EndoDA3: Unified Depth and Pose Estimation for Endoscopic Video

Jie Tan[1], Zijun Li[1], Hongmei Tuo[2], Yunhan Fu[3], Chengyang Li[1], and Li Lu[1]✉

[1] College of Computer Science, Sichuan University, Chengdu, China
[2] School of Gongli Hospital Medical Technology, University of Shanghai for Science and Technology, Shanghai, China
[3] Sichuan University-Pittsburgh Institute, Chengdu, China
luli@scu.edu.cn

Abstract. Three-dimensional (3D) scene reconstruction from monocular endoscopic videos is crucial for minimally invasive surgery, yet it remains challenging due to dynamic soft-tissue deformations, textureless surfaces, and illumination variations. Although recent self-supervised learning methods and adaptations of foundation models have shown promise, they often rely on complex multi-stage pipelines and external auxiliary networks for camera parameter estimation, while predominantly performing frame-wise inference, thereby neglecting inter-frame temporal consistency. To address these limitations, we propose EndoDA3, a unified and parameter-efficient framework that leverages the robust geometric priors of Depth Anything 3 for continuous endoscopic 3D reconstruction without requiring ground-truth depth. Specifically, we introduce a novel Dual-Vector Modulated Low-Rank Adaptation (DVM-LoRA) strategy to efficiently adapt the foundation model for joint scale-consistent depth prediction and camera pose estimation within a single shared network. Furthermore, to explicitly enforce temporal continuity and mitigate the adverse effects of dynamic outliers, we design an Uncertainty-aware Projection Loss that dynamically reweights projection-based training signals and suppresses unreliable regions caused by moving instruments and non-rigid tissue deformation. Extensive experiments demonstrate that EndoDA3 achieves strong performance across two public endoscopic datasets, with improved temporal consistency and robust camera pose estimation. Code is available at EndoDA3.

Keywords: Self-supervised learning · Endoscopic surgery · Video depth estimation.

1 Introduction

Three-dimensional (3D) scene reconstruction from monocular endoscopic videos plays a vital role in minimally invasive surgery (MIS). Recovering the 3D geometry of the surgical site can provide depth cues that are unavailable in standard 2D endoscopic videos. Such spatial awareness is essential for improving lesion

sizing, facilitating intraoperative navigation, and enabling robotic surgical automation [5,8,25].

However, recovering scale-consistent 3D structures from endoscopic scenes remains challenging. The surgical environment is highly complex, characterized by dynamic soft-tissue deformations, textureless surfaces, and illumination variations [8]. Consequently, conventional Structure-from-Motion [14] and Simultaneous Localization and Mapping [3] methods, which heavily rely on static scene assumptions, often yield suboptimal performance and catastrophic tracking failures in deformable surgical environments.

While deep learning methods [21,22] have significantly advanced depth estimation in natural scenes, acquiring ground-truth depth in endoscopy remains intrinsically challenging. To this end, self-supervised learning (SSL) leveraging adjacent-frame geometric consistency has emerged as a prevalent paradigm [2,6]. Yet, recent SfM-like multi-stage pipelines [15,23] still suffer from error accumulation and cumbersome offline optimization. Meanwhile, foundation models such as Depth Anything [4,10,21,22] and DUSt3R series [17,18] have shown strong zero-shot generalization in general domains, but adapting them to complex endoscopic scenes entails overcoming a substantial domain gap. To bridge this gap, recent methods [6,16,24] employ Low-Rank Adaptation (LoRA) [9] to adapt these generic models to surgical videos. However, these adaptations typically rely on external auxiliary networks or explicit prior knowledge to estimate camera parameters, or they fine-tune separate, redundant encoders for depth and pose regression. More critically, most of these methods perform isolated frame-wise inference, overlooking the rich temporal continuity inherent in surgical videos.

To address these limitations, we propose **EndoDA3**, a unified and parameter-efficient framework that adapts Depth Anything 3 (DA3) to continuous endoscopic 3D reconstruction without requiring ground-truth depth or an additional pose network. Specifically, we introduce a novel parameter-efficient fine-tuning strategy named **Dual-Vector Modulated Low-Rank Adaptation (DVM-LoRA)**. By training a minimal set of parameters, EndoDA3 enables the simultaneous regression of scale-consistent depth and reliable camera poses within a single network. Furthermore, to ensure temporal continuity and mitigate the adverse effects of tissue deformation and instrument motion, we introduce an **Uncertainty-aware Projection Loss**. This objective dynamically reweights supervision signals based on prediction confidence, effectively filtering out outliers induced by dynamic instruments and deformable tissues, thereby facilitating robust learning in complex surgical environments. Our main contributions are summarized as follows.

1. We propose EndoDA3, a unified endoscopic reconstruction framework derived from Depth Anything 3. By introducing a novel Dual-Vector Modulated LoRA (DVM-LoRA) and a shared encoder backbone, we adapt the foundation model to jointly infer video depth and camera parameters with high efficiency.
2. We introduce an Uncertainty-aware Projection Loss to enforce temporal continuity and improve robustness against dynamic outliers. By weighting the

reprojection depth error using confidence maps, this loss explicitly mitigates the impact of non-rigid tissue deformation and instrument motion.
3. Extensive experiments on two public endoscopic datasets demonstrate the effectiveness of EndoDA3 in improving video depth prediction, temporal consistency, and camera pose estimation.

2 Method

2.1 EndoDA3 Framework

Network Architecture Recognizing that depth estimation and ego-motion are inherently coupled tasks rather than independent processes, we introduce a unified architecture to simultaneously predict depth and camera parameters. To establish a robust geometric representation for endoscopic reconstruction, our framework builds upon the Depth Anything 3 [10] foundation model, aiming to exploit its geometric priors for spatially consistent 3D reconstruction. As illustrated in Fig. 1, we extend the input formulation from image pairs to video sequences to better capture temporal context.

Specifically, given a video sequence of T input images $\mathcal{I} = \{I_i\}_{i=1}^{T}$, a single shared ViT encoder [10,11] employing input-adaptive "within-view" and "cross-view" attention mechanisms is utilized to extract unified features. These features are subsequently decoded by two specialized branches. In the depth estimation branch, we employ a Multi-Scale Dense Prediction Transformer (DPT) head [10,12] to capture both global contextual information and local structural details. The feature representations are progressively reassembled and fused across different resolutions to predict pixel-aligned scalar depth maps $\mathbf{D}_i \in \mathbb{R}^{H\times W\times 1}$, alongside confidence maps $\mathbf{C}_i \in \mathbb{R}^{H\times W\times 1}$ that quantify the model's pixel-wise reliability regarding the depth predictions. Concurrently, the pose branch utilizes a lightweight regression head $\mathcal{D}_C$ to estimate explicit camera parameters. These parameters are then converted into standard rotation matrices $\mathbf{R}_i$, translation vectors $\mathbf{t}_i$, and intrinsic matrices $\mathbf{K}_i$.

Self-Supervised Learning Strategy Following the self-supervised learning (SSL) paradigm [6,15], the predicted depth and camera parameters are utilized to synthesize the target frame I_t by inversely warping the adjacent source views I_s ($s \in \{t-1, t+1\}$). Accordingly, a pixel $\mathbf{p}_t = [u, v, 1]^\top$ in the target image I_t is first back-projected to a 3D point $\mathbf{P}$ in the world coordinate system using the standard pinhole camera model:

$$\mathbf{P} = \mathbf{R}_t \left(\mathbf{D}_t(u,v)\mathbf{K}_t^{-1}\mathbf{p}_t\right) + \mathbf{t}_t. \tag{1}$$

Subsequently, these 3D coordinates are transformed into the source camera coordinate system and projected onto the source view I_s to obtain continuous 2D sampling coordinates. The synthesized target image $I_{s\to t}$ is then generated by differentiably sampling the color values from I_s at these projected coordinates via bilinear interpolation. Finally, we adopt the photometric reconstruction loss

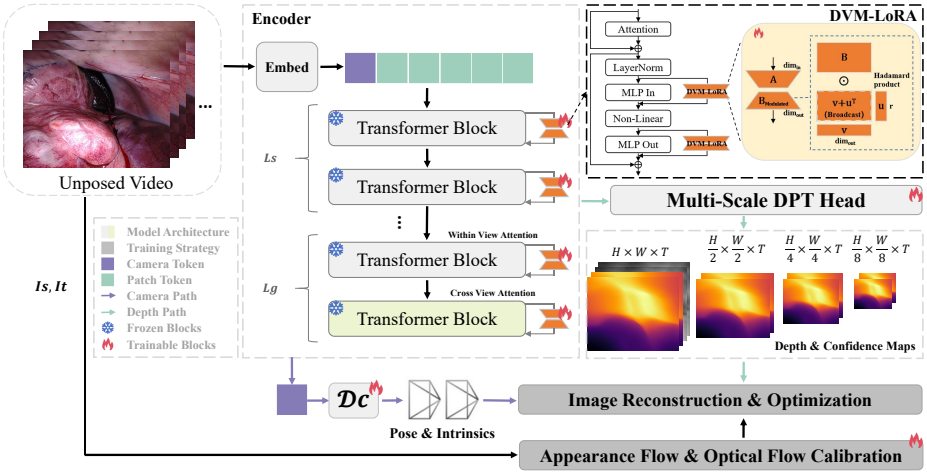

Fig. 1: Overview of the EndoDA3 self-supervised learning framework for depth and pose estimation, trained with multi-scale SSIM-based photometric reconstruction loss and the proposed uncertainty-aware projection loss.

$\mathcal{L}_p$ [6] and incorporate the illumination-compensation strategy of [15] to mitigate appearance inconsistencies, thereby encouraging $I_{s\to t}$ to align as closely as possible with the original I_t.

2.2 Dual-Vector Modulated Low-Rank Adaptation

The inherent domain gap frequently leads to performance degradation in the zero-shot depth estimation of DA3 in endoscopic scenes. This gap is highly heterogeneous in endoscopic videos, where specular highlights, weak textures, occlusions, and tissue deformation affect feature representations unevenly across channels. To address this challenge, we propose Dual-Vector Modulated Low-Rank Adaptation (DVM-LoRA), an efficient fine-tuning strategy that adapts robust geometric priors while preserving the generalizable representations of the pre-trained model.

Standard LoRA [9] adapts pre-trained models by introducing trainable low-rank update matrices while keeping the original weights frozen. Let $\mathbf{W}_0 \in \mathbb{R}^{d_{out}\times d_{in}}$ be a pre-trained weight matrix and $\mathbf{x} \in \mathbb{R}^{d_{in}}$ be the input. Standard LoRA modifies the forward pass as $\mathbf{h} = \mathbf{W}_0\mathbf{x} + s\mathbf{BAx}$, where $\mathbf{B} \in \mathbb{R}^{d_{out}\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times d_{in}}$ are low-rank matrices ($r \ll \min(d_{in}, d_{out})$), and $s = \frac{\alpha}{r}$ serves as a constant scaling factor.

Although efficient, standard LoRA controls the magnitude of the low-rank branch using a single global scaling factor, without explicitly modeling rank-wise or output-channel-wise modulation. This may limit its flexibility when adapting to heterogeneous domain shifts in endoscopic scenes. To overcome this limitation, DVM-LoRA introduces two learnable modulation vectors at both the intermediate low-rank space and the output feature space. Specifically, we retain $\mathbf{A}$ and

$\mathbf{B}$ and introduce two trainable column vectors: $\mathbf{u} \in \mathbb{R}^{r\times 1}$ and $\mathbf{v} \in \mathbb{R}^{d_{out}\times 1}$. The rank-wise vector $\mathbf{u}$ adjusts the contribution of individual low-rank components, while the output-channel-wise vector $\mathbf{v}$ provides channel-selective adaptation for the resulting feature updates. Conceptually, the weight update is modulated as follows:

$$\Delta\mathbf{W} = s\left(\mathbf{B}\mathrm{diag}(\mathbf{u})\mathbf{A} + \mathrm{diag}(\mathbf{v})\mathbf{B}\mathbf{A}\right). \tag{2}$$

In practice, to avoid the computational overhead of explicitly constructing diagonal matrices during the forward pass, we directly fuse the dual-modulation pathways within the weight space. By leveraging matrix broadcasting, the effective weight update $\Delta\mathbf{W}$ is efficiently formulated as:

$$\Delta\mathbf{W} = s\left(\mathbf{B} \odot (\mathbf{v}\mathbf{1}_r^\top + \mathbf{1}_{d_{out}}\mathbf{u}^\top)\right)\mathbf{A}, \tag{3}$$

where $\odot$ represents the Hadamard product, and $\mathbf{1}$ denotes a column vector of ones. The forward pass is then executed via a single unified matrix multiplication: $\mathbf{h} = (\mathbf{W}_0 + \Delta\mathbf{W})\mathbf{x}$. This design obviates the need to instantiate redundant activation tensors and allows for seamless zero-inference-overhead weight merging.

Regarding parameter initialization, $\mathbf{A}$, $\mathbf{u}$, and $\mathbf{v}$ are initialized using Kaiming uniform initialization, whereas $\mathbf{B}$ is strictly zero-initialized. This configuration guarantees that $\Delta\mathbf{W}$ remains zero at the onset of fine-tuning, thereby preserving the original behavior of the pre-trained model and ensuring training stability. DVM-LoRA introduces only $r + d_{out}$ additional parameters beyond standard LoRA, preserving parameter efficiency while enabling flexible adaptation to heterogeneous endoscopic domain gaps. In our architecture, DVM-LoRA is exclusively applied to the feed-forward networks (FFNs) within the Transformer blocks.

2.3 Uncertainty-aware Projection Loss

Standard self-supervised depth estimation frameworks often lack explicit temporal constraints, leading to temporally inconsistent predictions [6,24]. To address this, we introduce a temporal depth consistency constraint [25]. Given a source frame I_s and an adjacent target frame I_t, our network predicts their corresponding depth maps $\mathbf{D}_s$ and $\mathbf{D}_t$, alongside the relative camera pose $\mathbf{T}_{t\to s}$. Using the predicted pose and camera intrinsics, we project the target depth into the source viewpoint to synthesize the warped target depth $\hat{\mathbf{D}}_{t\to s}$. The base geometric consistency is enforced by minimizing the pixel-wise depth error, defined as:

$$e_{\mathbf{p}} = |\mathbf{D}_s(\mathbf{p}) - \hat{\mathbf{D}}_{t\to s}(\mathbf{p})|. \tag{4}$$

However, in endoscopic videos, the multi-view consistency assumption is frequently violated by tissue deformations and dynamic surgical instruments. Inspired by robust alignment strategies in recent reconstruction models [7], which down-weight dynamic outliers through confidence-aware filtering, we propose an uncertainty-aware projection loss tailored for surgical environments. By dynamically assigning lower confidence values to regions affected by rapid instrument

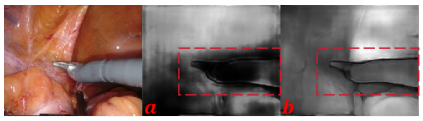

Fig. 2: Visualization of confidence maps for an example from [19]. Brighter regions correspond to higher confidence in the depth estimation. (*a*) predicted by DA3; (*b*) predicted by EndoDA3 with LoRA.

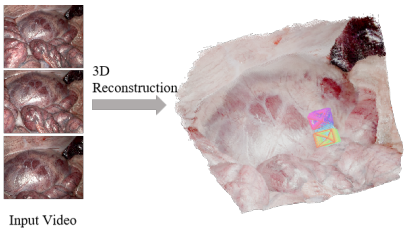

Fig. 3: Qualitative results of 3D reconstruction.

motion or severe tissue deformation, we attenuate their misleading contributions to the alignment process, while concentrating higher weights on relatively static anatomical backgrounds. This loss adaptively weights the geometric depth consistency using a network-predicted confidence map $\mathbf{C}$. The proposed uncertainty-aware projection loss is formulated as:

$$\mathcal{L}_{\text{conf}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \left(\mathbf{C}_{\mathbf{p}} \cdot \mathcal{L}_h(e_{\mathbf{p}}, \delta_h) - \lambda \log(\mathbf{C}_{\mathbf{p}}) \right), \tag{5}$$

where Ω denotes the set of valid projected pixels, and $\mathcal{L}_h(\cdot, \delta_h)$ is the Huber loss with a threshold δ_h to further suppress extreme errors.

The first term in Eq. 5 penalizes depth inconsistencies weighted by the predicted confidence, effectively encouraging the network to assign low confidence (i.e., high uncertainty) to unreliable regions. The second regularization term, $-\lambda \log(\mathbf{C}_{\mathbf{p}})$, prevents the network from converging to the trivial solution of predicting minimal confidence everywhere. The overall training objective is formulated as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{EndoDAC}} + \lambda_{\text{conf}} \mathcal{L}_{\text{conf}}, \tag{6}$$

where $\mathcal{L}_{\text{EndoDAC}}$ denotes the original self-supervised training objective of EndoDAC [6], and λ_{conf} controls the contribution of the proposed uncertainty-aware projection loss. In our implementation, we empirically set $\lambda_{\text{conf}} = 0.01$ and the Huber threshold $\delta_h = 1.0$. As shown in Fig. 2, the proposed loss assigns lower confidence to regions with moving instruments and tissue deformation.

3 Experiments and Results

3.1 Datasets and Evaluation

We evaluate our proposed method on two publicly available endoscopic and laparoscopic video datasets: the SCARED dataset and the Hamlyn dataset. **SCARED Dataset.** The SCARED [1] dataset provides a controlled environment for surgical data analysis. It comprises 35 endoscopic videos captured using a da Vinci Xi endoscope, yielding a total of 22,950 high-resolution frames.

Table 1: Quantitative video depth comparison on SCARED and Hamlyn datasets. The best results are in bold. "Total.(M)" and "Train.(M)" refer to the total and trainable parameters utilized in the depth network. Note that since the Hamlyn dataset does not provide the camera pose annotations, we do not evaluate the TAE metric on it.

	Methods	Year	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	TAE↓	Total.(M)	Train.(M)	Speed (ms)
SCARED	VDA [4]	2025	0.241	7.702	18.673	0.287	0.597	1.100	113.1	-	15.9
	DA3 [10]	2025	0.110	1.699	9.169	0.139	0.858	1.305	103.8	-	19.4
	EndoDAC [6]	2024	0.146	2.389	11.863	0.178	0.790	1.843	99.0	1.6	15.0
	EndoDAV [25]	2025	0.160	2.951	12.783	0.187	0.759	0.725	113.3	0.19	16.0
	EndoDA3	-	**0.088**	**1.014**	**7.002**	**0.110**	**0.933**	**0.335**	104.5	0.78	19.5
Hamlyn	DA3 [10]	2025	0.321	12.021	22.467	0.681	0.500	-	103.8	-	19.4
	EndoDAV [25]	2025	0.223	6.277	16.329	**0.281**	0.637	-	113.3	0.19	16.0
	EndoDA3	-	**0.220**	**6.224**	**16.095**	0.328	**0.648**	-	104.5	0.78	19.5

Table 2: Comparison of Pose Estimation on the SCARED Dataset.

Methods	ATE↓
EndoDAC [6]	0.124
DA3 [10]	0.208
Endo3R [8]	0.112
EndoDA3	**0.102**

Table 3: Ablation study on SCARED dataset. The best results are in bold. Specifically, we (i) use the original LoRA [9] to replace DVM-LoRA; (ii) remove Uncertainty-aware Projection Loss.

$\mathcal{L}_{\text{conf}}$	DVM-LoRA	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	TAE↓
×	×	0.099	1.252	7.711	0.129	0.898	0.538
✓	×	0.095	1.212	7.518	0.120	0.908	0.376
✓	✓	**0.088**	**1.014**	**7.002**	**0.110**	**0.933**	**0.335**

The dataset provides ground-truth depth maps generated via a structured light projector, alongside accurate camera poses and intrinsic parameters. Following previous work [25], the dataset is partitioned into 24 sequences for training, 3 for validation, and 8 for testing.

Hamlyn Dataset. The Hamlyn dataset offers a diverse collection of laparoscopic and endoscopic surgical videos. In our experiments, 21 selected video sequences [13] from this dataset are utilized for cross-dataset zero-shot validation.

Evaluation Metrics. Following [6,15,25], we utilize five standard metrics to assess the spatial accuracy of the depth predictions: Abs Rel, Sq Rel, RMSE, RMSE log, and $\delta < 1.25$. Furthermore, to quantitatively measure the temporal consistency of the video depth estimation, we report the Temporal Alignment Error (TAE) [20]. To evaluate the pose accuracy, we perform a 5-frame pose evaluation and adopt the Absolute Trajectory Error (ATE). Critically, for affine-invariant depth evaluation, depth predictions must be aligned to the ground truth. We perform alignment using the median and scale computed over the full video sequence [4], rather than per-frame alignment [6]. This rigorous sequence-level alignment is crucial to ensure that temporal consistency is accurately evaluated.

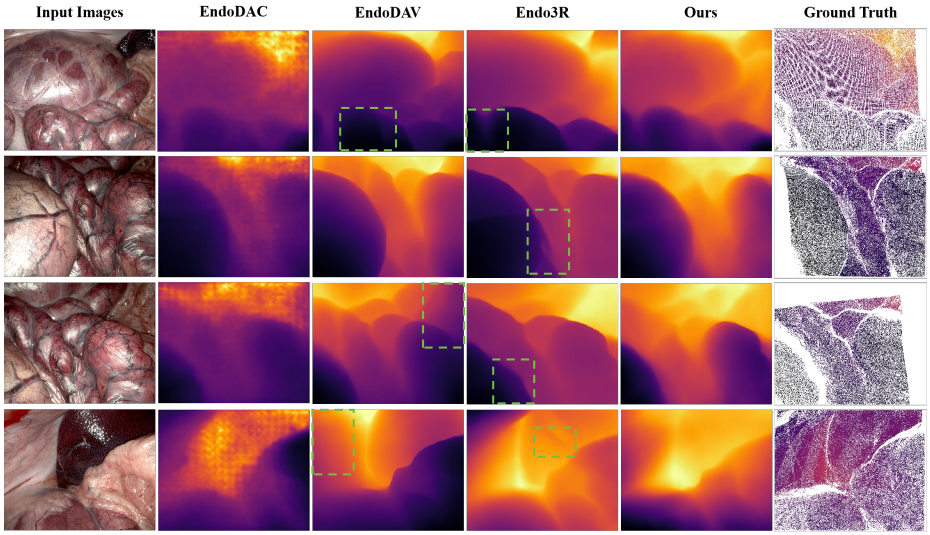

Fig. 4: Qualitative results using video sequences as input.

3.2 Experimental Results

Implementation Details. Implemented in PyTorch on a single RTX 3090 GPU, the framework is trained for 20 epochs with a sequence length of 16. We use the AdamW optimizer (initial learning rate 1×10^{-4}) and adopt standard data augmentations [6,25]. The DVM-LoRA rank is set to $r = 4$.

Quantitative Analysis. We quantitatively evaluate our proposed method on the SCARED and Hamlyn datasets against several competitive baselines. Some results are cited from the original papers. As shown in Table 1 and Table 2, our approach improves both video depth and pose prediction accuracy in complex endoscopic scenes, demonstrating high spatial accuracy and robust temporal consistency. Furthermore, benefiting from a parameter-efficient design, our method achieves an inference speed comparable to the aforementioned baselines, fully supporting real-time depth estimation.

Qualitative Analysis. We present qualitative results for video depth estimation in Fig. 4, which demonstrates that EndoDA3 can generate more accurate depth maps and better relative scale. In addition, as shown in Fig. 3, our accurate depth and pose estimates enable high-quality 3D reconstruction. For more visualization results, please refer to the supplementary video.

Ablation Studies. To validate the efficacy of individual components, we conducted ablation studies on the SCARED dataset, as shown in Table 3. Experimental results demonstrate that our proposed EndoDA3 effectively enhances both depth prediction accuracy and temporal consistency across videos. Notably, the Uncertainty-aware Projection Loss improves temporal continuity, while the proposed DVM-LoRA further enhances the overall depth estimation performance.

4 Conclusion

In this paper, we proposed EndoDA3, a unified and parameter-efficient framework for continuous 3D scene reconstruction from monocular endoscopic videos. By integrating a novel Dual-Vector Modulated LoRA (DVM-LoRA) with an Uncertainty-aware Projection Loss, our approach effectively adapts foundation models to complex surgical environments while maintaining robust temporal consistency against tissue deformations and instrument motion. Extensive experiments on two public datasets demonstrate that EndoDA3 achieves superior scale-consistent depth and accurate camera pose estimation with a single shared encoder. Future work will explore real-time intraoperative optimization and downstream integration with robotic navigation systems.

Acknowledgments. This study was funded by Chengdu Science and Technology Program (2026-YF05-00762-SN).

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. Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N.: Monocular depth estimation: A thorough review. IEEE Transactions on Pattern Analysis & Machine Intelligence **46**(04), 2396–2414 (2024)
3. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE transactions on robotics **37**(6), 1874–1890 (2021)
4. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22831–22840 (2025)
5. Collins, T., Pizarro, D., Gasparini, S., Bourdel, N., Chauvet, P., Canis, M., Calvet, L., Bartoli, A.: Augmented reality guided laparoscopic surgery of the uterus. IEEE Transactions on Medical Imaging **40**(1), 371–380 (2020)
6. 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)
7. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. arXiv preprint arXiv:2507.16443 (2025)
8. 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)

9. 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)
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. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
12. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
13. 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)
14. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
15. 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)
16. 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)
17. 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)
18. 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)
19. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International conference on medical image computing and computer-assisted intervention. pp. 431–441. Springer (2022)
20. Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., He, T.: Depth any video with scalable synthetic data. arXiv preprint arXiv:2410.10815 (2024)
21. 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)
22. 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)
23. 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)
24. Zeinoddin, M.S., Lena, C., Qu, J., Carlini, L., Magro, M., Kim, S., De Momi, E., Bano, S., Grech-Sollars, M., Mazomenos, E., et al.: Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model. arXiv preprint arXiv:2408.17433 (2024)

25. Zhou, Z., Yang, C., Yang, P., Yang, X., Shen, W.: Endodav: Depth any video in endoscopy with spatiotemporal accuracy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201. Springer (2025)