MMIC-EndoDepth: Self-Supervised Endoscopic Depth Estimation with Multi-Mechanism Illumination Correction

Ziang Xu[1⋆], Jialu Wang[2⋆], and Bing Wang[3]

[1] The Chinese University of Hong Kong, Hong Kong SAR
[2] University of Oxford, Oxford, United Kingdom
[3] The Hong Kong Polytechnic University, Hong Kong SAR
ziangxu@cuhk.edu.hk, maggirapple@qq.com

Abstract. Accurate and robust depth estimation from monocular endoscopic video is critical for advancing surgical navigation, 3D reconstruction, and augmented reality applications. However, the non-Lambertian reflectance, dynamic illumination, and interreflections prevalent in endoscopic scenes severely violate the brightness constancy assumption underlying conventional self-supervised methods, leading to significant performance degradation. To address this fundamental challenge, we propose MMIC-EndoDepth: a novel self-supervised framework for endoscopic depth estimation centered on Multi-Mechanism Illumination Correction. At its core lies the Illumination-Aware Photometric Correction Module (IAPC), a learnable and interpretable module that explicitly models and compensates for three distinct photometric distortions, including pixel-wise specular highlights, patch-level exposure variations, and depth-guided global lighting attenuation, prior to geometric reconstruction. Rather than treating illumination as a monolithic nuisance, IAPC decomposes it into physically grounded components and employs a Mixture-of-Experts (MoE) architecture with a spatially adaptive gating mechanism. This enables dynamic selection of the most appropriate correction strategy, aligning with clinical intuition and enhancing model explainability. Stable training is ensured through entropy and load-balancing constraints on the MoE, combined with a warmup strategy to promote effective expert utilization. Extensive experiments on standard endoscopic benchmarks (SimCol and C3VD) demonstrate that MMIC-EndoDepth achieves state-of-the-art performance, significantly outperforming existing self-supervised approaches by about 20%. The source code is publicly available at https://github.com/endoloc/endoloc.

Keywords: Monocular depth estimation · Self-supervised learning · Illumination-aware correction.

⋆ Equal contribution.

1 Introduction

Accurate depth estimation from monocular colonoscopy videos is a prerequisite for reliable 3D reconstruction, lesion localization, and augmented reality–assisted interventions in colorectal cancer screening. However, acquiring ground-truth depth in real-world colonoscopy is fundamentally infeasible due to the absence of active depth sensors, the constrained and deformable anatomy of the gastrointestinal tract, and the lack of calibrated stereo setups in clinical practice [13]. These constraints render supervised learning impractical and necessitate self-supervised approaches that leverage only raw video sequences.

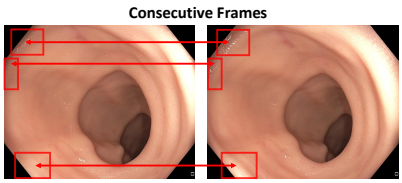

Fig. 1. Consecutive frames of the same mucosal region show strong photometric inconsistency due to dynamic illumination, specular reflections, and interreflections. Red boxes and arrows mark the same anatomy with notably different brightness.

Self-supervised monocular depth estimation has been extensively studied in autonomous driving [15, 16, 5] and general vision domains [7], where methods such as Monodepth2 [5] jointly learn depth and ego-motion by minimizing photometric error between warped and target views. Recent efforts have sought to transfer these paradigms to endoscopic environments. Turan et al. [12] and Liu et al. [6] applied standard self-supervised pipelines to surgical videos but observed significant degradation due to unmodeled illumination dynamics. To mitigate this, AF-SfMLearner [11] introduced an appearance flow module to model brightness changes, achieving improved robustness on surgical datasets. More recently, EndoDAC [3] proposed an efficient adaptation framework for foundation models using Dynamic Vector-Based Low-Rank Adaptation, reducing trainable parameters while estimating camera intrinsics in a self-supervised manner. Despite these advances, their performance in colonoscopy remains severely limited by the violation of the brightness constancy assumption, as show in Fig. 1.

To overcome these limitations, we propose MMIC-EndoDepth, a self-supervised depth estimation framework specifically designed for colonoscopy that explicitly models illumination heterogeneity through a Multi-Mechanism Illumination Corrector. Built upon an Illumination-Aware Photometric Correction Module (IAPC), our method decomposes photometric distortions into three physically grounded components and employs a Mixture-of-Experts (MoE) architecture with spatially adaptive gating to dynamically select the optimal correction strategy per image region.

Our contributions are summarized as follows:

- We present MMIC-EndoDepth, a self-supervised monocular depth method for colonoscopy that handles spatially varying illumination.
- We introduce IAPC, an MoE module that models photometric distortions at pixel, patch, and depth-bin levels for accurate and interpretable depth.
- Extensive experiments on SimCol and C3VD demonstrate state-of-the-art performance and strong cross-dataset generalization, outperforming recent state-of-the-art methods.

2 Methods

2.1 Self-Supervised Depth and Ego-Motion Estimation Pipeline

We build on a standard self-supervised monocular depth estimation framework [15] (Fig. 2-(a)), with a depth network and a pose network as the backbone. During training, we take adjacent frame pairs for view synthesis and insert an illumination-aware photometric correction module (IAPC) before computing the photometric consistency loss to correct appearance changes caused by dynamic lighting and reflections, thereby providing more reliable photometric supervision. During inference, we keep only the depth network to predict depth from a single image; IAPC is not used, so no additional runtime overhead is introduced.

2.2 Illumination-Aware Photometric Correction (IAPC) Module

Common photometric distortions in endoscopy violate the photometric consistency assumption, making reprojection-based photometric loss unreliable. To mitigate this issue, we introduce MoE-IAPC at each pyramid scale $s \in \{0, 1, 2, 3\}$. Three complementary experts predict per-pixel affine parameters (shift/scale), which are then adaptively fused by a pixel-wise gate in the parameter space.

Given a target frame I_t and an adjacent source frame $I_{t'}$, we warp the source to the target view using the predicted depth and pose, obtaining $I_{t'}(u)$, where $u = (x, y)$ denotes pixel coordinates. For each scale s, we apply an affine correction to the warped image, where $B_s(u)$ and $C_s(u)$ are the shift and scale at pixel u and scale s, and clamp$(\cdot)$ clips values to $[0, 1]$ (see Eq. (1)).

$$\hat{I}_t(u) = \text{clamp}\big(I_{t'}(u) \cdot C_s(u) + B_s(u)\big), \tag{1}$$

Expert 1: Pixel Expert (specular highlights / high-frequency spots). Specular highlights are local, sharp, and rapidly varying, thus requiring high-capacity modeling. Let x_s be the scale-s feature map (an intermediate representation from the depth/pose network). The Pixel Expert predicts per-pixel affine parameters:

$$[\beta_s^p(u), \alpha_s^p(u)] = \text{Conv}_{3\times3}(x_s), \tag{2}$$

where $\beta_s^p(u)$ is the shift and $\alpha_s^p(u)$ is the scale (constrained to be positive via $\exp(\cdot)$ or softplus$(\cdot)$). We denote its output as

$$[B_s^{(1)}(u), C_s^{(1)}(u)] = [\beta_s^p(u), \alpha_s^p(u)]. \tag{3}$$

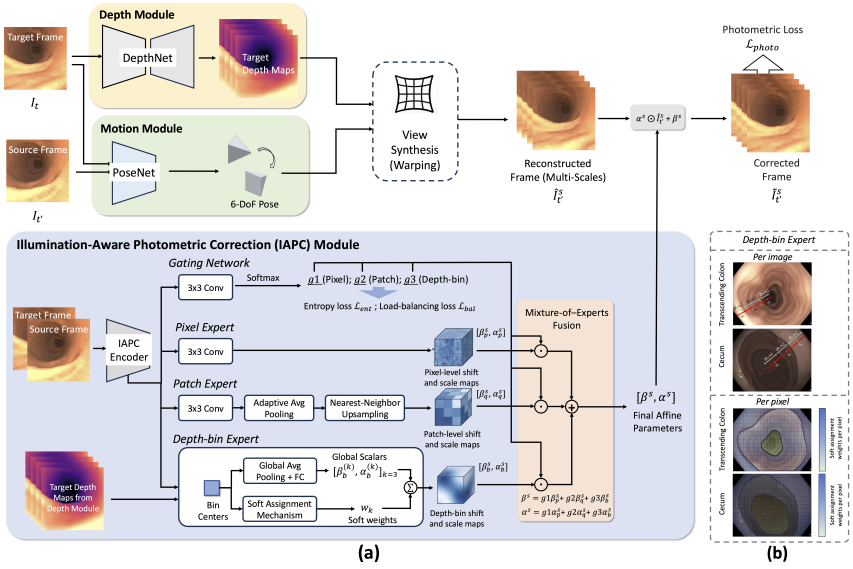

Fig. 2. (a) MMIC-EndoDepth overview:(a) depth-and-pose backbone with training-only IAPC (before photometric loss); inference uses only the depth network (no overhead). **(b) Depth-bin Expert (physical prior)**: Using the predicted depth as a distance cue, we compute soft bin assignments $w^k(u)$ with learnable global centers $\{d_k\}$ (Eq. 6), and use them to blend depth-bin affine templates $\{(\beta^s_{b,k}, \alpha^s_{b,k})\}$ dynamically generated from x_s, yielding pixel-wise correction parameters (Eq. 7).

Expert 2: Patch Expert (exposure drift / low-frequency shadows). Exposure drift and shadows vary smoothly in space; purely per-pixel parameters can be noisy and unstable. The Patch Expert first produces pixel-wise parameters and then enforces patch-wise sharing via average pooling followed by upsampling:

$$[\beta^q_s, \alpha^q_s] = \mathrm{Conv}_{3\times3}(x_s), \quad [\beta^q_s, \alpha^q_s] \leftarrow \mathrm{Upsample}\Big(\mathrm{AvgPool}_{p_s\times p_s}([\beta^q_s, \alpha^q_s])\Big), \quad (4)$$

with patch size $p_s = \max(1, p_{\text{base}}/2^s)$ and $p_{\text{base}} = 8$. The expert output is

$$[B^{(2)}_s(u), C^{(2)}_s(u)] = [\beta^q_s(u), \alpha^q_s(u)]. \quad (5)$$

Expert 3: Depth-bin Expert (physical prior). We exploit a simple observation in colonoscopy: brightness varies with camera–tissue distance (nearer regions tend to be brighter, farther regions darker). We therefore treat the *predicted depth* at each pixel as a distance cue and modulate the affine correction strength accordingly: **(1) per-image (global then local):** the continuous depth range is partitioned into K *depth bins* (we use K=3 by default, but larger K is possible) in a *learnable* manner, so the bin allocation adapts to each frame; **(2) per-pixel:** each pixel softly selects and blends the K depth-bin templates based

on its predicted depth, yielding a depth-continuous correction. As a result, pixels with similar depths are encouraged to share the same template and learn consistent compensation, improving stability in dark regions and depth-transition areas while avoiding discontinuities from fixed hard thresholds .

As shown in Fig. 2-(b), let D_t^s denote the predicted depth map at pyramid scale s and $D_t^s(u)$ the depth at pixel u. We represent K depth bins (near/mid/far) by learnable *global* centers $\{d_k\}_{k=1}^K$, shared across the dataset and updated during training. For each pixel u, its soft assignment to the k-th depth bin is computed from the distance to the centers:

$$w_s^k(u) = \mathrm{softmax}_k\Big(-\tau \big| D_t^s(u) - d_k \big| \Big), \qquad \sum_{k=1}^{K} w_s^k(u) = 1, \tag{6}$$

where τ controls the sharpness of the assignment. The expert predicts an image-level affine pair for each bin, $\{(\beta_{b,k}^s, \alpha_{b,k}^s)\}_{k=1}^K$, from the scale feature map x_s (shared across all pixels within the same image but dynamically generated per image), and mixes them into pixel-wise parameters:

$$\beta_b^s(u) = \sum_{k=1}^{K} w_s^k(u)\, \beta_{b,k}^s, \qquad \alpha_b^s(u) = \sum_{k=1}^{K} w_s^k(u)\, \alpha_{b,k}^s. \tag{7}$$

Gated fusion in parameter space. Rather than fusing corrected images, we fuse affine parameters for stability and interpretability. The pixel-wise gate predicts weights from x_s:

$$[g_s^1(u), g_s^2(u), g_s^3(u)] = \mathrm{softmax}\big(\mathrm{Conv}_{3\times3}(x_s)\big), \tag{8}$$

and produces the final affine parameters by convex combination:

$$B_s(u) = \sum_{j=1}^{3} g_s^j(u) B_s^{(j)}(u), \quad C_s(u) = \sum_{j=1}^{3} g_s^j(u) C_s^{(j)}(u), \tag{9}$$

where $g_s^j(u) \geq 0$ and $\sum_{j=1}^3 g_s^j(u) = 1$.

2.3 Loss Function and Training Strategy

To stabilize MoE gating and prevent expert collapse, we apply two common regularizers to the gating weights g: an entropy regularizer and a load-balancing regularizer. The entropy term encourages a more uniform expert assignment at early stages, while the load-balancing term enforces long-term balanced utilization across experts. We combine them as $\mathcal{L}_{\mathrm{moe}} = \mathcal{L}_{\mathrm{ent}} + \mathcal{L}_{\mathrm{bal}}$. To avoid overly strong regularization that may hinder convergence in the early phase, we linearly warm up its weight: $\lambda_{\mathrm{moe}}^{(t)} = \lambda_{\mathrm{moe}} \cdot \min\left(1, \frac{t}{N_{\mathrm{warm}}}\right)$, where $\lambda_{\mathrm{moe}} = 0.01$ and $N_{\mathrm{warm}} = 2000$. The final training objective is $\mathcal{L} = \frac{1}{4}\sum_{s=0}^{3}\left(\min \mathcal{L}_{\mathrm{photo}} + \lambda_{\mathrm{smooth}} \frac{\mathcal{L}_{\mathrm{smooth}}^s}{2^s}\right) + \lambda_{\mathrm{moe}}^{(t)} \mathcal{L}_{\mathrm{moe}}$, where $\mathcal{L}_{\mathrm{photo}}$ is the combined SSIM and L_1 photometric error, $\mathcal{L}_{\mathrm{smooth}}^s$ is the edge-aware smoothness term at scale s, and $\lambda_{\mathrm{smooth}} = 0.001$.

3 Experiments and results

3.1 Datasets and evaluation

Existing colonoscopy depth estimation research commonly employs synthetic data, as it provides RGB images paired with ground truth depth maps for direct evaluation of depth estimation models. We conduct experiments on two publicly available colonoscopy datasets: SimCol (synthetic) [9] and C3VD (phantom) [1]. The SimCol dataset comprises 33 synthetic colon scenes rendered from real CT scans, yielding 37,833 RGB frames, each paired with ground truth depth maps and camera poses. Following the standard protocol [9], we use trajectories from SyntheticColon I and II for training, validation, and the first test set. Additionally, all frames from SyntheticColon III (O1–O3) are reserved as an other test set. The C3VD dataset provides 22 clinically acquired video sequences registered to high-fidelity 3D colon models, resulting in 10,015 frames with ground-truth depth. To assess the model's generalization capability, we utilized the entire C3VD dataset as an out-of-sample test set, without any training or validation samples drawn from it.

We evaluate depth estimation accuracy using eight standard metrics as established in [9, 4]: Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), RMSE log, and the percentage of pixels with relative error below thresholds of $\delta < 1.25, \delta < 1.25^2, \delta < 1.25^3$. The performance of the pose estimation is assessed on the 5-frame using the absolute trajectory error (ATE), following the protocol in [3].

3.2 Implementation Details

We implement our framework in PyTorch and train on NVIDIA A100 GPUs. Our IAPC module is trained jointly with the main network. We use the AdamW optimizer with a weight decay of $1e^{-4}$. The model is trained for 30 epochs with a batch size of 12. We use the standard self-supervised losses: a combination of SSIM and photometric loss, and an edge-aware smoothness loss weighted by $1e^{-3}$. For the IAPC module, the auxiliary loss weight λ_{moe} is set to 0.01 with a linear warmup over the first 2000 steps. All comparison methods are trained and evaluated on the same SimCol dataset split [9].

3.3 Comparison with SOTA methods

We compare our method against a comprehensive set of state-of-the-art approaches on the SimCol [9] and C3VD [1] datasets, as shown in Tables 1. On the SyntheticColon I/II test set, our method achieves the best performance across all metrics, reducing RMSE by **24.3%** vs. EndoDAC and **36.7%** vs. Monodepth2. On the challenging SyntheticColon III split, the RMSE improvement over EndoDAC by **24%**. Pre-trained foundation models (Depth Pro [2], Depth Anything V2 [14]) underperform, emphasizing the distinct endoscopic challenges and the

Table 1. Depth estimation comparison on SimCol and unseen C3VD.

Dataset	Method	Depth Error (↓)				Depth Accuracy (↑)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SyntheticColon-I SyntheticColon-II	Monodepth2 [5]	0.133	0.169	0.614	0.171	0.869	0.955	0.969
	Lite-mono [15]	0.115	0.142	0.557	0.152	0.895	0.968	0.978
	MonoDiffusion [10]	0.118	0.145	0.565	0.155	0.890	0.965	0.975
	Depth Pro‡ [2]	0.142	0.182	0.640	0.185	0.855	0.945	0.960
	Depth Anything V2‡ [14]	0.135	0.172	0.627	0.173	0.865	0.952	0.967
	Endo-SFMLearner* [8]	0.125	0.158	0.590	0.165	0.881	0.962	0.972
	AF-SFMLearner* [11]	0.116	0.143	0.561	0.153	0.892	0.967	0.977
	EndoDAC* [3]	0.095	0.105	0.514	0.125	0.935	0.982	0.990
	MMIC-EndoDepth (Ours)	**0.086**	**0.088**	**0.389**	**0.103**	**0.960**	**0.989**	**0.995**
SyntheticColon-III	Monodepth2 [5]	0.119	0.137	0.533	0.151	0.899	0.960	0.977
	Lite-mono [15]	0.107	0.120	0.507	0.134	0.920	0.968	0.979
	MonoDiffusion [10]	0.111	0.129	0.513	0.139	0.924	0.970	0.979
	Depth Pro‡ [2]	0.132	0.154	0.545	0.165	0.878	0.949	0.961
	Depth Anything V2‡ [14]	0.130	0.151	0.539	0.162	0.889	0.957	0.965
	Endo-SFMLearner* [8]	0.113	0.126	0.520	0.141	0.923	0.975	0.982
	AF-SFMLearner* [11]	0.099	0.094	0.504	0.128	0.939	0.979	0.985
	EndoDAC* [3]	0.082	0.075	0.489	0.107	0.942	0.980	0.991
	MMIC-EndoDepth (Ours)	**0.057**	**0.028**	**0.366**	**0.086**	**0.967**	**0.988**	**0.998**
C3VD	Monodepth2 [5]	0.377	4.195	13.640	0.446	0.482	0.771	0.882
	Lite-mono [15]	0.299	3.572	9.845	0.327	0.557	0.815	0.909
	MonoDiffusion [10]	0.315	3.685	9.992	0.338	0.531	0.799	0.901
	Depth Pro‡ [2]	0.379	4.217	13.006	0.429	0.488	0.779	0.889
	Depth Anything V2‡ [14]	0.360	3.951	12.418	0.421	0.502	0.792	0.899
	Endo-SFMLearner* [8]	0.275	3.089	8.549	0.295	0.597	0.833	0.914
	AF-SFMLearner* [11]	0.241	2.830	8.278	0.271	0.619	0.851	0.935
	EndoDAC* [3]	0.207	2.472	7.385	0.247	0.655	0.873	0.940
	MMIC-EndoDepth (Ours)	**0.101**	**1.877**	**5.931**	**0.117**	**0.701**	**0.915**	**0.979**

* Endoscopy-specific self-supervised method; ‡ Pre-trained foundation model on natural images
Note: Depth error for SimClo dataset are reported in centimeters, while those for C3VD are in millimeters.

the importance of domain-specific design. Crucially, on the unseen C3VD phantom dataset, our method shows strong generalization, improving RMSE by 19.6% and Abs Rel by 48.3% over EndoDAC. These consistent and significant improvements validate that our multi-mechanism illumination correction strategy effectively addresses photometric inconsistency in endoscopic depth estimation.

Fig. 3 shows qualitative depth predictions on synthetic (SimCol) and phantom (C3VD) colonoscopy sequences. Our method produces significantly more coherent and anatomically plausible depth maps compared to baselines. Notably, in regions with severe specular highlights or interreflections (bottom row), competing methods exhibit distorted depth. In contrast, our IAPC module effectively corrects photometric inconsistencies, preserving structural integrity. Moreover, the depth boundaries of anatomical landmarks are sharper and more continuous in our results, demonstrating superior edge preservation and robustness to illumination variations.

We evaluate pose estimation on two selected sequences: SyntheticColon III O1 from SimCol and Sigmoid_t1_a from C3VD. As shown in Table 2, our method achieves the lowest ATE error on both sequences, outperforming all competing approaches.

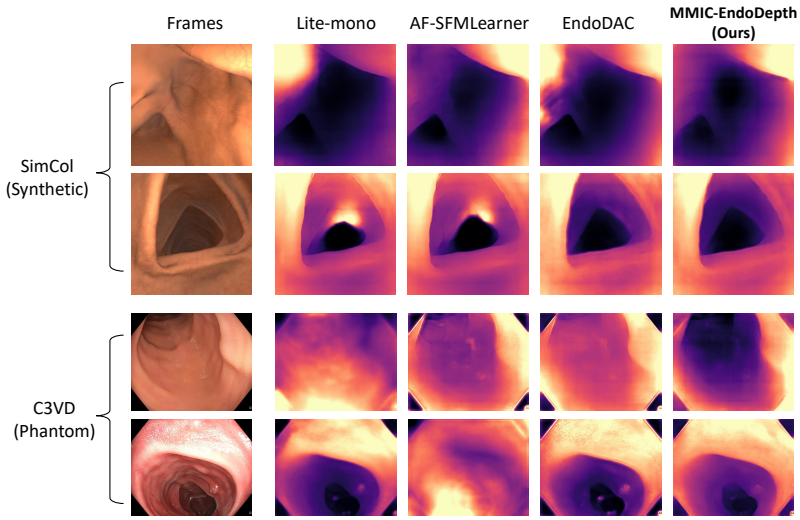

Fig. 3. Qualitative depth comparison on the SimCol and C3VD datasets.

Table 2. Quantitative pose estimation comparison on two selected sequences.

Method	ATE (↓)	
	SyntheticColon III (O1)	Sigmoid_t1_a
Lite-mono	0.3539	0.2108
AF-SFMLearner	0.3122	0.2076
EndoDAC	0.3218	0.2014
Ours	**0.2975**	**0.1901**

Table 3. Ablation study of our method under different expert-module settings on SimCol (SyntheticColon I/II split).

Expert Module			Depth Error (↓) (in cm)			
Pixel	Patch	Depth-bin	Abs Rel	Sq Rel	RMSE	RMSE log
✗	✗	✗	0.115	0.142	0.557	0.152
✓	✗	✗	0.101	0.125	0.455	0.137
✗	✓	✗	0.099	0.119	0.431	0.130
✗	✗	✓	0.108	0.131	0.474	0.139
✓	✓	✗	0.093	0.105	0.418	0.117
✗	✓	✓	0.098	0.107	0.425	0.122
✓	✗	✓	0.099	0.118	0.439	0.132
✓	✓	✓	**0.086**	**0.088**	**0.389**	**0.103**

3.4 Ablation study

We conduct an ablation study on the SimCol dataset to validate the contribution of each component in our IAPC module. As shown in Table 3, the full model with all three experts (Pixel, Patch, and Depth-bin) achieves the best performance, with an RMSE of 0.389, significantly outperforming the baseline without any expert (0.557). The combination of Pixel and Patch experts yields the strongest individual contribution among pairs.

4 Conclusion

In this paper, we propose MMIC-EndoDepth, a novel self-supervised framework for endoscopic depth estimation that explicitly addresses the fundamental challenge of spatially variant photometric inconsistency. Unlike prior methods, we introduce the IAPC module, which is the first to decompose lighting distortions

into three physically grounded mechanisms. By integrating these via a MoE architecture with adaptive gating, IAPC enables fine-grained, context-aware correction, leading to significantly more robust and accurate depth estimates on both synthetic and real-world phantom colonoscopy data. Extensive ablation and comparison studies confirm that this multi-mechanism design is essential for stable training and superior performance. In the future, we plan to extend our framework to broader surgical video domains (e.g., laparoscopy, bronchoscopy).

Disclaimer The concepts and information presented in this paper are based on research results that are not commercially available. Future availability cannot be guaranteed.

Disclosure of Interests The authors have no competing interests to declare.

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. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv (2024), https://arxiv.org/abs/2410.02073
3. 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)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
6. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE transactions on medical imaging **39**(5), 1438–1447 (2019)
7. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021)
8. 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)
9. 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)
10. Shao, S., Pei, Z., Chen, W., Sun, D., Chen, P.C., Li, Z.: Monodiffusion: self-supervised monocular depth estimation using diffusion model. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 3664–3678 (2024)

11. 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)
12. Turan, M., Ornek, E.P., Ibrahimli, N., Giracoglu, C., Almalioglu, Y., Yanik, M.F., Sitti, M.: Unsupervised odometry and depth learning for endoscopic capsule robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1801–1807. IEEE (2018)
13. Xu, Z., Li, B., Hu, Y., Zhang, C., East, J., Ali, S., Rittscher, J.: Self-supervised monocular depth and pose estimation for endoscopy with generative latent priors. arXiv preprint arXiv:2411.17790 (2024)
14. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
15. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18537–18546 (2023)
16. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: Monovit: Self-supervised monocular depth estimation with a vision transformer. In: 2022 international conference on 3D vision (3DV). pp. 668–678. IEEE (2022)