CoGE: Sim-to-Real Online Geometric Estimation for Monocular Colonoscopy

Liangjing Shao[1,2], Beilei Cui[1], and Hongliang Ren[⋆,1,2]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Shenzhen Loop Area Institute, China
{leonking-shaw, beileicui}@link.cuhk.edu.hk, hlren@ee.cuhk.edu.hk

Abstract. Geometric estimation including depth estimation and scene reconstruction is a crucial technique for colonoscopy which can provide surgeons with 3D spatial perception and navigation. However, geometric ground truth in colonoscopy is difficult to obtain due to narrow and enclosed space of the colon, while there is a large feature gap between simulated data and realistic data caused by artifacts and illumination. In this paper, we present CoGE, a novel framework for online monocular geometric estimation during colonoscopy. Firstly, we propose an illumination-aware supervision module based on the Retinex theory to address illumination diversity in different colonoscopy scenes. Moreover, a structure-aware perception module is proposed based on wavelet decomposition to extract common structural and local features of the colon. Both quantitative and qualitative results demonstrate that the proposed model solely trained on simulated data achieves state-of-the-art performance in geometric estimation for both simulated and realistic scenes.

Keywords: Depth Estimation · 3D Reconstruction · Colonoscopy.

1 Introduction

Geometric estimation from monocular endoscopic videos is crucial for 3D perception, spatial navigation and path planning in colonoscopy, due to 2D vision of endoscope and long narrow closed pathway of colon.

Recently, numerous 3D reconstruction foundation models [13, 8, 19, 14, 4] have been proposed for cross-scene geometric estimation. However, existing methods usually need a large amount of realistic data with geometric ground truth including depth map, point clouds and camera parameters. Such datasets with geometric ground truth are difficult and costly to obtain in realistic colonoscopy due to the narrow and closed space of colon. Thus, several self-supervised learning-based methods [6, 16] have been developed for colonoscopy geometric estimation. However, the performance of self-supervised models is limited compared with supervised models. Nevertheless, geometric ground truth can be obtained

⋆ Corresponding Author

by simulation software such as VR-Caps [5, 11]. Therefore, sim-to-real geometric estimation has become a potential approach for 3D perception in endoscopic scenes. Currently, some existing works [9, 10] have proposed sim-to-real methods for endoscopy geometric estimation. However, the proposed pipeline is evaluated after transforming the realistic data into the simulation-like data by a style translation network. Hence, the inference efficiency will be decreased due to data preprocessing, while inherent gaps between diverse realistic data and simulated data may degrade the cross-scene performance of appearance transformation models.

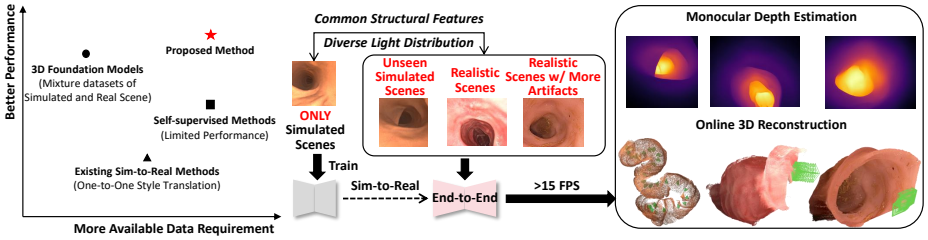


Fig. 1. Given observations from realistic and unseen scenes, our model only trained on simulated data outputs depth maps and 3D point clouds.

In this paper, we propose a feed-forward framework, **CoGE**, for sim-to-real **g**eometric **e**stimation in **co**lonoscopy. In previous work [17], wavelet decomposition has been used to perform frequency-based supervision for global tissue structure and high-frequency details. Inspired by this, a structure-aware perception module based on wavelet transformation is proposed to extract common structural features of colon from colonoscopy images. Moreover, features of colon surface are easily interfered by illumination, so a illumination-aware supervision module is proposed to control the confidence of the estimated results based on light distribution. Besides, considering long distance of colonoscopy sequence, we propose a memory forget mechanism based on attention map between feature of current frame and memory cache to prevent interference of redundant memory.

Our main contribution can be summarized as the following: **1)** A structure-aware perception block is embedded based on wavelet transformation to extract common structural features in both simulated and realistic scenes. **2)** An illumination-aware supervision module is proposed to control the confidence of the estimated results to prevent the influence of diverse illumination on colon surface. **3)** A memory forget mechanism is proposed to filter redundant irrelevant memory for current observation against interference. **4)** The proposed model only trained on simulated data achieves state-of-the-art performance for geometric estimation on both unseen simulated scenes and realistic scenes.

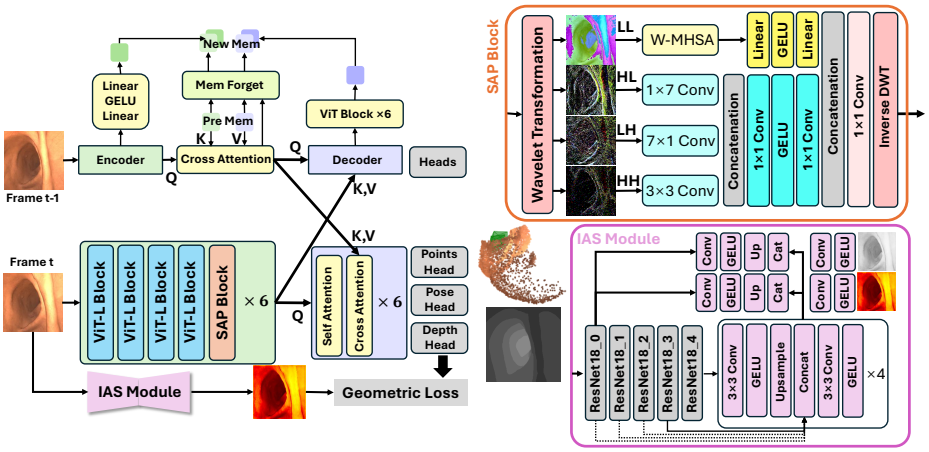


Fig. 2. The pipeline of the proposed framework.

2 Method

2.1 Overview

Given a pair of observations including cached previous observation I_{t-1} and current observation I_t, the proposed model will output the corresponding point map X_t, camera pose P_t, illumination map L_t and confidence map C_t. Meanwhile, inspired by Spann3R [13], two spatial memory caches M_k, M_v are set and they will be updated to $\hat{M}_k, \hat{M}_v$ after memory reading, memory forgetting and memory updating based on the observations. The model architecture is demonstrated in Fig. 2.

2.2 Structure-aware Encoding

Firstly, we utilize six ViT-based modules to extract scene features from the observations $f_{t-1} = \mathcal{F}_{enc}(I_{t-1}), f_t = \mathcal{F}_{enc}(I_t)$. Each module consists of four ViT-L blocks followed by a structure-aware perception (SAP) block, which could extract common structural features from colonoscopy images. In SAP block, we firstly utilize discrete wavelet transformation to decompose the input graph into four subgraphs including low-low I_{ll} with low-frequency main approximation of the input, low-high I_{lh} with horizontal high-frequency details, high-low I_{hl} with vertical high-frequency details and high-high I_{hh} with diagonal high-frequency details. Convolution can extract high-frequency local details while attention is better for low-frequency information extraction. Thus, we utilize window-based multi-head attention (W-MHSA) to extract features f_{ll} from I_{ll}, while vertical-band convolution with 7×1 kernel, horizontal-band convolution with 1×7 kernel and diagonal convolution with 3×3 kernel are respectively utilized to extract

features f_{hl}, f_{lh}, f_{hh} from I_{hl}, I_{lh} and I_{hh}. Then, features from different wavelet compositions are fused based on Eq. 1-3.

$$f_h = \mathcal{F}_{conv}^{1\times1}(\mathcal{F}_{GELU}(\mathcal{F}_{conv}^{1\times1}([f_{hl}, f_{lh}, f_{hh}]))) \tag{1}$$

$$f_l = \mathcal{F}_{lin}(\mathcal{F}_{GELU}(\mathcal{F}_{lin}(f_{ll}))) \tag{2}$$

$$f_{SAP} = \mathcal{F}_{idwt}(\mathcal{F}_{conv}^{1\times1}([f_h, f_l])) \tag{3}$$

where $\mathcal{F}_{lin}$ denotes a linear layer, $\mathcal{F}_{idwt}$ denotes inverse discrete wavelet transformation.

2.3 Memory-related Decoding

Memory Reading. The feature from previous observation f_{t-1} is utilized to retrieve the related memory information from caches by Eq. 5.

$$W_{mem} = softmax(\frac{f_{t-1}M_k^T}{\sqrt{d}}) \tag{4}$$

$$f_{t-1}^{mem} = W_{mem}M_v + f_{t-1} \tag{5}$$

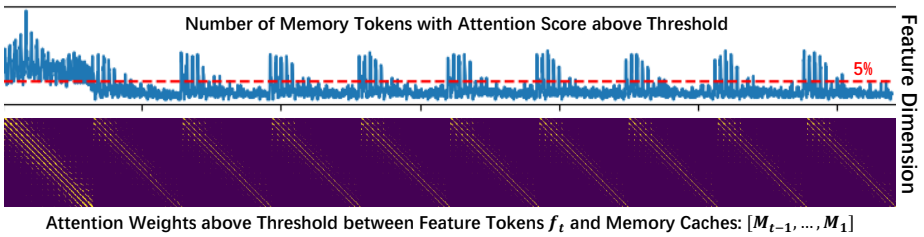


Fig. 3. Attention weights distribution between feature tokens and memory caches.

Memory Forget. For attention maps $W_{mem} \in \mathbb{R}^{d\times s}$ between feature $f_{t-1} \in \mathbb{R}^{d\times1}$ and memory cache $M_k, M_v \in \mathbb{R}^{s\times1}$, if there is at least 5% feature tokens $f_{t-1}(j)$ with an attention weight $W_{mem}(j,i)$ larger than 5e-4 for each memory token $M_k(i), M_v(i)$, the memory token is seemed as the relevant tokens to be retained, demonstrated as Fig. 3.

$$M_k(i), M_v(i) \rightarrow \begin{cases} retain & |\{W_{mem}(j,i) \geq \text{5e-4}\}| \geq 5\%|\{W_{mem}(\cdot,i)\}| \\ delete & \text{otherwise} \end{cases} \tag{6}$$

Decoding and Memory Updating. Given memory-related feature f_{t-1}^{mem} and current feature f_t, the output feature of current frame and previous frame

can be respectively obtained by six decoding blocks, each of which consists of self-attention and cross-attention $\hat{f}_t = \mathcal{F}_{dec}(f_t, f_{t-1}^{mem}), \hat{f}_{t-1} = \mathcal{F}_{dec}(f_{t-1}^{mem}, f_t)$. The new memory of key and value are encoded from memory-related feature $\Delta M_k = \mathcal{F}_{enc}^{key}(f_{t-1}^{mem})$ and output feature of previous frame $\Delta M_v = \mathcal{F}_{enc}^{val}(\hat{f}_{t-1})$. The memory cache is updated by concatenating the new memory and filtered cached memory, $\hat{M}_k, \hat{M}_v = [M_k, \Delta M_k], [M_v, \Delta M_v]$.

Downstream Heads. Given the output feature of the current observation $\hat{f}_t$, DPT heads $\mathcal{H}_{dpt}$ are utilized to output the corresponding point map X_t, camera pose P_t and confidence map C_t, following CUT3R [14]. Furthermore, the corresponding depth map D_t can be calculated based on the point map X_t, camera parameters P_t and K:

$$X_t, P_t, C_t = \mathcal{H}_{dpt}(\hat{f}_t), D_t = \mathcal{Z}(X_t, P_t, K) \tag{7}$$

2.4 Illumination-aware Supervision

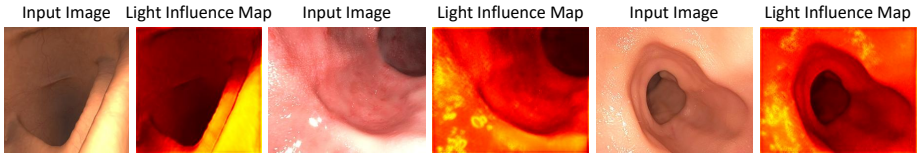


Fig. 4. Input images and corresponding light influence maps.

Considering diverse and unstable illumination in colonoscopy, we propose a self-supervised illumination-aware (IAS) module to adjust confidence map for supervision. Based on retinex theory [7], the invariance of intrinsic component A_t of image I_t to illumination condition in the gradient domain (∇) is considered to generate illumination influence distribution L_t. Given the input image I_t, the IAS module outputs the corresponding intrinsic graph A_t and light influence map L_t shown in Fig. 8. The IAS module is pretrained on the training split using the loss function Eq. 8 for 1 epoch with a batch size of 8. During the training of the whole network, L_t from the frozen IAS module is utilized to adjust confidence map C_t by Eq. 9 with a learnable weight α.

$$\mathcal{L}_{ias} = (1 - L_t) \log \nabla I_t - \log \nabla A_t \tag{8}$$

$$\hat{C}_t = \alpha C_t + (1 - \alpha) L_t \tag{9}$$

Our supervision loss is mainly based on the 3D geometric loss used in CUT3R [14], which can be formulated as Eq.10.

$$\mathcal{L} = \mathcal{L}_{conf}(X_t, \hat{C}_t) + \mathcal{L}_{pose} + \mathcal{L}_{rgb} \tag{10}$$

3 Experiments and Results

3.1 Experiment Implementation

We use 24 simulated sequences along with corresponding camera poses and depth maps obtained by VR-Caps [5, 11] for training. Simulated dataset SimCol3D [11] which includes 9 simulated scenes with 9009 frames and realistic datasets C3VD [1], C3VDv2 [3] including 11 sequences with over 3000 frames are utilized for evaluation. We also extract some in-house colonoscopy images for qualitative comparison. The evaluation splits mainly follow the setting in previous works [16, 6]. The proposed model is initialized with pretrained weights from DUSt3R [15] and trained for 10 epochs with a batch size of 4 on a NVIDIA H100 GPU. All evaluations are implemented on a NVIDIA RTX 4090 GPU.

Table 1. Quantitative results of monocular depth estimation.

	Methods	Year	$Rel_{Abs} \downarrow$	$Rel_{Sq} \downarrow$	$RMSE \downarrow$	$RMSE_{log} \downarrow$	$\delta \uparrow$	FPS
SimCol	**AF-SfM [12]**	2022	0.086	0.036	0.585	0.104	0.954	193
	DepthAnything [18]	2024	0.089	0.042	0.599	0.112	0.948	89
	EndoDAC [2]	2024	0.107	0.050	0.530	0.133	0.924	53
	ColonAdapter [6]	2025	0.062	0.034	0.373	0.094	0.969	-
	MonoPCC [16]	2025	0.058	0.028	0.347	0.090	0.975	-
	Spann3R [13]	2025	0.122	0.103	0.577	0.181	0.795	30
	MonST3R [19]	2025	0.032	0.020	0.205	0.068	0.985	22
	CUT3R [14]	2025	0.036	0.032	0.231	0.070	0.985	34
	STream3R [8]	2026	0.034	0.013	0.204	0.061	0.990	35
	Ours	2026	**0.027**	**0.008**	**0.169**	**0.049**	**0.995**	19
C3VD	**AF-SfM [12]**	2022	0.117	1.324	7.520	0.171	0.882	189
	DepthAnything [18]	2024	0.121	1.155	6.239	0.165	0.874	87
	EndoDAC [2]	2024	0.103	0.652	5.306	0.147	0.890	50
	ColonAdapter [6]	2025	0.139	0.956	5.592	0.175	0.832	-
	Spann3R [13]	2025	0.090	0.523	4.152	0.132	0.891	26
	MonST3R [19]	2025	0.086	0.541	4.207	0.134	0.889	20
	CUT3R [14]	2025	0.087	0.511	4.144	0.129	0.896	29
	STream3R [8]	2026	0.085	0.493	4.105	0.130	0.900	30
	Ours	2026	**0.083**	**0.476**	**4.061**	**0.126**	**0.913**	18
C3VDv2	**Spann3R [13]**	2025	0.108	2.200	8.582	0.305	0.802	20
	MonST3R [19]	2025	0.109	2.014	8.269	0.260	0.825	26
	CUT3R [14]	2025	0.132	**1.713**	9.014	0.275	0.749	29
	STream3R [8]	2026	0.124	2.090	9.205	0.274	0.765	30
	Ours	2026	**0.098**	2.008	**7.772**	**0.250**	**0.865**	18

3.2 Results

Monocular Depth Estimation. The proposed method is compared with SOTA 3D geometric estimation methods on both simulated data and realistic data for

monocular depth estimation. Following previous works [12, 2], we use Abs Rel (Rel_{Abs}), Sq Rel (Rel_{Sq}), RMSE, RMSE Log ($RMSE_{log}$) and $\delta < 1.25$ (δ) as evaluation metrics. Results of MonoPCC [16] and ColonAdapter [6] are directly from public papers, while others are trained for similar time based on the same implementation by our own. Quantitative results in Table 1 and qualitative results in Fig. 5 demonstrate the outstanding performance of our model on monocular depth estimation, only with training on simulated data.

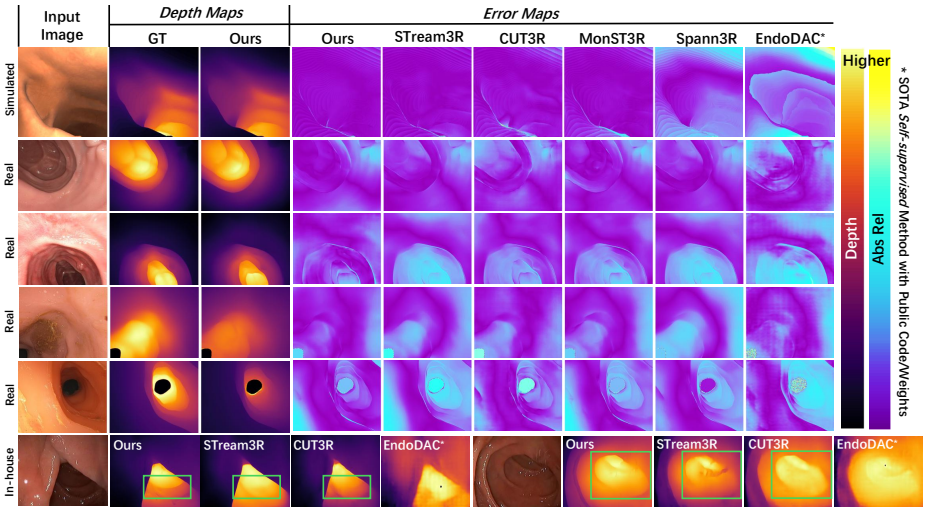


Fig. 5. Qualitative results of monocular depth estimation.

3D Reconstruction. The proposed method is also compared with several 3D foundation models for 3D reconstruction based on monocular videos. We utilize mean Euclidean Distance (mED) to the point cloud generated from ground truth and ratio of points with ED$<N$ mm ($\delta < N$) to evaluate the reconstruction accuracy. Results in Fig.7 demonstrate that our method provides more accurate point cloud reconstruction compared with existing reconstruction methods. Fig. 6 further displays some example results from our method.

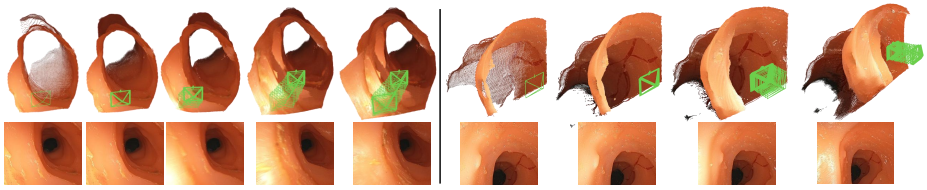
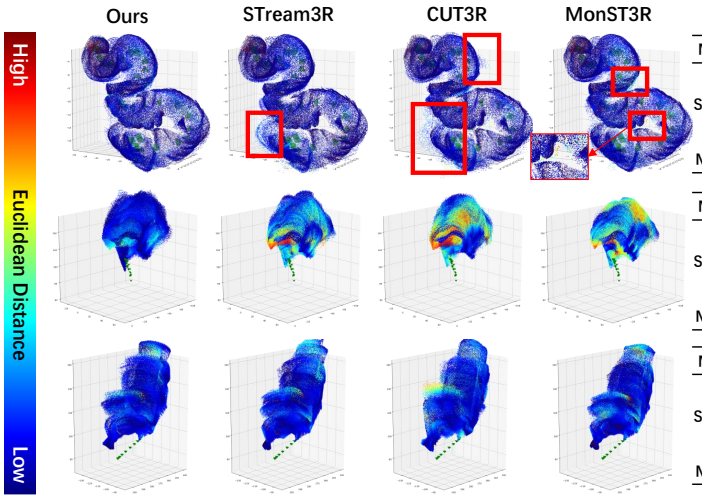


Fig. 6. Qualitative Display of Our 3D reconstruction.



Methods	mED	δ<0.2
Ours	**0.193**	**0.656**
STream3R	0.240	0.457
CUT3R	0.221	0.417
MonST3R	0.217	0.424

Methods	mED	δ<4.0
Ours	**3.957**	**0.645**
STream3R	4.469	0.490
CUT3R	4.243	0.323
MonST3R	4.583	0.215

Methods	mED	δ<3.0
Ours	**2.805**	**0.736**
STream3R	2.952	0.531
CUT3R	3.193	0.301
MonST3R	3.087	0.264

Fig. 7. Qualitative Comparison of 3D reconstruction.

Ablation Studies. Our main contributions include memory forget gate (MemF), illumination-aware supervision module (IAS) and structure-aware perception Block (SAP) for challenges of sim-to-real geometric estimation in colonoscopy. Results shown in Table. 2 and Fig. 8 demonstrate the effectiveness of each main component we propose in this paper.

Table 2. Results of Ablation Studies

Ablations			SimCol		C3VD		C3VDv2	
MemF	IAS	SAP	Abs Rel	δ	Abs Rel	δ	Abs Rel	δ
✔	✔	✔	**0.024**	**0.994**	**0.085**	**0.913**	**0.098**	**0.865**
×	✔	✔	0.026	0.991	0.087	0.905	0.100	0.861
✔	×	✔	0.030	0.988	0.087	0.900	0.104	0.853
✔	✔	×	0.032	0.993	0.089	0.903	0.099	0.860

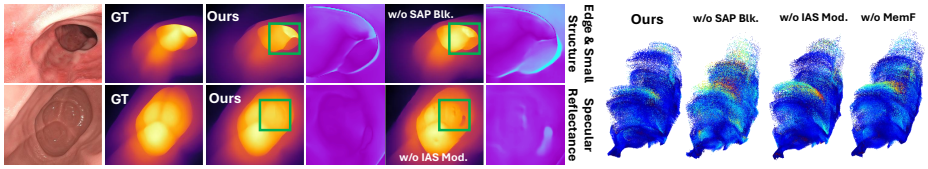


Fig. 8. Qualitative results of ablation studies.

4 Conclusion

In this paper, we propose a geometric estimation model, which is only trained on simulated data, to perform online depth estimation and 3D reconstruction in realistic colonoscopy. We focus on illumination diversity and common structure feature extraction for sim-to-real geometric estimation. Our method outperforms state-of-the-art methods based on quantitative and qualitative experimental results on realistic data, with effects of our contributions. Despite of real-time inference speed, the inference efficiency of the proposed model is still decreased due to extra structure-aware perception blocks. Moreover, the performance of the proposed model will be degraded due to motion blur, acute dynamic change and specific lesion tissue in colonoscopy.

Acknowledgments. This work was supported by Ministry of Science and Technology (MOST) of China Key Project 2025YFE0122500, Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGCX20250526153900001, Hong Kong RGC CRF C4026-21G, RIF R4020-22, GRF 14211420, 14216022 & 14203323.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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)
2. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218. Springer (2024)
3. Golhar, M.V., Fretes, L.S.G., Ayers, L., Akshintala, V.S., Bobrow, T.L., Durr, N.J.: C3vdv2–colonoscopy 3d video dataset with enhanced realism. arXiv preprint arXiv:2506.24074 (2025)
4. 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)
5. İncetan, K., Celik, I.O., Obeid, A., Gokceler, G.I., Ozyoruk, K.B., Almalioglu, Y., Chen, R.J., Mahmood, F., Gilbert, H., Durr, N.J., et al.: Vr-caps: a virtual environment for capsule endoscopy. Medical image analysis **70**, 101990 (2021)
6. Jiang, Z., Wang, Y., Cheng, X., Ge, Z.: Colonadapter: Geometry estimation through foundation model adaptation for colonoscopy. IEEE Robotics and Automation Letters (2025)
7. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: European conference on computer vision. pp. 143–159. Springer (2016)
8. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: Stream3r: Scalable sequential 3d reconstruction with causal transformer. arXiv preprint arXiv:2508.10893 (2025)

9. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE transactions on medical imaging **37**(12), 2572–2581 (2018)
10. Mahmood, F., Durr, N.J.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical image analysis **48**, 230–243 (2018)
11. 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)
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. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 2025 International Conference on 3D Vision (3DV). pp. 78–89. IEEE (2025)
14. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
15. 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)
16. Wang, Z., Zhou, Y., He, S., Li, T., Huang, F., Ding, Q., Feng, X., Liu, M., Li, Q.: Monopcc: Photometric-invariant cycle constraint for monocular depth estimation of endoscopic images. Medical Image Analysis **102**, 103534 (2025)
17. Wu, T., Miao, Y., Guo, J., Chen, Z., Zhao, S., Li, Z., Tang, Z., Huang, B., Yu, L.: Endowave: 4d gaussian splatting with rational wavelet for endoscopic reconstruction. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 6125–6132 (2025)
18. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
19. 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)