Adapting Generalizable Visual Geometry Models for Endoscopic 3D Reconstruction

Yanzhe Liu[1,3], Yicheng Yu[2], Junyang Wu[2,3], and Yun Gu[1,3] (✉)

[1] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
geron762@sjtu.edu.cn

Abstract. Accurate endoscopic 3D reconstruction is important for intraoperative scene understanding, navigation, and robotic assistance. Recent feed-forward visual geometry models generalize well in natural scenes, but their direct transfer to endoscopic videos is hindered by specular highlights, weak textures, tissue deformation, narrow fields of view, and imaging degradations. We formulate endoscopic 3D reconstruction as domain adaptation of a pre-trained visual geometry model and adapt VGGT through supervised depth-pose fine-tuning, perturbation-enhanced training, and teacher-student geometric consistency. Experiments on multiple public endoscopic datasets show that endoscopic fine-tuning substantially improves camera pose estimation over zero-shot transfer, while perturbation augmentation and teacher-student consistency provide complementary robustness gains. These results suggest that general visual geometry priors can effectively support endoscopic reconstruction when combined with domain-specific adaptation.

Keywords: 3D reconstruction · Visual geometry model · Depth estimation · Endoscopic surgery · Camera pose estimation

1 Introduction

Endoscopic intervention can be viewed as a sequential process of local visual exploration. Although each frame provides only a narrow and two-dimensional observation that depends on the viewing perspective, many computer-assisted surgical applications require a temporally consistent and spatially organized understanding of the anatomical scene. Recent studies on real-time endoscopic reconstruction and geometry-aware surgical perception suggest that recovering spatial structure from endoscopic videos can provide a common representation for scene understanding, navigation, coverage analysis, and robot-assisted intervention [1–5]. Therefore, estimating three-dimensional structure and camera motion from endoscopic videos is not merely a geometric vision task, but a key step toward intraoperative spatial understanding and future blind-spot-aware endoscopic systems.

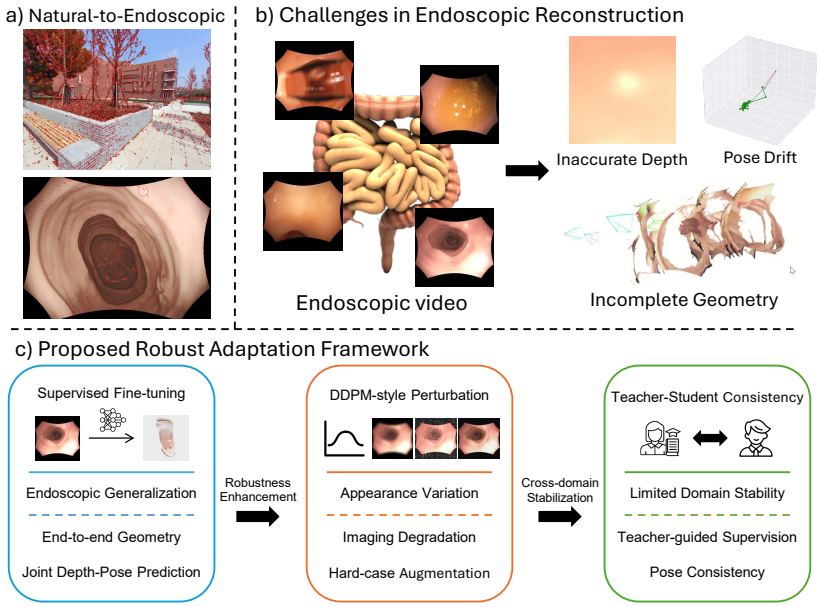


Fig. 1: **Motivation and framework overview.** (a) SIFT comparison reveals the endoscopic domain gap. (b) Imaging artifacts and tissue deformation hinder reliable reconstruction. (c) VGGT is adapted through supervised fine-tuning, DDPM-style perturbations, and teacher–student consistency.

However, endoscopic 3D reconstruction remains considerably more challenging than reconstruction in natural scenes. Endoscopic videos often contain weak or repetitive textures, specular highlights, non-uniform illumination, motion blur, liquid artifacts, narrow fields of view, and non-rigid tissue deformation. These factors make stable feature matching, rigid-scene assumptions, and photometric consistency unreliable, which limits the performance of traditional SfM-, SLAM-, and multi-view stereo-based pipelines [6–10]. Learning-based monocular depth and ego-motion methods have improved robustness in endoscopic scenarios [11–14], but many of them still rely on task-specific assumptions, carefully designed self-supervised losses, or additional priors that may be fragile under severe appearance changes and tissue deformation.

In parallel, general 3D vision has recently shifted from multi-stage geometric optimization toward feed-forward visual geometry models. Methods such as DUSt3R [15], MASt3R [16], MonST3R [17], VGGT [18], and VGGT-Ω [19] show that dense geometry, camera parameters, point correspondences, and 3D structure can be inferred directly from images using large-scale learned priors, reducing the dependence on explicit feature matching, bundle adjustment, and test-time optimization. This trend has also extended to surgical video reconstruction, as demonstrated by Endo3R, which predicts scale-consistent depth, camera parameters, and globally aligned geometry from monocular endoscopic videos [5].

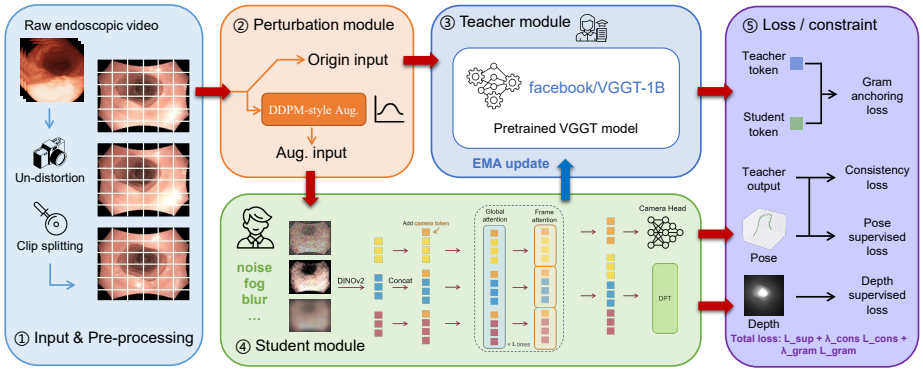


Fig. 2: **Our framework**. The proposed framework adapts a pre-trained VGGT model to endoscopic 3D reconstruction through preprocessing, perturbation augmentation, teacher-student geometric consistency, and supervised depth-pose constraints.

Motivated by these advances, we formulate endoscopic 3D reconstruction as the domain adaptation of a pre-trained visual geometry model. Despite their strong geometric priors, such models face substantial domain shifts caused by weak textures, specular reflections, tissue deformation, narrow fields of view, and limited annotations. As shown in Fig. 1, we adapt VGGT [18] through supervised depth-pose fine-tuning, diffusion-inspired perturbation augmentation, and teacher-student consistency to improve appearance robustness and stabilize pose-related representations. This setting also provides a meaningful testbed for evaluating the transferability and limitations of visual geometry foundation models in medical scenarios.

We evaluate our framework on four public endoscopic datasets, including C3VD [20], C3VDv2 [21], SimColon [2], and SCARED [22]. These datasets cover diverse imaging conditions, anatomical structures, motion patterns, and annotation settings. Our experiments show that direct zero-shot transfer from the pre-trained model is insufficient, while endoscopic fine-tuning significantly improves camera pose estimation. Perturbation-enhanced training and teacher-student consistency further improve robustness under different dataset conditions, suggesting that endoscopic reconstruction benefits from both general geometric priors and domain-specific adaptation.

Our contributions are summarized as follows: 1) We investigate endoscopic 3D reconstruction as a domain adaptation problem for feed-forward visual geometry models, highlighting their potential in challenging medical scenes. 2) We propose an adaptation framework that integrates supervised fine-tuning, perturbation-enhanced training, and teacher-student geometric consistency for robust endoscopic depth-pose estimation. 3) We demonstrate on multiple public endoscopic datasets that general visual geometry priors can substantially improve endoscopic reconstruction after domain adaptation, while highlighting the need for stronger multi-domain generalization in more challenging endoscopic scenarios.

2 Methodology

We adapt a pre-trained feed-forward visual geometry model to endoscopic 3D reconstruction. Given an endoscopic image sequence $I_{1:T} = \{I_i\}_{i=1}^{T}$, the model predicts the corresponding depth maps $\hat{D}_{1:T}$ and camera poses $\hat{P}_{1:T}$. For supervised training, each sample provides depth and camera pose annotations, denoted as $D_{1:T}$ and $P_{1:T}$. Following the training formulation of VGGT [18], the basic training objective is

$$\mathcal{L}_{\text{sup}} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}(\hat{D}_{1:T}, D_{1:T}) + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}}(\hat{P}_{1:T}, P_{1:T}). \tag{1}$$

As shown in Fig. 2, the proposed framework contains five main components: input preprocessing, perturbation augmentation, a clean-input teacher branch, an augmented-input student branch, and a loss constraint module. The raw endoscopic video is first undistorted and split into short clips. The perturbation module keeps the original sequence for the teacher branch and generates a degraded sequence for the student branch. The student model is supervised by depth and pose annotations, while the teacher model provides stop-gradient references for representation-level consistency. This design allows the model to learn from endoscopic appearance degradation while preserving stable geometric representations.

2.1 Perturbation-enhanced Endoscopic Adaptation

Endoscopic images often contain blur, illumination variation, haze-like degradation, specular disturbance, and other appearance artifacts. Although these artifacts change the image appearance, they should not change the underlying 3D structure or camera motion. Based on this geometric label invariance, we apply perturbation augmentation only to the input images while keeping the original depth and pose labels unchanged.

Formally, the augmented sequence is generated by

$$\tilde{I}_{1:T} = \mathcal{A}_{s,\alpha}(I_{1:T}), \tag{2}$$

where $\mathcal{A}_{s,\alpha}$ denotes an augmentation function controlled by the degradation style s and perturbation parameter α. The student model then predicts depth and pose from the augmented sequence:

$$(\hat{D}_{1:T}, \hat{P}_{1:T}) = f_\phi(\tilde{I}_{1:T}). \tag{3}$$

Since the augmentation only modifies image appearance, Eq. 1 can still be used with the original geometric annotations.

In practice, the augmentation module includes diffusion-inspired appearance degradation [23] and corruptions such as blur, noise, and fog-like degradation. By training on both clean and degraded inputs, the model is encouraged to rely less on local texture and brightness patterns, and more on temporally consistent structural cues and global geometric relationships.

2.2 Teacher-student Geometric Consistency

To stabilize the geometric representation, we introduce a teacher-student consistency framework. The teacher branch receives the clean sequence, while the student branch receives the augmented sequence:

$$(\hat{D}^T, \hat{P}^T, E^T, F^T) = f_{\theta_T}(I_{1:T}), \tag{4}$$

$$(\hat{D}^S, \hat{P}^S, E^S, F^S) = f_{\theta_S}(\tilde{I}_{1:T}), \tag{5}$$

where E denotes the pose-related encoding and F denotes intermediate geometric tokens. The teacher outputs are treated as stop-gradient references.

Instead of directly constraining the decoded rotation and translation, we impose consistency on the pose encoding space:

$$\mathcal{L}_{\text{cons}} = d_{\text{cons}}\left(E^S, \text{sg}(E^T)\right), \tag{6}$$

where $d_{\text{cons}}(\cdot)$ is a feature distance and $\text{sg}(\cdot)$ denotes the stop-gradient operation. This representation-level constraint allows the student model to maintain stable pose-related geometry under degraded observations without over-constraining the final camera parameters.

Following DINOv3 [24], we apply Gram Anchoring to preserve the correlation structure of intermediate geometric tokens. For a token feature $F_l \in \mathbb{R}^{C_l \times N_l}$, its Gram matrix is

$$G(F_l) = \frac{1}{C_l N_l} F_l F_l^\top. \tag{7}$$

The Gram Anchoring loss is defined as

$$\mathcal{L}_{\text{gram}} = \frac{1}{|\mathcal{L}_{\text{ga}}|} \sum_{l \in \mathcal{L}_{\text{ga}}} \left\| G(F_l^S) - \text{sg}\left(G(F_l^T)\right) \right\|_F^2, \tag{8}$$

where $\mathcal{L}_{\text{ga}}$ denotes the selected token layers. Unlike element-wise feature matching, Gram Anchoring aligns the feature correlation structure between teacher and student branches, allowing local feature adaptation while preserving global geometric statistics.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{gram}}\mathcal{L}_{\text{gram}}. \tag{9}$$

3 Experiments

3.1 Implementation Details

We initialize the model from pre-trained VGGT [18]. All input images are resized to 224×224 with a patch size of 14. For teacher-student supervision, the teacher branch is updated by EMA with a decay rate of 0.999 at each training step. The pose encoding consistency loss is computed on `pose_enc` using MSE with

Table 1: Comparison with representative monocular depth-and-motion baselines.

Camera pose estimation: APE RMSE ↓.

Method	C3VD	C3VDv2	SimColon	SCARED
Ours	**0.4198**	**0.6128**	**0.2341**	**0.8054**
AF-SfMLearner [12]	2.2008	3.4627	1.0139	2.1374
Monodepth2 [25]	2.2891	2.8098	0.7731	1.9431
EndoDAC [14]	1.2769	2.8617	0.8995	1.6060

(b) Depth estimation: AbsRel ↓ / $\delta_{1.25}$ ↑.

Method	C3VD	C3VDv2	SimColon	SCARED
Ours	**0.1000** / **0.9320**	**0.0951** / **0.9329**	**0.0301** / **0.9954**	0.0633 / 0.9576
AF-SfMLearner [12]	0.9041 / 0.2248	0.7209 / 0.2485	0.1334 / 0.9303	**0.0362** / **0.9954**
Monodepth2 [25]	0.6353 / 0.3117	0.7303 / 0.2670	0.1331 / 0.8996	0.0826 / 0.9398
EndoDAC [14]	5.1775 / 0.1184	5.5017 / 0.1210	0.1823 / 0.7877	0.0932 / 0.9230

a weight of 0.075. Gram Anchoring is applied to intermediate geometric tokens from layers 0, 7, 15, and 23, with a loss weight of 0.05.

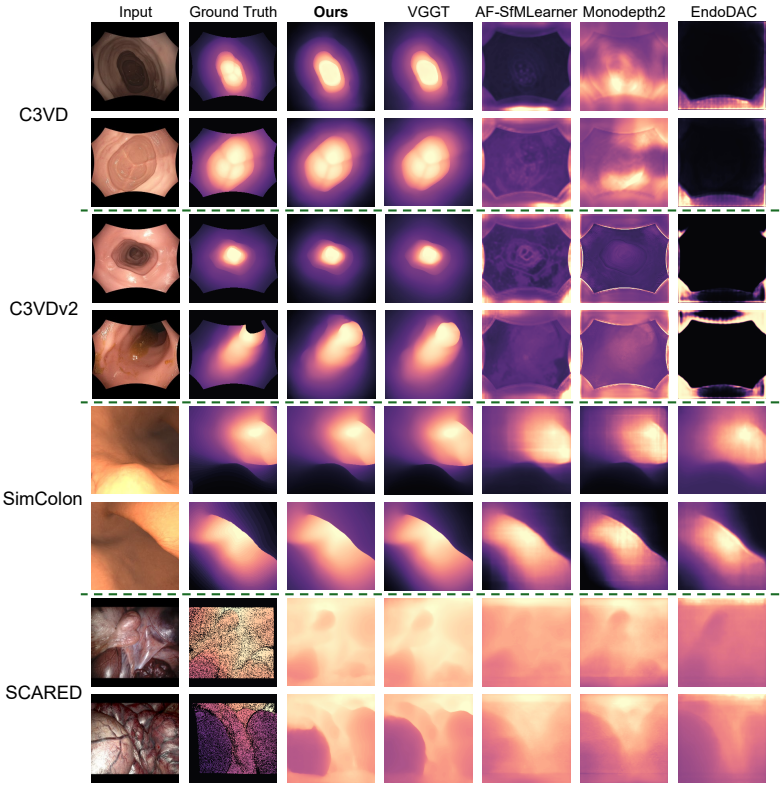
Datasets. We train the adaptation framework using five public endoscopic datasets, including C3VD [20], C3VDv2 [21], SimColon [2], SCARED [22], and EndoDepth [26]. Depth maps, camera intrinsics, and camera poses are used as geometric supervision when available. Unless otherwise specified, quantitative results are reported on C3VD, C3VDv2, SimColon, and SCARED.

Evaluation Metrics. For camera pose estimation, we report Absolute Pose Error Root Mean Square Error (APE RMSE) following standard trajectory evaluation protocols [27], where lower values indicate more accurate trajectories. For depth estimation, we report Absolute Relative Error (AbsRel) and threshold accuracy $\delta_{1.25}$, where lower AbsRel and higher $\delta_{1.25}$ indicate better depth prediction.

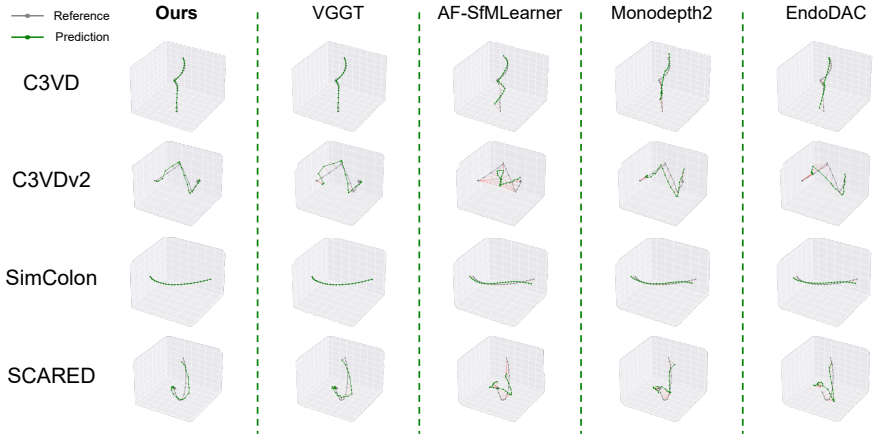
3.2 Experimental Results

Comparison with Existing Methods. Table 1 compares our model with representative monocular depth methods, including AF-SfMLearner [12], Monodepth2 [25], and EndoDAC [14]. Our method consistently achieves the lowest APE RMSE across all four datasets, showing that adapting a feed-forward visual geometry model provides stronger camera pose estimation than conventional monocular self-supervised pipelines. For depth estimation, our method obtains the best AbsRel on C3VD, C3VDv2, and SimColon, while maintaining competitive performance on SCARED.

Qualitative Visualization. Fig. 3 shows qualitative comparisons of depth maps and camera trajectories on C3VD, C3VDv2, SimColon, and SCARED. Our method produces depth maps with more consistent global structure and fewer artifacts than the compared monocular methods. In pose estimation, our predicted trajectories are generally better aligned with the reference trajectories, showing more stable camera motion estimation across different endoscopic datasets. These visual results support the quantitative improvements reported above.



(a) Depth visualization results.



(b) Pose visualization results.

Fig. 3: Qualitative comparison of endoscopic 3D reconstruction results. VGGT denotes the endoscopically fine-tuned VGGT model.

Table 2: Ablation study of different adaptation components.

(a) Camera pose estimation: APE RMSE ↓.

Perturbation	T-S Consistency	C3VD	C3VDv2	SimColon	SCARED
✓	✓	0.4198	0.6128	**0.2341**	**0.8054**
✓	×	**0.4176**	**0.5990**	0.2936	0.8388
×	×	0.4504	0.6731	0.3155	0.8444

(b) Depth estimation: AbsRel ↓ / $\delta_{1.25}$ ↑.

Perturbation	T-S Consistency	C3VD	C3VDv2	SimColon	SCARED
✓	✓	**0.1000** / **0.9320**	**0.0951** / **0.9329**	**0.0301** / **0.9954**	0.0633 / 0.9576
✓	×	0.1012 / 0.9294	0.0972 / 0.9276	0.0304 / 0.9953	0.0636 / 0.9682
×	×	0.1030 / 0.9304	0.0977 / 0.9261	0.0319 / 0.9949	**0.0580** / **0.9690**

Ablation Study. Table 2 evaluates the proposed adaptation components, with supervised endoscopic fine-tuning used in all variants. Perturbation enhancement improves pose estimation across all datasets and depth estimation on three of the four datasets. Teacher-student consistency further improves pose accuracy on SimColon and SCARED and depth accuracy on C3VD, C3VDv2, and SimColon, while the slight regressions on the remaining metrics indicate dataset-dependent trade-offs.

4 Conclusion

We presented a framework for adapting a general visual geometry model to endoscopic 3D reconstruction through supervised fine-tuning, perturbation-enhanced training, and teacher-student geometric consistency. Experiments on multiple public endoscopic datasets demonstrate that domain-specific fine-tuning substantially improves camera pose estimation over zero-shot transfer, indicating the feasibility of adapting pre-trained 3D visual geometry models to challenging endoscopic scenes with limited training data.

Future work will focus on stronger cross-dataset generalization, more comprehensive depth evaluation, uncertainty-aware reconstruction, and temporally consistent online modeling. The adapted framework may further support blind-spot detection, active robotic re-scanning, and robot-assisted endoscopic navigation.

Acknowledgments. This work is supported by National Key R&D Program of China (2022ZD0212400), Science and Technology Commission of Shanghai Municipality (No.25511102602), Natural Science Foundation of China (62373243), Open Project of the Shanghai Key Laboratory of Flexible Medical Robotics (Grant No. SKLFMR-0211), and Shanghai Tongren Hospital Medical-Engineering Collaboration Project (lhyjzx2024-xm03).

Disclosure of Interests. The authors declare no competing interests.

References

1. Richter, A., Steinmann, T., Rosenthal, J.-C., Rupitsch, S.J.: Advances in real-time 3D reconstruction for medical endoscopy. Journal of Imaging **10**(5), 120 (2024)
2. Rau, A., Bano, S., Jin, Y., Azagra, P., Morlana, J., Kader, R., Sanderson, E., Matuszewski, B.J., Lee, J.Y., Lee, D.J., et al.: SimCol3D—3D reconstruction during colonoscopy challenge. Medical Image Analysis **96**, 103195 (2024)
3. Mangulabnan, J.E., Soberanis-Mukul, R.D., Teufel, T., Sahu, M., Porras, J.L., Vedula, S.S., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: An endoscopic chisel: Intraoperative imaging carves 3D anatomical models. International Journal of Computer Assisted Radiology and Surgery **19**(7), 1359–1366 (2024)
4. Hardy, R., Berzin, T.M., Rajpurkar, P.: ColonCrafter: A depth estimation model for colonoscopy videos using diffusion priors. arXiv preprint arXiv:2509.13525 (2025)
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 (MICCAI). Springer (2025)
6. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104–4113 (2016)
7. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
8. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)
9. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: Depth inference for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV), pp. 767–783. Springer (2018)
10. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2495–2504 (2020)
11. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis **71**, 102058 (2021)
12. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical Image Analysis **77**, 102338 (2022)
13. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J.M.M., Civera, J.: Endo-Depth-and-Motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021)
14. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: EndoDAC: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, pp. 208–218. Springer (2024)
15. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20697–20709 (2024)

16. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: European Conference on Computer Vision (ECCV), pp. 71–91. Springer (2024)
17. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.-H.: MonST3R: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
18. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual Geometry Grounded Transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5294–5306 (2025)
19. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J.L., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: VGGT-Ω. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21486–21499 (2026)
20. 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)
21. Golhar, M.V., Galeano Fretes, L.S., Ayers, L., Akshintala, V.S., Bobrow, T.L., Durr, N.J.: C3VDv2—Colonoscopy 3D video dataset with enhanced realism. Scientific Data (2026). doi:10.1038/s41597-026-07471-1
22. 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)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851 (2020)
24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
25. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3828–3838 (2019)
26. Reyes-Amezcua, I., Espinosa, R., Daul, C., Ochoa-Ruiz, G., Méndez-Vázquez, A.: EndoDepth: A benchmark for assessing robustness in endoscopic depth prediction. arXiv preprint arXiv:2409.19930 (2024)
27. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 573–580 (2012)