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

MEDE-TTA: Cross-Scenario Monocular Endoscopy Depth Estimation via Test Time Adaptation

Liangjing Shao[1,2], Beilei Cui[1], Long Bai[1], Mobarak I. Hoque[3], 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, Shenzhen, China
[3] Division of Informatics, Imaging and Data Science, University of Manchester, UK
{leonking-shaw, beileicui}@link.cuhk.edu.hk, hlren@ee.cuhk.edu.hk

Abstract. Monocular depth estimation supports navigation, reconstruction, and augmented reality in endoscopic surgery, but models trained on one clinical setting often fail when deployed across hospitals, scopes, and anatomical sites. To this end, we present MEDE-TTA, a unified framework for cross-scenario monocular endoscopic depth estimation via test-time adaptation (TTA), to adapt a source-trained network for inference on unseen data without ground-truth labels. We propose Aug-Teacher to construct robust pseudo labels by averaging EMA predictions over color-jittered views, using anchor-confidence gating to switch between augmentation-averaged and standard EMA teachers. To preserve pre-trained knowledge and limiting catastrophic drift, a stochastic reset (Stoch-Reset) mechanism is proposed to restores adapted parameters toward the source snapshot after each update. Besides, a scale-invariant depth consistency loss is utilized to further stabilize adaptation. Zero-shot experiments on Hamlyn and C3VD with the baseline model respectively trained on SCARED and SimCol verify that MEDE-TTA outperforms existing TTA methods and state-of-the-art endoscopy depth estimation methods on **training-free monocular depth estimation in unseen endoscopic environment**. The code has been released on https://github.com/BaymaxShao/MEDE-TTA

Keywords: Test Time Adaptation · Depth Estimation · Endoscopy.

1 Introduction

Monocular depth estimation (MDE) has emerged as a transformative technology in minimally invasive surgery (MIS) for critical applications such as autonomous robotic navigation and safe control by providing three-dimensional spatial awareness.

⋆ Corresponding Author

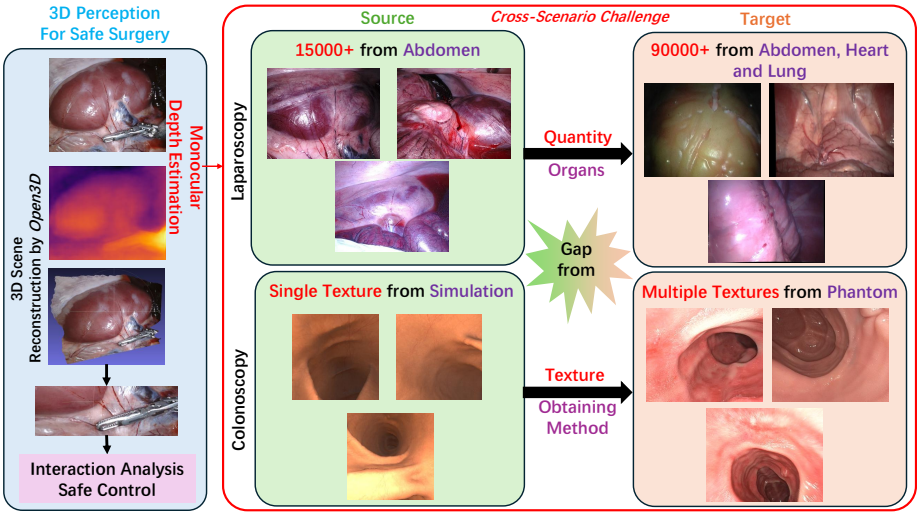


Fig. 1. Application of MDE for safe surgery and cross-scenario challenges for MDE in endoscopy.

While foundation models such as Depth Anything[16], VGGT[12], Scal3R[15], have demonstrated remarkable potential in learning intricate geometric structures from large-scale data, their direct application to surgical domains requires labeled target data, which is prohibitively expensive and difficult to obtain in clinical settings due to the lack of reliable ground-truth depth sensors in standard endoscopic setups. [4] Meanwhile, for application in unseen environment, performance of self-supervised methods such as [18], EndoDAC[3] and MonoPCC[14] can be severely limited by large feature gap between diverse endoscopic scenes, which needs to be solved by extra retraining. Consequently, there is an urgent need for robust adaptation strategies that can align model predictions with unseen target domains without relying on ground-truth labels or extensive retraining processes.

Test-time adaptation (TTA) [8, 19, 5] offers a promising solution by enabling online model adjustment during inference using only incoming test data. Continual TTA (CTTA) extends this setting to long test streams in which domain statistics drift over time that closely matches real endoscopic deployment [13]. However, directly transferring CTTA recipes from natural-image classification to endoscopic depth estimation remains challenging. Pseudo-supervision from a single teacher view is brittle under specular highlights, smoke, and color shifts. [11, 7] Moreover, disparity-space losses are sensitive to unknown global scale. Besides, unconstrained online updates accumulate self-training errors, eventually drifting away from useful source knowledge.

In this work, we present **MEDE-TTA** to target continual, label-free adaptation by updating only parameter-efficient finetuning modules and decoder pa-

rameters for monocular depth estimation in unseen endoscopic scenario. Our main contributions can be summarized as the following:

- an *augmentation-averaged teacher* (**Aug-Teacher**) aggregates EMA predictions over photometrically jittered views and selects between augmentation-averaged and standard teacher targets via anchor-confidence gating, yielding more stable pseudo labels under appearance perturbations.
- *stochastic parameter restoration* (**Stoch-Reset**) randomly reverts adapted weights toward the source snapshot after each update, mitigating error accumulation over long test sequences.
- The proposed method achieve **state-of-the-art performance** on training-free depth estimation in unseen endoscopic environments from colonoscopy (SimCol→C3VD) and laparoscopy (SCARED→Hamlyn), compared with existing TTA methods and endoscopy depth estimation methods.

2 Proposed Method

2.1 Problem Formulation

Let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote the source and target endoscopic domains, respectively. We assume a monocular depth network f_Θ pre-trained on $\mathcal{D}_s$ and deployed on a stream of unlabeled test frames $\{x_t\}_{t=1}^{T}$ from $\mathcal{D}_t$. Our goal is *continual test-time adaptation* (CTTA): at each time step t, the model observes x_t, updates a small set of adaptation parameters, and predicts depth $\hat{d}_t = f_{\Theta_t}(x_t)$ without access to ground-truth depth. Following EndoDAC, f_Θ is built upon a ViT encoder with Dynamic Vector-based LoRA (DV-LoRA) and a DPT-style depth head; only LoRA and depth-head parameters are unfrozen during adaptation.

2.2 Overview of MEDE-TTA

We propose **MEDE-TTA**, a label-free CTTA method that stabilizes online adaptation through two complementary mechanisms: (i) **Aug-Teacher**, a robust pseudo-supervision strategy that aggregates teacher predictions over photometric augmentations and selects the teacher via anchor-confidence gating; and (ii) **Stoch-Reset**, a stochastic parameter restoration scheme that periodically reverts adapted weights toward the source snapshot to mitigate error accumulation. MEDE-TTA further optimizes a scale-invariant depth consistency objective and employs grouped learning rates for domain-routing and task-specific parameters. The whole pipeline is shown as Fig. 2

2.3 Aug-Teacher: Augmentation-Averaged Pseudo Supervision

MEDE-TTA maintains three copies of the network: the *student* f_{Θ_t}, a frozen *anchor* f_{Θ_0}, and an exponential moving average (EMA) *teacher* $f_{\Theta_t^{\text{ema}}}$. Given input

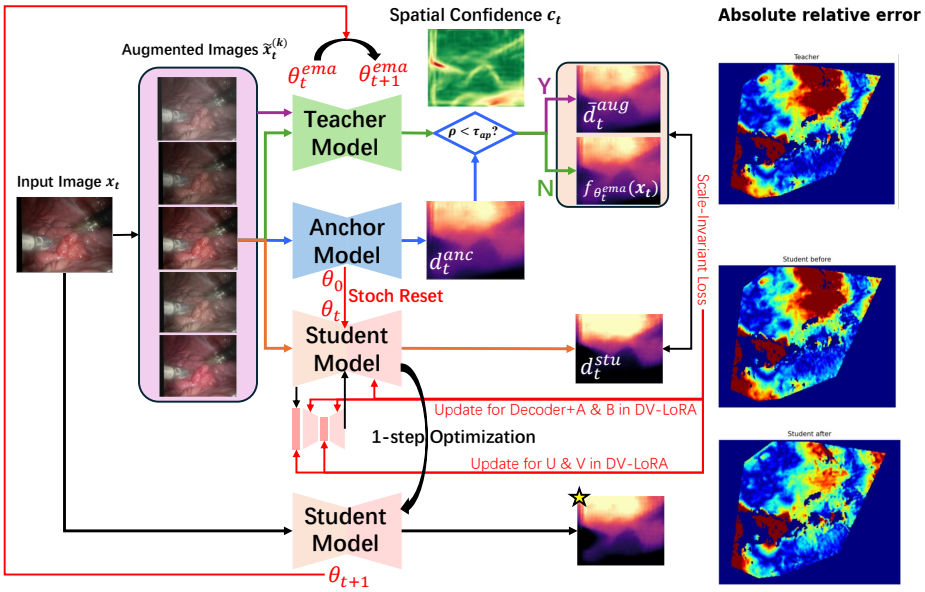


Fig. 2. The pipeline of MEDE-TTA

x_t, we first generate K color-jittered views $\{\tilde{x}_t^{(k)}\}_{k=1}^{K}$ and form an augmented batch. The augmentation-averaged teacher disparity is

$$\bar{d}_t^{\text{aug}} = \frac{1}{K} \sum_{k=1}^{K} f_{\Theta_t^{\text{ema}}} \left(\tilde{x}_t^{(k)} \right), \tag{1}$$

which reduces variance induced by appearance perturbations and yields more stable pseudo labels under endoscopic illumination changes. To prevent over-smoothing when the anchor prediction is already reliable, we introduce *anchor-confidence gating*. Let $d_t^{\text{anc}} = f_{\Theta_0}(x_t)$ denote the anchor disparity. We define a spatial confidence map as

$$c_t = 1 - \text{Std}_{\text{sp}}(d_t^{\text{anc}}), \tag{2}$$

where $\text{Std}_{\text{sp}}(\cdot)$ is the spatial standard deviation over each disparity map. The scalar anchor confidence is $\rho_t = \mathbb{E}[c_t]$. The final teacher target is selected as

$$d_t^{\text{tea}} = \begin{cases} \bar{d}_t^{\text{aug}}, & \text{if } \rho_t < \tau_{\text{ap}}, \\ f_{\Theta_t^{\text{ema}}}(x_t), & \text{otherwise}, \end{cases} \tag{3}$$

where τ_{ap} is a confidence threshold. When the anchor is uncertain ($\rho_t < \tau_{\text{ap}}$), MEDE-TTA favors the augmentation-averaged teacher; otherwise it uses the standard EMA prediction on the clean input, preserving sharp structure.

2.4 Stoch-Reset: Stochastic Parameter Restoration

In long test streams, self-training errors can accumulate and gradually drift Θ_t away from useful source knowledge. Inspired by stochastic weight restoration, we introduce **Reset** after each adaptation step. Let Θ_0 denote the source snapshot taken at the beginning of testing. For each trainable parameter $\theta \in \Theta_t$, we independently sample a binary mask $m \sim \text{Bernoulli}(p_{\text{rst}})$ and update

$$\theta \leftarrow m \cdot \theta_0 + (1 - m) \cdot \theta, \tag{4}$$

where p_{rst} controls the restoration probability. Reset acts as a soft regularizer: it allows local adaptation while preventing irreversible drift, which is particularly important for continual deployment in endoscopy.

2.5 Scale-Invariant Consistency Loss

Monocular depth is defined up to an unknown global scale under domain shift. Direct ℓ_2 consistency in disparity space is sensitive to such scale mismatch. Therefore, MEDE-TTA converts student and teacher disparities to depth via the standard bounded inverse mapping and minimizes a scale-invariant log-depth loss:

$$\mathcal{L}_{\text{SI}} = \sqrt{\mathbb{E}[\Delta_t^2] - \lambda\, \mathbb{E}[\Delta_t]^2} \cdot s, \quad \Delta_t = \log d_t^{\text{stu}} - \log d_t^{\text{tea}}, \tag{5}$$

where $d_t^{\text{stu}} = f_{\Theta_t}(x_t)$, $\lambda \in [0, 1)$ is a variance-focus hyper-parameter, and s is a loss scale factor. The overall adaptation objective at step t is $\mathcal{L} = \mathcal{L}_{\text{SI}}$.

2.6 Optimization and Model Maintenance

Parameter subset. Only DV-LoRA branches and depth-head parameters are updated; the ViT backbone remains frozen except for injected adapters.

Grouped learning rates. DV-LoRA contains domain-routing vectors (lora_U/V) and task-specific low-rank factors (lora_A/B). We assign a higher learning rate η_{uv} to domain-routing vectors and a lower rate η_{base} to remaining adaptation parameters.

Dual-rate EMA. After each gradient step, the teacher is updated by

$$\Theta_{t+1}^{\text{ema}} \leftarrow \alpha(\phi)\, \Theta_t^{\text{ema}} + \big(1 - \alpha(\phi)\big)\, \Theta_{t+1}, \tag{6}$$

where $\alpha(\phi) = \alpha_{\text{uv}}$ if $\phi \in \{\texttt{lora_U}, \texttt{lora_V}\}$, and $\alpha(\phi) = \alpha_{\text{tea}}$ otherwise.

3 Experiments

3.1 Implementation

We evaluate the proposed method in both laparoscopic and colonoscopic scenes, while datasets used for experiments are introduced and compared in Table 1. All experiments are implemented on single NVIDIA RTX 4090 GPU. MEDE-TTA is built on EndoDAC, while only LoRA parameters and the DPT head are updated at test time. Inputs are resized to 256×320. We perform continual online adaptation with one Adam step per frame, using learning rates $\eta_{\text{uv}} = 10^{-3}$ for `lora_U/V` and $\eta_{\text{base}} = 10^{-4}$ for other adaptation parameters. The EMA teacher uses $\alpha_{\text{tea}} = 0.999$ and $\alpha_{\text{uv}} = 0.99$ for routing vectors. Reset is applied with probability $p_{\text{rst}} = 0.01$. Aug-Teacher uses color jitter and anchor threshold $\tau_{\text{ap}} = 1.0$. The scale-invariant consistency loss uses $\lambda = 0.85$ and $s = 10$. For augmentation, we set $K = 2$ for Hamlyn data, while we set $K = 6$ for C3VD data.

Table 1. Datsets Distribution

Source(Train)			**Target(Eval)**		
Name	**Scenario**	**Number**	**Name**	**Scenario**	**Number**
SCARED [1]	Abdomen	15351	Hamlyn [10]	Heart+Lung+Abdomen	91866
SimCol [9]	Simulation	28824	C3VD [2]	Colon Phantom	3166

Table 2. Zero-shot results on **Hamlyn dataset** by models trained on **SCARED dataset**. **Mem**: required GPU memory (G). *: Relative to the baseline (EndoDAC). **Best results** are in bold, while the second best results are underlined.

Methods	$Rel_{Abs}\downarrow$	$Rel_{Sq}\downarrow$	$RMSE\downarrow$	$RMSE_{Log}\downarrow$	δ<1.25↑	**Mem**
DA v2† $_{NeuIPS'24}$[17]	0.182	4.994	15.067	0.219	0.740	0.43
IID-SfMLearner $_{JBHI'24}$ [7]	0.171	4.526	14.066	0.206	0.767	0.12
MonoPCC $_{MedIA'25}$ [14]	0.158	3.889	13.205	0.194	0.782	0.29
EndoDAC $_{MICCAI'24}$ [3]	0.156	3.860	12.943	0.193	0.791	0.45
+CoMoDA$_{WACV'21}$ [6]	0.162	<u>3.511</u>	13.525	0.203	0.745	1.7
+CoTTA $_{CVPR'22}$ [13]	<u>0.152</u>	3.572	<u>12.567</u>	<u>0.188</u>	<u>0.794</u>	1.6
+AdaDepth $_{ICRA'23}$ [8]	0.161	3.498	13.478	0.202	0.749	1.7
+TEA $_{CVPR'24}$ [19]	0.162	3.512	13.526	0.203	0.745	1.9
+REM $_{ICML'25}$ [5]	0.176	4.342	14.669	0.378	0.731	2.3
+Ours	**0.142** ↓9.0%*	**3.003** ↓22.2%	**11.788** ↓8.9%	**0.177** ↓8.3%	**0.806** ↑1.9%	1.9

3.2 Comparison Results

Quantitative results from our method, existing TTA methods with EndoDAC and state-of-the-art endoscopy depth estimation are shown in Table 2 for laparoscopy and Table 3 for colonoscopy. Moreover, several samples are selected for qualitative comparison, results of which are shown as Fig. 3 and Fig. 4. Experimental results demonstrate that the proposed method outperforms existing methods for training-free depth estimation in unseen endoscopic scenes obtained from diverse subjects by different endoscopes.

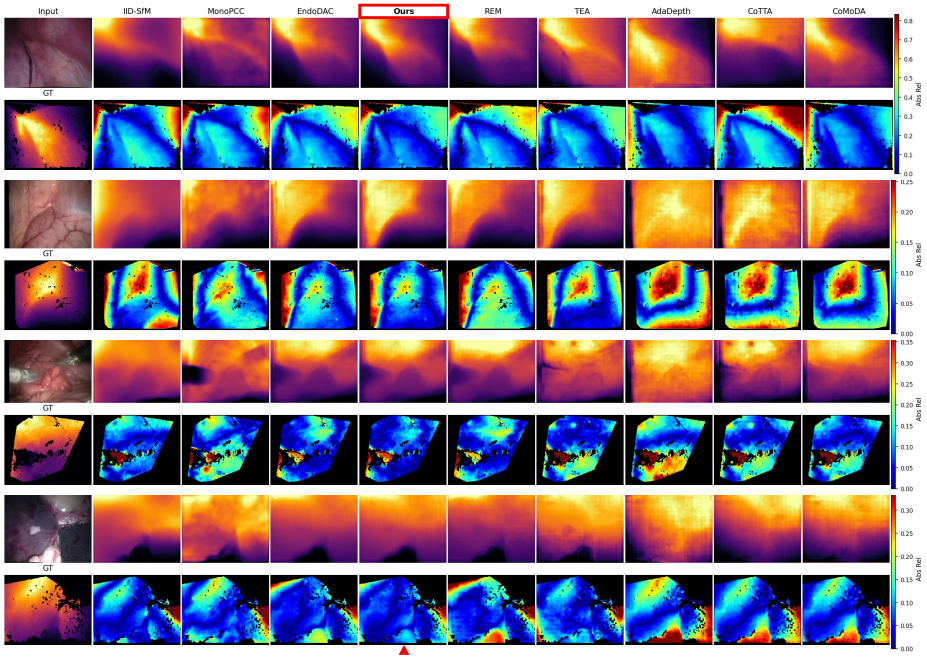


Fig. 3. Qualitative Comparison on Hamlyn Dataset

3.3 Ablation Studies

Ablation studies are performed to evaluate the effect of augmentation-averaged pseudo supervision (Aug-Teacher), stochastic parameter reset (Stoch-Reset) and scale-invariant consistency loss (SI Loss). To prove the effect of SI Loss, we utilize MSE Loss instead of SI Loss to perform the same evaluation procedure. Experimental results on laparoscopy and colonoscopy are shown in Table 4, which demonstrates the proposed test-time adaptation benefits from our three main contributions.

Table 3. Zero-shot results on **C3VD dataset** by models trained on **SimCol dataset**. **ms/fr**: inference time on per frame. *: Relative to the baseline (EndoDAC). **Best results** are in bold, while the second best results are underlined.

Methods	$Rel_{Abs}\downarrow$	$Rel_{Sq}\downarrow$	$RMSE\downarrow$	$RMSE_{Log}\downarrow$	$\delta\uparrow$	**ms/fr**
IID-SfMLearner $_{JBHI'24}$ [7]	0.140	1.484	7.733	0.186	0.810	4.0
EndoDAC $_{MICCAI'24}$ [3]	0.127	0.693	4.700	0.154	0.853	11.7
+CoMoDA$_{WACV'21}$ [6]	0.182	1.072	7.336	0.248	0.692	172.3
+CoTTA $_{CVPR'22}$ [13]	0.166	1.528	7.586	0.214	0.746	188.4
+AdaDepth $_{ICRA'23}$ [8]	0.143	0.895	4.810	0.210	0.750	139.7
+TEA $_{CVPR'24}$ [19]	0.183	1.094	7.360	0.249	0.694	588.3
+REM $_{ICML'25}$ [5]	0.185	1.710	7.444	0.276	0.690	114.0
+Ours	**0.118** ↓7.1%*	**0.630** ↓9.1%	**4.655** ↓0.96%	**0.145** ↓5.8%	**0.873** ↑2.3%	106.4

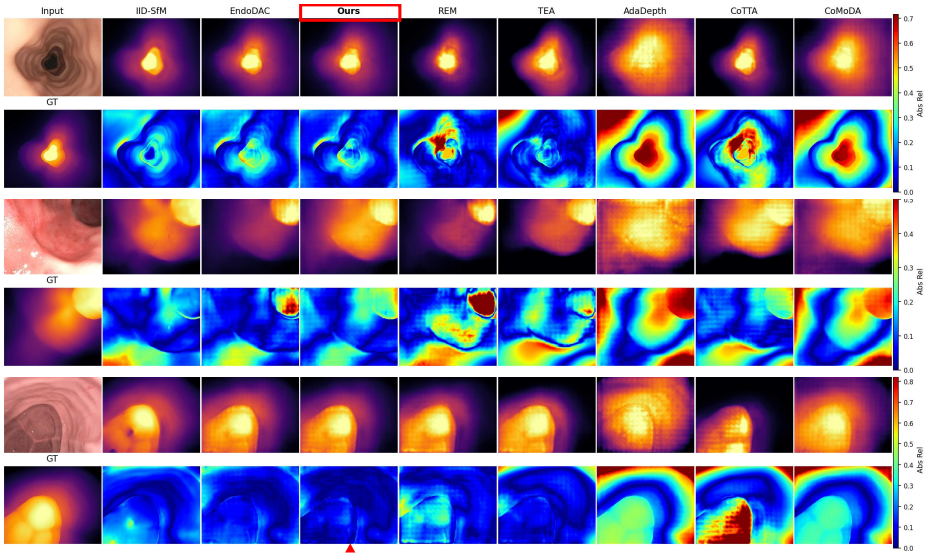


Fig. 4. Qualitative Comparison on C3VD Dataset

Table 4. Ablation studies on **C3VD/Hamlyn dataset** by models trained on **SimCol/SCARED dataset**. **Best results** are in bold.

Ablations	$Rel_{Abs}\downarrow$	$Rel_{Sq}\downarrow$	$RMSE\downarrow$	$RMSE_{Log}\downarrow$	$\delta\uparrow$
w/o SI Loss	0.125/0.150	0.721/3.317	4.951/12.234	0.151/0.187	0.861/0.784
w/o Stoch-Reset	0.121/0.155	0.811/3.384	5.490/12.796	0.155/0.192	0.859/0.776
w/o Aug-Teacher	0.127/0.147	0.693/3.269	4.700/12.120	0.154/0.182	0.853/0.800
Ours	**0.118/0.142**	**0.630/3.003**	**4.655/11.788**	**0.145/0.177**	**0.873/0.806**

4 Conclusion

This work proposes MEDE-TTA, a label-free continual test-time adaptation framework for cross-scenario monocular endoscopic depth estimation. The Aug-

Teacher generates robust pseudo-labels under endoscopic visual disturbances, while Stoch-Reset alleviates catastrophic parameter drift. Extensive zero-shot experiments on laparoscopy and colonoscopy datasets validate its state-of-the-art performance for MDE in unseen endoscopic environment against existing TTA and endoscopic depth estimation methods, facilitating practical unlabeled intraoperative 3D perception. However, the computational cost of the proposed method increases due to extra modules and optimization steps, but the efficiency of the proposed TTA method is still better than existing TTA methods.

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. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. 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. 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. 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. Han, J., Na, J., Hwang, W.: Ranked entropy minimization for continual test-time adaptation. In: International Conference on Machine Learning. pp. 21798–21813. PMLR (2025)
6. Kuznietsov, Y., Proesmans, M., Van Gool, L.: Comoda: Continuous monocular depth adaptation using past experiences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2907–2917 (2021)
7. Li, B., Liu, B., Zhu, M., Luo, X., Zhou, F.: Image intrinsic-based unsupervised monocular depth estimation in endoscopy. IEEE Journal of Biomedical and Health Informatics pp. 1–11 (2024)
8. Li, Z., Shi, S., Schiele, B., Dai, D.: Test-time domain adaptation for monocular depth estimation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4873–4879. IEEE (2023)
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. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J.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)
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. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
13. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7201–7211 (2022)
14. 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)
15. Xie, T., Yang, P., Jin, Y., Cai, Y., Yin, W., Ren, W., Zhang, Q., Hua, W., Peng, S., Guo, X., et al.: Scal3r: Scalable test-time training for large-scale 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21760–21771 (2026)
16. 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)
17. 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)
18. Yang, Z., Pan, J., Dai, J., Sun, Z., Xiao, Y.: Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. IEEE Transactions on Medical Imaging **43**(5), 1934–1944 (2024)
19. Yuan, Y., Xu, B., Hou, L., Sun, F., Shen, H., Cheng, X.: Tea: Test-time energy adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23901–23911 (2024)