ColScale3R: Parameter-Efficient Native-Scale Adaptation of Feed-Forward 3D Reconstruction for Colonoscopy

Saad Khalil and Youngbae Hwang*

Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju 28644, Republic of Korea
ybhwang@cbnu.ac.kr

Abstract. Feed-forward 3D foundation models jointly predict dense geometry and camera parameters, but their direct transfer to colonoscopy is hindered by domain-specific appearance, non-rigid tissue motion, and close-range imaging. We introduce ColScale3R, a parameter-efficient adaptation of MapAnything for native-scale reconstruction from calibrated three-view colonoscopy windows. Weight-Decomposed Low-Rank Adaptation (DoRA) with rank-stabilized LoRA (rsLoRA) scaling is applied to 72 image-encoder projections, while all remaining components are frozen. This leaves only 5.886M parameters (0.477% of the network) trainable. A metric-preserving objective combines reference-view depth and point-map supervision, metric-factor re-anchoring, and an adapter-disabled structural anchor that preserves pretrained ray and camera-pose predictions. On C3VD, ColScale3R achieves a mean three-view depth RMSE of 3.263 mm and a native fused Chamfer distance of 2.172 mm, compared with 8.580 mm and 6.069 mm, respectively, for a C3VD-trained Endo3R checkpoint using a single scale factor estimated only from the training split. It also outperforms zero-shot MapAnything and DA3 under the same evaluation protocol. Without target-domain training or test-time median scaling, ColScale3R reduces one-view SimCol depth RMSE from 155.94 cm to 1.12 cm. On the physical subset of PolypSense3D, the frozen model achieves a supplied-mask sizing MAE of 4.02 mm. Virtual-only Geometry-Preserving Scale Calibration and Aggregation (GSCA), applied solely to downstream diameter estimates, reduces the MAE to 2.90 mm. These results demonstrate parameter-efficient native-scale colonoscopic reconstruction with cross-dataset transfer and quantitative sizing capability.

Keywords: Colonoscopy · Feed-forward 3D reconstruction · Native metric scale · Parameter-efficient adaptation · Multi-view geometry

1 Introduction

Colonoscopy is central to colorectal cancer screening, diagnosis, and surveillance [1,19]. Although colonoscopy videos provide detailed mucosal appearance,

* Corresponding author

applications such as lesion measurement, mucosal coverage assessment, and retrospective spatial inspection require three-dimensional geometric information. In particular, the apparent size of a polyp varies with camera distance and viewpoint even when its physical dimensions remain unchanged [3]. Recovering geometry in native physical units is therefore essential for quantitative measurement.

Colonoscopic reconstruction is challenging because mucosal surfaces are weakly textured and locally repetitive, while the co-located camera and illumination source produce strong view-dependent reflections. Motion blur, fluid, abrupt depth transitions, narrow camera baselines, and non-rigid tissue deformation further violate the photometric and rigid-scene assumptions of conventional structure-from-motion and simultaneous localization and mapping. Consequently, correspondence estimation, camera tracking, and long-sequence fusion remain unreliable [18]. Learning-based methods have improved endoscopic depth and motion estimation through self-supervised view synthesis, recurrent reconstruction, and domain-specific adaptation [4,16,18]. Nevertheless, many approaches remain ambiguous in physical scale: monocular depth is often evaluated after per-image median scaling, while trajectories and surfaces are assessed after similarity alignment. Such protocols measure relative geometry but remove the scale error that determines whether predictions can be fused directly or used for physical measurement. We therefore define *native-scale* prediction as depth, camera translation, and reconstructed geometry evaluated directly in the calibrated physical coordinate system, without ground-truth-derived scaling, test-time scale fitting, or similarity alignment.

Recent feed-forward 3D models, including DUSt3R [22], VGGT [21], and DA3 [10], jointly infer dense geometry and camera parameters from multiple images in a single forward pass. MapAnything [8] further factors its outputs into ray depth, relative camera poses, and a shared metric factor, making it suitable for native-scale adaptation. However, its general-scene prior transfers poorly to the distinct appearance, motion, and close-range geometry of colonoscopy. Zero-shot inference may preserve relative structure after alignment while producing large native-scale errors in depth and camera translation.

We present ColScale3R, a parameter-efficient MapAnything adaptation for calibrated short-window colonoscopy reconstruction. DoRA with rank-stabilized LoRA scaling is applied only to selected image-encoder projections, while a metric-preserving objective retains pretrained geometric structure. For downstream sizing, virtual-only GSCA calibrates view-level diameter estimates without altering the reconstruction. Our contributions are:

- We introduce ColScale3R, a parameter-efficient adaptation of MapAnything for native-scale short-window colonoscopic reconstruction, optimizing only 0.477% of the network while keeping the multi-view transformer, dense decoder, and geometric heads frozen.
- We propose a metric-preserving objective that combines reference-view native geometry supervision, metric-factor re-anchoring, and an adapter-disabled structural anchor for preserving pretrained ray and camera-pose predictions.

– We evaluate ColScale3R across three benchmark datasets: native-scale depth, camera pose, and local 3D reconstruction on C3VD; zero-shot native-depth transfer on SimCol; and supplied-mask physical polyp sizing on PolypSense3D with virtual-only GSCA.

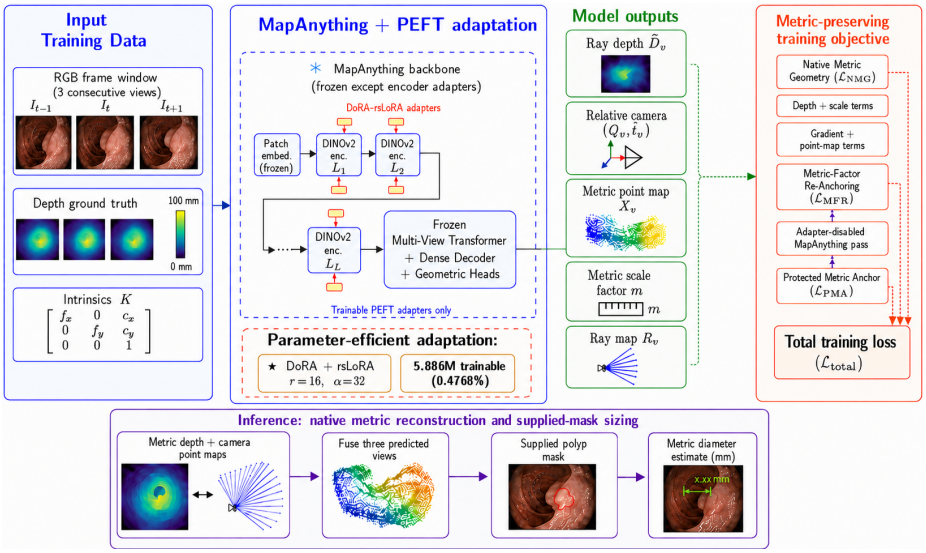

Fig. 1. Overview of ColScale3R. DoRA–rsLoRA adapters modify selected projections in the DINOv2 image encoder, while the multi-view transformer, dense decoder, and geometric prediction heads remain frozen. Training combines reference-view native metric geometry supervision, metric-factor re-anchoring, and an adapter-disabled structural anchor for ray and camera pose preservation. At inference, only the adapted branch is retained to predict native metric point maps for local fusion and polyp sizing.

2 Related Work

Endoscopic Geometry and Metric Scale. Endoscopic reconstruction has been explored through self-supervised depth and ego-motion estimation, efficient foundation-model adaptation, and recurrent or online reconstruction [16,4,20,5]. Stereo capsule endoscopy has demonstrated the clinical feasibility of hardware-enabled 3D reconstruction and lesion-size measurement [14]. AI-based removal of poorly visualized capsule frames has also reduced reading time while preserving diagnostic concordance [15], highlighting the importance of robustness to degraded endoscopic observations. Camera-intrinsics-guided learning has further enabled native-scale monocular colonoscopy reconstruction [13]. Unlike recurrent sequence-level approaches, we reconstruct calibrated short windows and

directly evaluate depth, camera translation, and fused geometry in physical units. PolypSense3D [11] provides metric geometry and lesion measurements; we use its physical-subset masks to isolate geometric sizing errors.

Feed-Forward 3D Foundation Models and Efficient Adaptation. DUSt3R [22], VGGT [21], DA3 [10], and MapAnything [8] jointly estimate multi-view geometry and camera parameters. MapAnything factorizes scene geometry into ray depth, relative camera poses, and a shared metric-scale factor. LoRA [6], DoRA [12], and rsLoRA [7] enable parameter-efficient adaptation of large pretrained models. Prior studies [4,9] adapted Depth Anything models for self-supervised monocular depth and ego-motion estimation in endoscopy. In contrast, we adapt a unified feed-forward multi-view reconstruction model to recover native metric scale while preserving its pretrained geometric priors.

3 Method

Figure 1 presents the architecture of ColScale3R. During training, an adapted MapAnything branch processes a calibrated three-view window. An adapter-disabled pass provides frozen geometric targets.

3.1 Factored Native-Scale Reconstruction

Given $\mathcal{W} = \{(I_v, K_v)\}_{v=1}^{V}$, MapAnything predicts a unit ray map R_v, up-to-scale ray depth $\widetilde{D}_v$, relative camera pose $(q_v, \widetilde{t}_v)$ and a shared metric factor m. The metric point map in the first-view coordinate frame is

$$X_v(\mathbf{u}) = m\Big[\mathcal{R}(q_v)\big(R_v(\mathbf{u})\widetilde{D}_v(\mathbf{u})\big) + \widetilde{t}_v\Big], \tag{1}$$

where $\mathcal{R}(q_v)$ is the rotation matrix corresponding to the unit quaternion q_v. We train with three consecutive calibrated frames. Dense metric supervision is applied only to the reference output, while all three depth and camera-pose predictions are evaluated and fused at inference.

3.2 Parameter-Efficient Adaptation

We apply DoRA [12] with rank-stabilized LoRA scaling [7] to 72 attention and MLP projections in the image encoder. For a frozen weight W_0,

$$\widetilde{W} = W_0 + \frac{\alpha}{\sqrt{r}}BA, \qquad W_{\text{eff}} = \operatorname{diag}(\gamma)\frac{\widetilde{W}}{\|\widetilde{W}\|_{2,\text{row}} + \epsilon}, \tag{2}$$

where A and B are low-rank factors and γ contains the trainable DoRA magnitude. We use $r = 16$, $\alpha = 32$, and dropout 0.05. The patch embedding, multi-view transformer, dense decoder, and geometric heads remain frozen, leaving 5.886M trainable parameters (0.477% of the network).

3.3 Metric-Preserving Objective

Let $\widehat{D}, D^*$ and $\widehat{X}, X^*$ denote the predicted and ground-truth reference-view metric depths, and point maps. Native Metric Geometry supervision is

$$\mathcal{L}_{\mathrm{NMG}} = \mathcal{L}_{\log D} + 0.5\mathcal{L}_{\mathrm{scale}} + 0.05\mathcal{L}_{\mathrm{grad}} + 0.05\mathcal{L}_{\mathrm{point}}, \tag{3}$$

combining robust log-depth, median log-scale, depth-gradient, and metric point-map errors. The median term is used only during training; no median scaling was applied during inference.

An adapter-disabled pass provides detached predictions $(m^0, D^0, R^0, \widetilde{t}^0, q^0)$. For valid reference pixels Ω, the metric factor is re-anchored using

$$m^* = m^0 \,\mathrm{med}_{\mathbf{u}\in\Omega}\left[\frac{D^*(\mathbf{u})+\epsilon}{D^0(\mathbf{u})+\epsilon}\right], \qquad \mathcal{L}_{\mathrm{MFR}} = \left|\log \widehat{m} - \log m^*\right|. \tag{4}$$

To preserve the pretrained multi-view geometry, the Protected Metric Anchor constrains adapted rays, translations, and quaternions to the adapter-disabled predictions:

$$\mathcal{L}_{\mathrm{PMA}} = 0.10\mathcal{L}_R + 0.50\mathcal{L}_t + 0.25\mathcal{L}_q, \tag{5}$$

where $\mathcal{L}_R$, $\mathcal{L}_t$, and $\mathcal{L}_q$ are, respectively, cosine ray, SmoothL1 translation, and sign-invariant quaternion losses. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{NMG}} + 0.25\mathcal{L}_{\mathrm{MFR}} + 0.10\mathcal{L}_{\mathrm{PMA}}. \tag{6}$$

No ground-truth camera pose loss was used. At inference, the anchor pass is discarded, and valid point maps are fused in the first-view frame without test-time scale recovery or geometric optimization.

4 Experiments

Datasets and evaluation protocol. C3VD [2] provides colonoscopy videos registered to physical phantom geometry, with metric depth, calibrated intrinsics, and camera poses. From 6491 target–source pairs, eligible three-frame windows are formed using the target, source, and next frame in the same temporal direction. We evaluate on 1761 matched test windows at 476×588 resolution and valid depths of 0.05–10 cm. All methods use identical frames, masks, units, and point sampling. MapAnything and DA3 are evaluated zero-shot. The released Endo3R checkpoint [5] includes C3VD training; because its scale is arbitrary, a single scalar estimated only on the C3VD training split is applied to its depth and camera translation. SimCol [17] is used for one-view zero-shot transfer. Physical sizing uses 15 PolypSense3D [11] phantom clips (13 unique polyps; three views per clip). The largest supplied-mask contour in each view is fitted by a minimum-area rectangle and converted to millimeters using predicted metric depth and calibrated intrinsics; view estimates are combined by a weighted median. GSCA is a 13–32–16–2 MLP trained only on virtual cases to predict diameter correction and uncertainty without modifying the reconstruction.

Table 1. Matched three-view native-scale evaluation on C3VD ($N = 1761$ windows). Depth metrics are averaged over the three outputs. RMSE, MAE, translation, and Chamfer distance (CD) are in millimeters. Endo3R† uses a fixed scale factor estimated only on the C3VD training split; no test-set scale alignment is used. Aligned CD is reported only as a relative-shape diagnostic.

Method	AbsRel↓	RMSE↓	MAE↓	$\lvert \log s \rvert$ ↓	Rot.° ↓	Trans.↓	Fused CD↓	Align. CD↓
MapAnything ZS [8]	41.040	1073.132	1034.879	3.406	0.336	11.200	1004.611	4.606
DA3 ZS [10]	15.995	421.024	419.963	2.764	**0.185**	7.805	396.705	6.783
Endo3R† [5]	0.288	8.580	7.525	0.263	3.221	2.576	6.069	2.135
ColScale3R	**0.092**	**3.263**	**2.628**	**0.074**	**0.185**	**0.276**	**2.172**	**1.218**

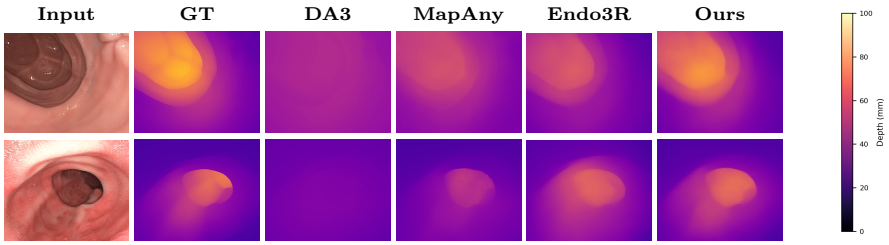

Fig. 2. Qualitative depth comparison of the C3VD test set. Baselines are median-scaled only for visualization; ground truth and ColScale3R use the native 0–100 mm scale.

Implementation details. The adapters are optimized with AdamW for 9600 steps using a learning rate of 10^{-5}, weight decay of 10^{-4}, effective batch size of 4, and bfloat16 precision. No monocular teacher, ground-truth pose supervision, target-domain scale fitting, or test-time median scaling is used.

Evaluation metrics. Depth is evaluated using AbsRel, RMSE, MAE, and absolute log-scale error, while camera pose uses geodesic rotation and translation errors. Symmetric Chamfer distance is computed after sampling every eight pixels, transforming points to the first-view frame, and voxel-downsampling the fused cloud at 1 mm. Similarity-aligned Chamfer is reported only as a relative-shape diagnostic.

5 Results

Native Metric Depth, Camera Pose, and 3D Reconstruction. Table 1 shows large native-scale errors for zero-shot MapAnything and DA3, whose depth RMSEs are 1073.1 and 421.0 mm, respectively. With the fixed training-split scale calibration described above, the C3VD-trained Endo3R checkpoint obtains a depth RMSE of 8.580 mm and fused Chamfer distance of 6.069 mm. In contrast, ColScale3R achieves a depth RMSE of 3.263 mm, MAE of 2.628 mm, fused Chamfer distance of 2.172 mm, and aligned Chamfer distance of 1.218 mm. Despite using a frozen camera-pose head without pose supervision, ColScale3R

Table 2. Zero-shot SimCol native-depth transfer and supplied-mask sizing on the physical PolypSense3D subset. Our model uses no target-domain fine-tuning or test-time scale fitting. GSCA is trained only on virtual PolypSense3D cases.

Dataset / task	Method	Views	Metric	Result
SimCol depth	MapAnything ZS	1	RMSE (cm)	155.938
SimCol depth	**ColScale3R**	1	RMSE (cm)	**1.123**
PolypSense3D sizing	Frozen ColScale3R	3	MAE (mm)	4.020
PolypSense3D sizing	**ColScale3R + GSCA**	3	MAE (mm)	**2.897**

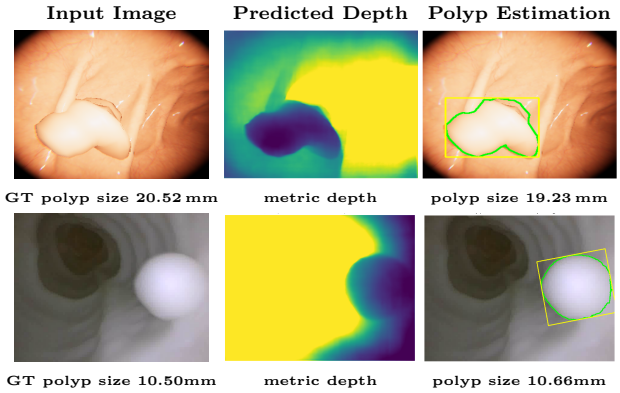

Fig. 3. Qualitative polyp size estimation on virtual (top) and physical (bottom) examples from PolypSense3D. Native metric depth predictions (center) enable local surface reconstruction and physical polyp sizing (right).

obtains rotation and translation errors of 0.185° and 0.276 mm, compared with 3.221° and 2.576 mm for Endo3R. Fusion reduces Chamfer distance from 2.218 to 2.172 mm, improving 66.1% of windows with a paired mean gain of 0.046 mm. Because C3VD inter-frame translations are mostly sub-millimeter, these results indicate local pose preservation rather than long-horizon navigation. Figure 2 and Figure4 provide qualitative depth and reconstruction comparisons on a matched C3VD test-set window.

Input-view cardinality ablation. Using the same three-view-trained checkpoint, one-, three-, and five-view inference produces depth RMSEs of 7.985, 3.263, and 3.581 mm and fused Chamfer distances of 5.365, 2.172, and 2.414 mm, respectively. Three views provide the best operating point and the one and five-view results represent inference-time cardinality sensitivity rather than separately trained models.

Cross-Dataset Transfer and Polyp-Size Estimation. Table 2 shows that the C3VD-trained ColScale3R reduces zero-shot one-view SimCol RMSE from 155.938 to 1.123 cm without target-domain training or scale fitting. On physical

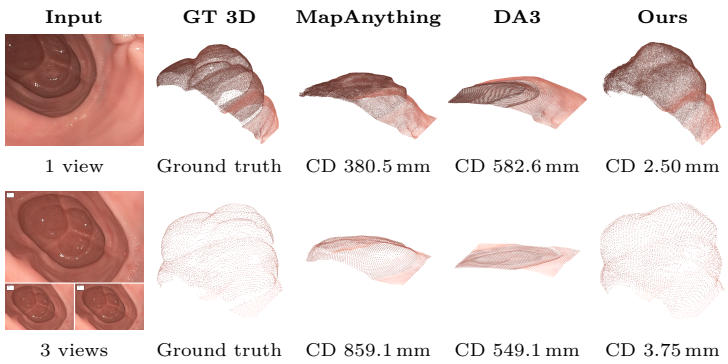

Fig. 4. 1-view and 3-view fused reconstructions from C3VD test set. Predictions use each method's camera poses; native-scale CD is shown below each cloud.

PolypSense3D, frozen three-view geometry obtains a supplied-mask sizing MAE of 4.020 mm, which virtual-only GSCA reduces to 2.897 mm. Because supplied masks and phantom measurements are used, this experiment evaluates metric sizing independently of automatic polyp detection and segmentation. Figure 3 shows representative virtual and physical cases.

6 Conclusion

We present ColScale3R, a parameter-efficient MapAnything adaptation for native-scale reconstruction from calibrated short colonoscopy windows. Across C3VD, SimCol, and physical PolypSense3D, it recovers metric geometry and supports sizing without test-time scale recovery while training only 0.477% of the model. Limitations include calibrated local windows, reference-view supervision, and phantom or simulated validation. Future work will address all-view supervision, deformation-aware fusion, uncertainty-aware frame selection, and in vivo validation.

Acknowledgments. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (Ministry of Science and ICT) under Grant IITP-2026-RS-2020-II201462 (50%), in part by the Korea Evaluation Institute of Industrial Technology (KEIT) under Grant 20026447, and in part by the Chungbuk National University BK21 program (2026).

Disclosure of Interests. The authors declare no competing interests.

References

1. Atkin, W.S., Edwards, R., Kralj-Hans, K., Wooldrage, K., Hart, A.R., Northover, J.M.A., Parkin, D.M., Wardle, J., Duffy, S.W., Cuzick, J.: Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: A multicentre randomised controlled trial. The Lancet **375**(9726), 1624–1633 (2010)
2. 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)
3. Chaptini, L., Chaaya, A., DePalma, F., Hunter, K., Peikin, S., Laine, L.: Variation in polyp size estimation among endoscopists and impact on surveillance intervals. Gastrointestinal Endoscopy **80**(4), 652–659 (2014)
4. 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. pp. 208–218. Springer (2024)
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: Medical Image Computing and Computer Assisted Intervention – MICCAI. pp. 170–180. Springer (2025)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
7. Kalajdzievski, D.: A rank stabilization scaling factor for fine-tuning with lora. arXiv preprint arXiv:2312.03732 (2023)
8. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
9. Khalil, S., Kim, S., Lee, B.I., Hwang, Y.: Self-supervised depth estimation and 3d reconstruction with layer-wise lora of foundation model in endoscopy. IEEE Access (2025)
10. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
11. Liu, R., Wang, L., Zhou, M., Zhang, J., Zhang, H., Liu, X., Cheng, X., Chan, S., Shen, Y., Dai, S., et al.: Polypsense3d: A multi-source benchmark dataset for depth-aware polyp size measurement in endoscopy. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2025)
12. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: International Conference on Machine Learning (2024)
13. Liu, Y., Yu, D., Xiao, Y., He, T., Liu, L.: Scale-consistent 3d reconstruction in monocular colonoscopy via camera-intrinsics-guided learning. Computerized Medical Imaging and Graphics p. 102734 (2026)
14. Nam, S.J., Lim, Y.J., Nam, J.H., Lee, H.S., Hwang, Y., Park, J., Chun, H.J.: 3d reconstruction of small bowel lesions using stereo camera-based capsule endoscopy. Scientific reports **10**(1), 6025 (2020)
15. Oh, D.J., Hwang, Y., Kim, S.H., Nam, J.H., Jung, M.K., Lim, Y.J.: Reading of small bowel capsule endoscopy after frame reduction using an artificial intelligence algorithm. BMC gastroenterology **24**(1), 80 (2024)

16. 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)
17. 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)
18. 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)
19. Shaukat, A., Kahi, C.J., Burke, C.A., Rabeneck, L., Sauer, B.G., Rex, D.K.: Acg clinical guidelines: Colorectal cancer screening 2021. The American Journal of Gastroenterology **116**(3), 458–479 (2021)
20. Sheikh Zeinoddin, M., Hoque, M.I., Tandogdu, Z., Shaw, G.L., Clarkson, M.J., Mazomenos, E.B., Stoyanov, D.: Endo-fast3r: Endoscopic foundation model adaptation for structure from motion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 117–126. Springer (2025)
21. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
22. 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)