This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

EndoDefLoc: Descriptor-Free Endoscopic Relocalization in Minimally Invasive Surgery

Jinjing Xu[1,3], Chongyang Jiang[1], Peng Liu[1,2], Chenyang Li[1,3], Nithya Bhasker[1,2], Shuwei Shao[4], Micha Pfeiffer[1,2], and Stefanie Speidel[1,2,3,5⋆]

[1] Department of Translational Surgical Oncology, NCT/UCC Dresden, a partnership between DKFZ, MF and UKDD, TUD, and HZDR, Germany
[2] German Cancer Research Center (DKFZ), Heidelberg, Germany
[3] Faculty of Medicine (MF) and University Hospital Carl Gustav Carus (UKDD), TU Dresden, Dresden, Germany
[4] Shandong University, Jinan, China
[5] The Centre for Tactile Internet with Human-in-the-Loop (CeTI), TU Dresden, Dresden, Germany
stefanie.speidel@nct-dresden.de

Abstract. Accurate endoscopic camera relocalization in a given map is critical in minimally invasive surgery (MIS), yet remains challenging under extreme viewpoints and ambiguous anatomical geometry. Recent feed-forward 3D reconstruction methods have emerged as feasible and efficient scene representations, but they produce dense scene maps without storing feature descriptors, rendering conventional descriptor-based localization pipelines incompatible. To enable descriptor-free relocalization on such maps, we propose **EndoDefLoc**, a framework that novelly formulates endoscope relocalization as a dense registration problem. Crucially, to improve localization accuracy and robustness under large viewpoint changes and in anatomical scenes with low discriminability, we extend this formulation with two task-driven innovations. First, a rotation-invariant feature crafting (RIFC) stage, utilizing point-pair features to enhance geometry-based feature matching across extreme viewpoint changes. Second, a geometry–appearance fusion (GAF) module integrates color cues with the aforementioned rotation-invariant geometry in a dual-branch hierarchical design to enhance feature distinctiveness. Experiments on public ex-vivo datasets demonstrate the efficacy of the proposed designs, while empirical studies further validate EndoDefLoc as a pioneering step towards descriptor-free localization for feed-forward endoscopic scene representations.

Keywords: Endoscopy localization · 2D-3D matching · Point cloud registration.

1 Introduction

Camera localization is a crucial task [24,15] in computer-assisted minimally invasive surgery, enabling recovery from tracking failures in intraoperative naviga-

⋆ Corresponding author.

tion [6], augmented reality, and simultaneous localization and mapping (SLAM) [9]. Unlike camera tracking, which estimates relative poses between adjacent frames, camera localization estimates the query pose in a pre-built 3D map, where large viewpoint changes are common after tracking failure. Conventional localization pipelines [24,19,9] detect 2D keypoints [3,17], match them to pre-computed visual descriptors stored in the map, and estimate the 6DoF camera pose. However, recent feed-forward 3D reconstruction methods [26] produce dense scene maps storing 3D coordinates and RGB rather than explicit visual descriptors, rendering descriptor-based localization incompatible and motivating descriptor-free localization on modern feed-forward maps [5,21,23].

Existing descriptor-free localization methods adopt different strategies. Methods based on rendering, e.g. GS-CPR [13], render novel views from the map and perform keypoint matching, incurring additional computation while requiring a reliable initial pose for sufficient overlap. Other methods [28,2] directly establish sparse 2D–3D correspondences without descriptors, under-utilizing the available color cues in the map. Follow-up methods [22,27] further exploit geometry and color, yet remain designed for sparse structure-from-motion (SfM) maps [18], making them less suitable for dense feed-forward scene representations. Moreover, these methods target large-scale indoor or urban environments, resulting in a substantial domain gap to MIS scenes.

Unlike sparse structure-from-motion (SfM) maps, feed-forward reconstruction preserves dense scene geometry rather than sparse visual descriptors, naturally recasting localization as a dense 3D correspondence problem. Motivated by this paradigm shift, we reformulate endoscopic localization as a dense 3D–3D registration problem. This reformulation is non-trivial because endoscopic localization in MIS scenes presents distinct challenges and opportunities beyond conventional geometric registration. Specifically, (1) localization must remain robust to the large viewpoint changes encountered after tracking failure; and (2) repetitive anatomical structures often lead to ambiguous geometry-only correspondences, while localization naturally provides complementary appearance cues from both the query image and reconstructed map for more discriminative correspondence estimation. Existing registration methods improve robustness to large transformations through data augmentation [8,11] or carefully designed attention mechanisms [16], yet lack an inherently rotation-invariant representation. Meanwhile, appearance remains largely underexplored in 3D registration methods [8,11,16,14,25], which predominantly rely on geometric representations.

In this paper, we present EndoDefLoc, a descriptor-free endoscopic localization framework for dense feed-forward scene maps. Specifically, we (i) reformulate localization as dense 3D–3D registration and introduce a localization pipeline tailored to feed-forward scene representations; (ii) propose a rotation-invariant feature crafting (RIFC) stage for robust relocalization under large viewpoint shifts; and (iii) propose a geometry–appearance fusion (GAF) module to exploit complementary color cues for more discriminative localization in anatomical scenes. Evaluation experiments and empirical studies demonstrate the efficacy of our method for localization in modern feed-forward maps of MIS scenes.

2 Method

In this section, we present the proposed localization pipeline. We first introduce our novel reformulation of the conventional endoscopy localization problem as a 3D–3D registration problem under a descriptor-free setup (Sec. 2.1). Subsequent sections detail the network design of EndoDefLoc, beginning with the rotation-invariant feature crafting stage. We then introduce the geometry–appearance fusion module, which hierarchically fuses these geometric features with color cues—information unique to the endoscopic localization task—via a dual-branch hierarchical design. Finally, Sec. 2.4 and Sec. 2.5 detail the adoption of best practice for differentiable matching and loss designs from existing registration literature for the complete method.

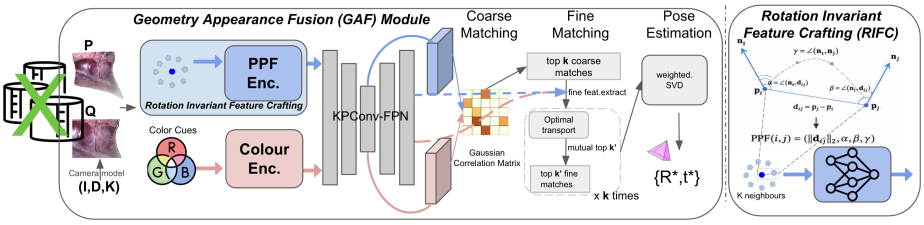


Fig. 1. Overview of our descriptor-free relocalization pipeline: EndoDefLoc. On top of a reformulation of localization as registration, we highlight the proposed task-driven innovations: (1) the rotation-invariant feature crafting (RIFC) stage and (2) the geometry-appearance fusion (GAF) backbone.

2.1 Problem Reformulation: Localization via Registration

Given a query image $I : \Omega \to \mathbb{R}^3$, a depth map $D : \Omega \to \mathbb{R}_+$, and camera intrinsics $K \in \mathbb{R}^{3\times3}$, we estimate the rigid 6-DoF transformation $T^\star$ that aligns the query observation to a pre-built 3D scene map $\mathcal{S}$. We reformulate localization as 3D–3D registration between a source point cloud Q reconstructed from the query image and a target point cloud $P \subset \mathcal{S}$ sampled from the 3D scene map. The query cloud is back-projected from valid image pixels as

$$Q = \left\{ q(u,v) = D(u,v)K^{-1}[u,v,1]^\top \mid (u,v) \in \Omega,\ D(u,v) > 0 \right\}. \tag{1}$$

Both inputs are represented as oriented point clouds with RGB cues:

$$P = \{(p_i, n_i, c_i^P)\}_{i=1}^{N_p}, \quad Q = \{(q_j, m_j, c_j^Q)\}_{j=1}^{N_q}, \tag{2}$$

where N_p and N_q denote the numbers of map and query points, $p_i, q_j \in \mathbb{R}^3$ are 3D coordinates, $n_i, m_j \in S^2$ are unit surface normals, and $c_i^P, c_j^Q \in [0,1]^3$ are RGB

color cues. Surface normals are estimated by PCA over local 3D neighborhoods after voxelization, using the finest-scale point cloud. Our objective is to recover

$$T^\star(q) = R^\star q + t^\star, \quad R^\star \in SO(3), \quad t^\star \in \mathbb{R}^3, \tag{3}$$

such that transformed query points are aligned with their corresponding map points. The proposed network learns multi-scale point descriptors for the subsequent coarse-to-fine matching stage.

2.2 Rotation-invariant Feature Crafting Stage

In the proposed rotation-invariant feature crafting (RIFC) stage, we explicitly compute the Point Pair Feature (PPF) [4] to serve as a rotation-invariant geometric representation. For an oriented reference point (p_i, n_i) and one neighbor $(p_j, n_j) \in \mathcal{N}(p_i)$, the 4D PPF is

$$\mathrm{PPF}(i,j) = \big(\|d_{ij}\|_2, \alpha, \beta, \gamma\big), \tag{4}$$

where

$$d_{ij} = p_j - p_i, \quad \alpha = \angle(n_i, d_{ij}), \quad \beta = \angle(n_j, d_{ij}), \quad \gamma = \angle(n_i, n_j). \tag{5}$$

Here, $\angle(a,b) \in [0,\pi]$ denotes the unsigned angle between two vectors. More explicitly, $d_{ij} \in \mathbb{R}^3$, $\|d_{ij}\|_2 \in \mathbb{R}_+$, and $\alpha, \beta, \gamma \in [0,\pi]$, so each point pair contributes a 4D vector. Because rigid rotations preserve Euclidean distances and pairwise angles between rotated vectors, the PPF tuple is invariant to the global camera rotation. The distance term $\|d_{ij}\|_2$ is normalized by the local search radius to ensure scale consistency. These features are computed within each local neighborhood $\mathcal{N}(p_i)$; in our setup, the RIFC input uses $K = 8$ neighbors per point. The resulting $K \times 4$ local feature tensor is processed by a two-layer MLP with Group Normalization to aggregate neighborhood information into one rotation-invariant embedding per point.

2.3 Geometry-Appearance Fusion Module

Different from the geometry feature, which requires special crafting for rotation robustness, the appearance, indicated by the color cue, is naturally less sensitive to viewpoint transformation. In the GAF module, we integrate color cues with the invariant geometric features via a dual-branch design. Specifically, the per-point RGB colors are first projected in a higher-dimensional feature space via a 2-layer MLP and fused with the geometry feature via a learnable weighted summation. The fused feature is then processed by a Group-Normalized, Feature Pyramid Network [12] style Kernel Point Convolution [20] backbone to enhance the fusion and obtain multi-scale features for coarse-to-fine matching. This fusion improves discriminability in anatomically ambiguous regions, where geometry or appearance alone may be insufficient for robust correspondence estimation.

2.4 Coarse-to-fine Matching

Following established registration frameworks [16,10], we adopt a standard hierarchical coarse-to-fine matching pipeline to instantiate the proposed localization formulation. Let $P_c = \{x_i\}_{i=1}^{N_c^P}$ and $Q_c = \{y_j\}_{j=1}^{N_c^Q}$ denote the downsampled superpoints. The GAF backbone produces multi-scale descriptors, where $h_i^P, h_j^Q \in \mathbb{R}^{d_c}$ denote the coarse-level descriptors, and $f_a^P, f_b^Q \in \mathbb{R}^{d_f}$ denote the fine-level descriptors. We first compute the coarse similarity matrix

$$S \in \mathbb{R}^{N_c^P \times N_c^Q}, \quad s_{i,j} = \exp\left(-\|h_i^P - h_j^Q\|_2^2\right), \tag{6}$$

and select the top-k entries as coarse superpoint anchors. For each anchor (x_i, y_i), let $G_{x_i}^P$ and $G_{y_i}^Q$ denote the corresponding fine-resolution point patches containing $n_i = |G_{x_i}^P|$ and $m_i = |G_{y_i}^Q|$ points, respectively. Dense correspondences are estimated by an optimal-transport layer with Sinkhorn normalization:

$$\mathbf{Z}_i = \text{Sinkhorn}\left(\exp(\mathbf{A}_i/\tau)\right), \tag{7}$$

where τ is the temperature and $\mathbf{A}_i = [a_{a,b}^i] \in \mathbb{R}^{n_i \times m_i}$ is the score matrix whose entries are computed from normalized cosine similarities,

$$a_{a,b}^i = \frac{\langle f_a^P, f_b^Q \rangle}{\|f_a^P\|_2 \|f_b^Q\|_2}. \tag{8}$$

The final point correspondences between $p_a \in G_{x_i}^P$ and $q_b \in G_{y_i}^Q$ are extracted from the mutual top-k' entries of $\mathbf{Z}_i$:

$$\mathcal{C}_i = \left\{(p_a, q_b, z_{a,b}^i) \;\middle|\; (a,b) \in \text{mutual top-}k'(\mathbf{Z}_i)\right\}. \tag{9}$$

2.5 Loss Computation and 6DoF Pose Estimation

Following the standard registration pipeline [16], we adopt the same coarse and fine matching losses for end-to-end training:

$$\mathcal{L} = \lambda_c \mathcal{L}_\text{c} + \lambda_f \mathcal{L}_\text{f}. \tag{10}$$

The coarse loss is an overlap-aware circle loss over the coarse descriptors:

$$\mathcal{L}_\text{c} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \log\left(1 + \sum_{j \in \mathcal{P}_i} e^{\alpha_{ij}^p (\delta_{ij} - \Delta_p)} \sum_{k \in \mathcal{N}_i} e^{\alpha_{ik}^n (\Delta_n - \delta_{ik})}\right), \tag{11}$$

where $\mathcal{A}$ denotes the anchor superpoints, $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the positive and negative sets associated with anchor i, and $\delta_{ij} = \|h_i^P - h_j^Q\|_2$. The fine loss supervises the Sinkhorn transport matrix using the ground-truth correspondence set $\mathcal{M}_i$:

$$\mathcal{L}_\text{f} = -\frac{1}{\sum_i |\mathcal{M}_i|} \sum_i \sum_{(a,b) \in \mathcal{M}_i} \log z_{a,b}^i. \tag{12}$$

The final 6DoF pose is estimated from the aggregated correspondence set $\mathcal{C} = \bigcup_i \mathcal{C}_i$ via weighted SVD-based Procrustes alignment:

$$(R^\star, t^\star) = \arg \min_{R \in SO(3),\, t \in \mathbb{R}^3} \sum_{(i,j) \in \mathcal{C}} w_{ij} \left\| p_i - (R q_j + t) \right\|_2^2 . \tag{13}$$

where w_{ij} denotes the matching confidence associated with correspondence (i, j).

3 Experiments

3.1 Experiment Setting

Datasets and Evaluation. We evaluate the proposed method on the ex-vivo SCARED dataset [1], a widely used benchmark. The dataset provides endoscopic stereo sequences of porcine abdominal anatomy with ground truth labels of camera intrinsics and extrinsics. We construct the query–map pointcloud pairs following the practice in [10,28]. Specifically, given metric depth annotations as computed in[7], each query frame is back-projected as the source cloud, and keyframes with 5-frame fused fragments are used as target maps. These clouds are voxelized at 1mm resolution and selected with valid overlap ratios in the range 30–90% so that the pairs have sufficient geometric correspondence while not being overly difficult. The number of pairs for train/val is 6000/900, respectively, and we prepare 1500 test samples representing large viewpoint shifts to simulate the tracking-failure scenarios where localization algorithms are typically invoked. We report localization performance using Relative Translation Error (RTE, mm), Relative Rotation Error (RRE, °), Inlier Ratio (IR, %), and Chamfer Distance (CD, mm).

Implementation Details. We implemented our method with PyTorch, and all experiments are conducted on an NVIDIA RTX-A5000 GPU. Following the setup in [16], we train our network with batch size 1 and weight decay 10^{-6}. The learning rate starts from 10^{-4} and decays exponentially by 0.05 every epoch.

3.2 Validation of the Localization Formulation

Table 1. Evaluation on SCARED dataset

	IR (%)↑	RTE (mm)↓	RRE (deg)↓	CD (mm)↓
Baseline [16]	63.7	1.97	2.73	4.27
Ours	**70.6**	**1.76**	**1.23**	**3.81**

We select GeoTransformer [16] as the baseline, as existing localization methods rely on visual descriptors stored in sparse SfM maps and are therefore not

directly comparable to our descriptor-free setting. GeoTransformer is a representative transformation-invariant 3D–3D registration framework with a KPConv backbone and hierarchical coarse-to-fine matching. As the closest generic registration solution, it provides a representative and sufficient baseline for evaluating the benefit of our localization-specific designs beyond the problem reformulation. For a fair comparison, we retrain GeoTransformer using the same training protocol. As a generic registration framework, GeoTransformer does not exploit the localization-specific properties of endoscopic scenes, while EndoDefLoc explicitly crafts rotation-invariant geometric features and exploits endoscopy-specific color cues. As shown in Tab. 1, EndoDefLoc consistently improves all metrics. These results validate the feasibility of descriptor-free localization on feed-forward scene maps and further demonstrate the benefit of the proposed localization-specific designs over a generic registration framework.

3.3 Compatibility with Feed-forward Reconstruction

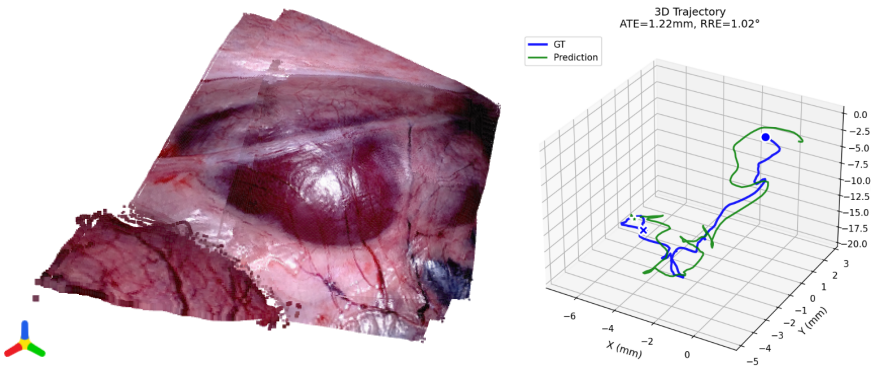


Fig. 2. Compatibility evaluation with feed-forward reconstruction. Left: Endo3R dense map for SCARED dataset_1 keyframe_1 sequence. Right: predicted camera trajectory from our descriptor-free localization on the same feed-forward scene representation. We highly recommend the supplementary video for better animation.

We also evaluate the practical applicability of our proposed method on a dense map generated by Endo3R [5], a recent endoscopy feed-forward reconstruction method. We conduct the experiment on the keyframe_1 sequence (197 frames) from the dataset_1 split of the SCARED dataset. The qualitative result in Fig.2 shows stable and consistent camera tracking, demonstrating that the proposed descriptor-free localization formulation can operate directly on modern feed-forward scene representations without requiring descriptor reconstruction.

Table 2. Architectural ablation and training augmentation on SCARED

(a) Architectural ablation

GAF	RIFC	IR↑	RTE↓	RRE↓	CD↓
✓	✓	**61.7**	**2.57**	**1.85**	**4.74**
✗	✓	55.7	3.36	2.34	6.17
✗	✗	36.6	9.25	7.16	14.74

(b) Training augmentation

	IR↑	RTE↓	RRE↓	CD↓
w/o aug.	61.7	2.57	1.85	4.74
large rot.	**70.6**	**1.76**	**1.23**	**3.81**

3.4 Ablation Study on Architectural Design

We investigate the contributions of RIFC and GAF on SCARED. As shown in Table 2(a), each component consistently improves localization performance: RIFC provides the dominant IR gain, indicating improved robustness to large rotations, while GAF adds a further IR gain, showing that appearance cues provide complementary discriminative information in anatomically ambiguous regions.

We further investigate the effect of large-rotation augmentation during training. As shown in Table 2(b), large-rotation augmentation consistently improves all evaluation metrics on the hard test set. The improvement suggests that the model does not merely memorize the augmented hard cases, but learns representations that generalize better to large rotational variations.

4 Conclusion

We presented EndoDefLoc, a descriptor-free localization framework for dense feed-forward scene maps in minimally invasive surgery. Motivated by the paradigm shift from sparse SfM maps storing visual descriptors to feed-forward scene representations preserving dense geometry and color, we reformulated endoscopic localization as a dense 3D–3D registration problem compatible with descriptor-free scene representations. To address localization-specific challenges beyond generic registration, EndoDefLoc introduces a rotation-invariant feature crafting (RIFC) stage for robust feature matching under large viewpoint changes, together with a geometry–appearance fusion (GAF) module that exploits complementary color cues to improve correspondence discriminability in anatomically ambiguous scenes. Extensive experiments on public ex-vivo dataset demonstrate the effectiveness of both the proposed formulation and the task-driven designs. We hope this work provides a foundation for future research on descriptor-free localization using modern feed-forward scene representations.

Limitations This work establishes descriptor-free localization under a registration formulation for dense feed-forward scene maps. The proposed formulation does not explicitly incorporate mechanisms for suppressing dynamic observations during localization. Future work could extend the framework with correspondence filtering or outlier rejection strategies that enable pose estimation to rely

primarily on geometrically consistent static structures in more challenging surgical scenes involving tissue deformation, surgical tool motion, and tool–tissue interactions.

Acknowledgments. This work was funded by the Bavarian and Saxon Ministries for "Next Generation AI Computing (GAIn)". It also received funding from the German Research Foundation (DFG, EXC 2050/2, Project ID 390696704, CeTI), DFG Reinhart Koselleck-project (Project ID 560101272, SIMSURGE), and the BMFTR (6G-life project, ID: 16KIS2413K).

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 preprint arXiv:2101.01133 (2021)
2. Campbell, D., Liu, L., Gould, S.: Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In: European Conference on Computer Vision. pp. 244–261. Springer (2020)
3. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
4. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 998–1005. Ieee (2010)
5. 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)
6. Hayashi, Y., Misawa, K., Oda, M., Hawkes, D.J., Mori, K.: Clinical application of a surgical navigation system based on virtual laparoscopy in laparoscopic gastrectomy for gastric cancer. International journal of computer assisted radiology and surgery **11**(5), 827–836 (2016)
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 **18**(7), 1185–1192 (2023)
8. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)
9. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.M.: Defslam: Tracking and mapping of deforming scenes from monocular sequences. IEEE Transactions on robotics **37**(1), 291–303 (2020)
10. Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y., Xu, K.: 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14128–14138 (2023)

11. Li, Y., Harada, T.: Lepard: Learning partial point cloud matching in rigid and deformable scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5554–5564 (2022)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
13. Liu, C., Chen, S., Bhalgat, Y., Hu, S., Cheng, M., Wang, Z., Prisacariu, V.A., Braud, T.: Gs-cpr: Efficient camera pose refinement via 3d gaussian splatting. arXiv preprint arXiv:2408.11085 (2024)
14. Liu, P., Golde, J., Morgenstern, J., Bodenstedt, S., Li, C., Hu, Y., Chen, Z., Koch, E., Neudert, M., Speidel, S.: Non-rigid point cloud registration for middle ear diagnostics with endoscopic optical coherence tomography. International Journal of Computer Assisted Radiology and Surgery **19**(1), 139–145 (2024)
15. Ma, R., McGill, S.K., Wang, R., Rosenman, J., Frahm, J.M., Zhang, Y., Pizer, S.: Colon10k: a benchmark for place recognition in colonoscopy. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1279–1283. IEEE (2021)
16. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9806–9821 (2023)
17. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)
18. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
19. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. IEEE Robotics and Automation Letters **3**(4), 4068–4075 (2018)
20. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019)
21. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
22. Wang, S., Kannala, J., Barath, D.: Dgc-gnn: Leveraging geometry and color cues for visual descriptor-free 2d-3d matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20881–20891 (2024)
23. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
24. Wei, R., Li, B., Mo, H., Lu, B., Long, Y., Yang, B., Dou, Q., Liu, Y., Sun, D.: Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery. IEEE Transactions on Biomedical Engineering **70**(2), 488–500 (2022)
25. Yang, Z., Simon, R., Linte, C.A.: Learning feature descriptors for pre-and intraoperative point cloud matching for laparoscopic liver registration. International journal of computer assisted radiology and surgery **18**(6), 1025–1032 (2023)

26. Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.X., Liang, H., Xu, Z., Su, H., et al.: Advances in feed-forward 3d reconstruction and view synthesis: A survey. arXiv preprint arXiv:2507.14501 (2025)
27. Zhang, Y., Wang, S., Kannala, J.: A2-gnn: Angle-annular gnn for visual descriptor-free camera relocalization. In: 2025 International Conference on 3D Vision (3DV). pp. 357–368. IEEE (2025)
28. Zhou, Q., Agostinho, S., Ošep, A., Leal-Taixé, L.: Is geometry enough for matching in visual localization? In: European Conference on Computer Vision. pp. 407–425. Springer (2022)