming that deformation priors are critical for stable reconstruction from limited viewpoints. Disabling motion gating also degrades performance, especially under heavy noise, as local tissue motion can be absorbed into the camera trajectory. Removing pose optimization further reduces quality when pose noise is large. Overall, the full model performs best, showing that deformation modeling, motion-aware gating, and pose refinement are complementary.

Limitations and assumptions. The current implementation of Track2Map is not real time, running at 6 seconds per frame, so we recommend keyframe processing. The motion gate assumes that coherent global flow is more likely to indicate camera motion, whereas heterogeneous or local flow is more likely to indicate tissue/tool deformation. This assumption may fail under optical-axis motion in tubular structures, large coherent tissue motion, tracker failure, substantial tool occlusion, or weakly textured tissue. Finally, evaluation is mainly performed on StereoMIS, so broader generalization should be validated on additional datasets and surgical scenarios.

4 Conclusion

We presented Track2Map, an online deformable reconstruction system for surgical video that remains stable when camera pose priors are missing or corrupted. Across StereoMIS, it improves reconstruction quality under clean, noisy, and no-pose settings, and yields more accurate camera trajectories. Robust deformable mapping in RAMIS can therefore be achieved without precise kinematics by leveraging constraints from stereo depth, consistent 2D correspondences, and assumptions about typical endoscopic motion during surgical procedures.

Acknowledgments. This work was supported in whole, or in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z], EPSRC Optical and Acoustic Imaging for Surgical and Interventional Sciences (OASIS) [UKRI145], the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies programme, EPSRC [EP/W00805X/1, EP/Y01958X/1], the EPSRC HuMIRoS grant [EP/Z534754/1], and the UCL Centre for Digital Innovation (CDI) Scholarship. For open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

Disclosure of Interests. Some authors have current or prior affiliations with an industry partner involved in robotic surgery. The authors declare no other competing interests relevant to the content of this work.

References

1. Agudo, A.: Total estimation from rgb video: on-line camera self-calibration, non-rigid shape and motion. In: 2020 25th international conference on pattern recognition (ICPR). pp. 8140–8147. IEEE (2021)
2. Agudo, A., Moreno-Noguer, F., Calvo, B., Montiel, J.M.M.: Sequential non-rigid structure from motion using physical priors. *IEEE transactions on pattern analysis and machine intelligence* **38**(5), 979–994 (2015)
3. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis* **90**, 102956 (2023)
4. Bonilla, S., Zhang, S., Psychogyios, D., Stoyanov, D., Vasconcelos, F., Bano, S.: Gaussian pancakes: geometrically-regularized 3d gaussian splatting for realistic endoscopic reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 274–283. Springer (2024)
5. Cartucho, J., Weld, A., Tukra, S., Xu, H., Matsuzaki, H., Ishikawa, T., Kwon, M., Jang, Y.E., Kim, K.J., Lee, G., et al.: Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery. *Medical image analysis* **91**, 102985 (2024)
6. Guo, J., Dong, W., Huang, T., Ding, H., Wang, Z., Kuang, H., Dou, Q., Liu, Y.H.: Endo3r: Unified online reconstruction from dynamic monocular endoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 170–180. Springer (2025)
7. 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* (2023)

8. Hayoz, M., Hahne, C., Kurmann, T., Allan, M., Beldi, G., Candinas, D., Márquez-Neila, P., Sznitman, R.: Online 3d reconstruction and dense tracking in endoscopic videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 444–454. Springer (2024)
9. Huang, Y., Cui, B., Bai, L., Chen, Z., Wu, J., Li, Z., Liu, H., Ren, H.: Advancing dense endoscopic reconstruction with gaussian splatting-driven surface normal-aware tracking and mapping. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 4084–4091. IEEE (2025)
10. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2024)
11. Johari, M.M., Carta, C., Fleuret, F.: Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17408–17419 (2023)
12. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
13. Karaoglu, M.A., Ji, W., Abbas, A., Navab, N., Busam, B., Ladikos, A.: Litetracker: Leveraging temporal causality for accurate low-latency tissue tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 308–317. Springer (2025)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
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. Neoral, M., Šerých, J., Matas, J.: Mft: Long-term tracking of every pixel. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6837–6847 (2024)
17. Schmidt, A., Karaoglu, M.A., Sinha, S., Jang, M., Ha, H.G., Jung, K., Gu, K., Ullah, I., Lee, H., Šerých, J., et al.: Point tracking in surgery—the 2024 surgical tattoos in infrared (stir) challenge. *arXiv preprint arXiv:2503.24306* (2025)
18. Schmidt, A., Mohareri, O., DiMaio, S., Yip, M.C., Salcudean, S.E.: Tracking and mapping in medical computer vision: A review. *Medical Image Analysis* (2024)
19. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
20. Seenivasan, L., Cheng, J., Fang, J., Bai, J., Soberanis-Mukul, R.D., Mangulabnan, J.E., Vedula, S.S., Ishii, M., Hager, G., Taylor, R.H., et al.: Endotracker: Robustly tracking any point in endoscopic surgical scene. In: International Workshop on Foundation Models for General Medical AI. pp. 140–149. Springer (2025)
21. Song, T., Stoyanov, D., Mazomenos, E., Vasconcelos, F.: Diff2dgs: Reliable reconstruction of occluded surgical scenes via 2d gaussian splatting. *arXiv preprint arXiv:2602.18314* (2026)
22. Taylor, J., Jepson, A.D., Kutulakos, K.N.: Non-rigid structure from locally-rigid motion. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2761–2768. IEEE (2010)
23. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision (ECCV). pp. 402–419. Springer (2020)

24. Wang, K., Yang, C., Lin, R., Yang, X., Shen, W.: Endosd-slam: Real-time deformable endoscopic slam via sparse-dense hybrid representation. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 4214–4219. IEEE (2025)
25. Wang, K., Yang, C., Wang, Y., Li, S., Wang, Y., Dou, Q., Yang, X., Shen, W.: Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 219–229. Springer (2024)
26. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundation-stereo: Zero-shot stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5249–5260 (2025)
27. Wu, T., Miao, Y., Li, Z., Zhao, H., Dang, K., Su, J., Yu, L., Li, H.: Endoflow-slam: Real-time endoscopic slam with flow-constrained gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 202–212. Springer (2025)
28. Xu, M., Zhao, L., Huang, S., Hao, Q.: Active slam in 3d deformable environments. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7952–7958. IEEE (2022)
29. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *Advances in Neural Information Processing Systems* **37**, 21875–21911 (2024)