Endo3R: Unified Online Reconstruction from Dynamic Monocular Endoscopic Video

Jiaxin Guo[1], Wenzhen Dong[2], Tianyu Huang[1,2], Hao Ding[3], Ziyi Wang[1], Haomin Kuang[4], Qi Dou[1], and Yun-Hui Liu[1,2](✉)

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Hong Kong Center for Logistics Robotics, Hong Kong SAR, China
[3] Johns Hopkins University, Baltimore MD 21218, USA
[4] Shanghai Jiao Tong University, Shanghai, China

Abstract. Reconstructing 3D scenes from monocular surgical videos can enhance surgeon's perception and therefore plays a vital role in various computer-assisted surgery tasks. However, achieving scale-consistent reconstruction remains an open challenge due to inherent issues in endoscopic videos, such as dynamic deformations and textureless surfaces. Despite recent advances, current methods either rely on calibration or instrument priors to estimate scale, or employ SfM-like multi-stage pipelines, leading to error accumulation and requiring offline optimization. In this paper, we present Endo3R, a unified 3D foundation model for online scale-consistent reconstruction from monocular surgical video, without any priors or extra optimization. Our model unifies the tasks by predicting globally aligned pointmaps, scale-consistent video depths, and camera parameters without any offline optimization. The core contribution of our method is expanding the capability of the recent pairwise reconstruction model to long-term incremental dynamic reconstruction by an uncertainty-aware dual memory mechanism. The mechanism maintains history tokens of both short-term dynamics and long-term spatial consistency. Notably, to tackle the highly dynamic nature of surgical scenes, we measure the uncertainty of tokens via Sampson distance and filter out tokens with high uncertainty. Regarding the scarcity of endoscopic datasets with ground-truth depth and camera poses, we further devise a self-supervised mechanism with a novel dynamics-aware flow loss. Abundant experiments on SCARED and Hamlyn datasets demonstrate our superior performance in zero-shot surgical video depth prediction and camera pose estimation with online efficiency. Project page: https://wrld.github.io/Endo3R/.

Keywords: 3D foundation model · Video depth estimation · 3D Reconstruction · Pose estimation · Endoscopic surgery.

1 Introduction

Reconstructing surgical scenes from endoscopic videos is crucial for minimally invasive surgery, benefiting various downstream tasks including surgical planning, intraoperative navigation, and robotic surgical automation [24, 35]. This

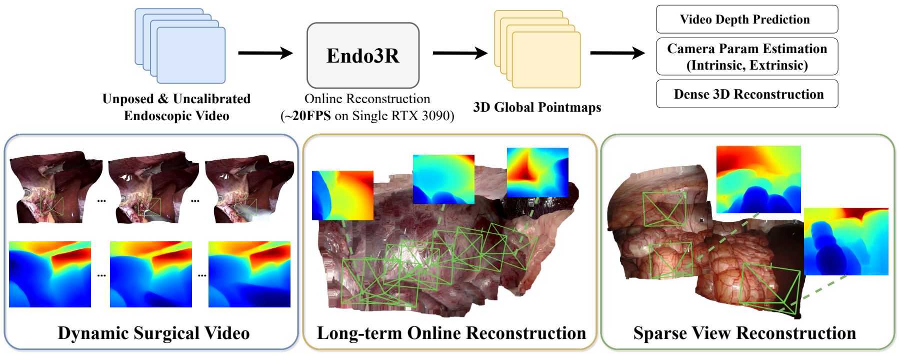
Unposed & Uncalibrated Endoscopic Video
Endo3R
Online Reconstruction (~20FPS on Single RTX 3090)
3D Global Pointmaps
Video Depth Prediction
Camera Param Estimation (Intrinsic, Extrinsic)
Dense 3D Reconstruction
Dynamic Surgical Video
Long-term Online Reconstruction
Sparse View Reconstruction

Fig. 1. Given monocular surgical video as input, our Endo3R allows feed-forward output of global pointmaps, scale-consistent depth, and camera parameters.

topic has been studied for decades, with relevant areas including depth estimation [11,26], pose estimation [9,12], multi-view stereo (MVS) [33,36], novel view synthesis (NVS) [4,13,14], Structure-from-Motion (SfM) [15], and Simultaneous localization and mapping (SLAM) [6,38].

However, estimating scale-consistent 3D structures from dynamic monocular surgical video remains a challenging and ill-posed problem. This challenge arises from sparse features, the lack of multi-view constraints, and the complexity of the surgical environment, which involves factors such as illumination variance, homogeneous surfaces, motion blur, and dynamic deformations from surgical interventions. Traditional methods [8, 15], which are developed under the assumption of rigid scenes, struggle to extract reliable features and match correspondences across frames in such dynamic environments. Although recent monocular depth foundation models [11,26] have made significant progress, they degrade when applied to surgical scenes and fail to predict accurate relative geometry. Some methods attempt to transfer general-domain models to surgical video, but they either require prior information (e.g., camera parameters or instrument models) [27,28,30], or adopt an SfM-like multi-stage pipeline to learn both motion and geometry by estimating correspondences, camera poses and intrinsics for higher relative scale consistency [3, 30]. Moreover, such SfM-like multi-stage pipeline will accumulate errors in every stage or require offline optimization, leading to sub-optimal accuracy and consistency.

In this paper, we address these challenges and present **Endo3R**, a unified 3D surgical foundation model for online scale-consistent reconstruction from monocular endoscopic video *without any prior information or extra optimization*, predicting globally aligned pointmaps, scale-consistent video depth, camera poses and intrinsics, as shown in Fig. 1. The key contribution of our method is devising an uncertainty-aware dual memory mechanism to expand the pairwise reconstruction ability from DUSt3R [31] to long-term incremental dynamic reconstruction, by capturing both short-term dynamics and long-term spatial

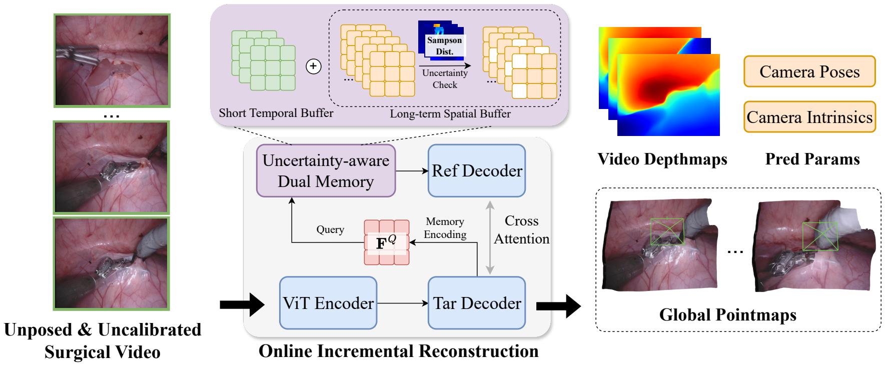

Fig. 2. Overview of Endo3R. Given monocular surgical video as input, we present a 3D surgical foundation model to enable online reconstruction from video.

memory. We employ a memory encoder to save history tokens as memory keys and values, retrieving the relative information by cross attention. We measure the uncertainty of tokens by calculating the Sampson distance and filtering out tokens with high uncertainty. Regarding the lack of training datasets, we introduce a self-supervised training scheme for data without ground-truth poses and depths. Namely, a dynamics-aware flow loss is designed to enforce the cross-frame temporal consistency.

Our contribution is summarized as follows: **1**) We present Endo3R, a 3D surgical foundation model to enable real-time reconstruction from monocular video, unifying the prediction of globally aligned pointmaps, scale-consistent video depth, camera poses, and intrinsics. **2**) We present an uncertainty-aware dual memory mechanism to enable long-term online dynamic reconstruction. **3**) A self-supervised scheme is introduced to allow for scaling to more surgical datasets without ground truth. **4**) Experimental results demonstrate our superior performance in video depth estimation and pose estimation with online efficiency.

2 Methodology

In this paper, we aim to build a unified framework to solve online 3D reconstruction from endoscopic video, by adapting the static pairwise reconstruction from DUSt3R to long-term endoscopic videos. To enhance the robustness in long-term learning, our main insight is to enable incremental online reconstruction by an uncertainty-aware dual memory mechanism, predicting globally aligned pointmaps, temporally consistent video depth, camera poses, and intrinsics. Due to the scarcity of surgical datasets, we further employ a hybrid training mechanism and devise a flow-guided self-supervised learning to help scale up our network to more surgical datasets with different scenes.

As shown in Fig. 2, given a sequence of images $\{\mathbf{I}_i\}_{i=1}^{N} \in \mathbb{R}^{W\times H\times 3}$ as input, our goal is to train a network $\mathcal{F}$ to output the corresponding pointmaps

$\{\mathbf{X}_{i,1}\}_{i=1}^{N} \in \mathbb{R}^{W\times H\times 3}$, confidence maps $\{\mathbf{C}_{i,1}\}_{i=1}^{N} \in \mathbb{R}^{W\times H}$ in the coordinate frame of $t = 1$ (Sec. 2.1). To tackle the highly dynamic long-term surgical video, we incorporate an uncertainty-aware spatial-temporal memory to preserve history tokens, capturing both short-term dynamics and long-term spatial consistency (Sec. 2.2). Regarding the scarcity of surgical datasets with groundtruth, we design a self-supervised flow loss by decomposing the optical flow into scene flow and camera projection flow, enforcing the scale continuity and pose smoothness between consecutive frames (Sec. 2.3).

2.1 Network architecture

Encoding. Given a sequence of images as input, our network first encodes every frame $\mathbf{I}_i$ into tokens $\mathbf{F}_i$ with a ViT encoder [1]: $\mathbf{F}_i = \text{Encoder}(\mathbf{I}_i)$.

Memory retrieval. Unlike [31] conducting pairwise prediction, we enable the incremental reconstruction by incorporating an attention-based memory mechanism inspired by Spann3R [10]. The memory bank stores the historical key features and value features. Given every new frame, we leverage a previous query feature $\mathbf{F}_{t-1}^{Q}$ to retrieve relative contexts from the memory bank to output the fused tokens $\mathbf{F}_{t-1}^{G}$:

$$\mathbf{F}_{i-1}^{G} = \text{Softmax}(\frac{\mathbf{F}_{i-1}^{Q}(\mathbf{F}^{K})^{\mathrm{T}}}{\sqrt{C}})\mathbf{F}^{V} + \mathbf{F}_{i-1}^{Q}, \tag{1}$$

where $\mathbf{F}^{K}$ and $\mathbf{F}^{V}$ are key and value features saved in the memory bank.

Decoding. After encoding, two transformer decoders sequentially perform self-attention and cross-attention on both encoded feature $\mathbf{F}_t$ and fused feature $\mathbf{F}_{t-1}^{G}$ to predict the 3D geometry: $\mathbf{F'}_i, \mathbf{F'}_{i-1}^{G} = \text{Decoder}(\mathbf{F}_i, \mathbf{F}_{i-1}^{G})$, where $\mathbf{F'}_i$ and $\mathbf{F'}_{i-1}^{G}$ denote the features after the cross-view interaction.

Regression Head. After decoding, the 3D representations are predicted from the decoded features. Following [31], we employ DPT [26] head to predict the 3D pointmap and associated confidence map. We compute the camera pose $\hat{\mathbf{T}}_{i,1}$ based on PnP. Then the depth $\hat{\mathbf{D}}_i$ could be estimated by transforming the global pointmap to the local coordinate with $\hat{\mathbf{T}}_{i,1}$:

$$\hat{\mathbf{X}}_{i,1}, \hat{\mathbf{C}}_{i,1} = \text{Head}_{\text{output}}(\mathbf{F'}_i), \quad \hat{\mathbf{D}}_i = (\hat{\mathbf{T}}_{i,1}\hat{\mathbf{X}}_{i,1})_z. \tag{2}$$

2.2 Uncertainty-aware Dual Memory

Dual Memory. To extend [31] to sequential reconstruction, we introduce an uncertainty-aware dual memory mechanism consisting of a long-term spatial buffer and short-term temporal buffer. Namely, global keyframe tokens and stable 3D information are stored in the long-term spatial buffer, maintaining spatial consistency over time. The short-term temporal buffer stores tokens from the recent frames, ensuring temporal consistency across consecutive frames.

Memory Encoding. At the end of each step, the decoded feature $\mathbf{F'}_i$ and encoded feature $\mathbf{F}_i$ are used to generate the query feature for the next step. The information of the current frame is preserved in the short-term temporal buffer of the memory bank as key and value features. When more frames come in, the older memory keys and values will be moved to the long-term spatial buffer.

Uncertainty Check. Unlike static reconstruction, dynamic surgical scenes present additional challenges, e.g., non-rigid tissue deformations, surgical instruments frequently appearing and disappearing, and occlusions due to sudden camera movements or interactions with anatomical structures. Therefore, we aim to filter the memory bank to eliminate the 3D information of transient objects and occlusions, to enhance the global 3D consistency and robustness for the new incoming frames. To filter out dynamic tokens and disturbances, we use the Sampson distance to assess the reliability of the tokens stored in the long-term spatial memory. We follow [13] to leverage the optical flow $\mathbf{O}_{i\to i+1}$ to assess the epipolar geometry with the estimated poses $\hat{\mathbf{T}}_i$ and $\hat{\mathbf{T}}_{i+1}$. Therefore, given every encoded memory as input, the tokens with high Sampson distance (i.e., larger than threshold β) indicate unreliable matches and will be eliminated from the memory bank $\mathbf{F}^K$ and $\mathbf{F}^Q$. For long sequence inference, we leverage confidence map $\mathbf{C}$ to select top K tokens in the memory bank and prune the others.

2.3 Self-supervised Losses

Despite the success of DUSt3R-related methods, they require supervised training on large-scale datasets with both GT depth and poses. However, in surgical scenes, there are limited datasets containing the GT depth and poses, which hinders the training for diverse scenes or surgeries with monocular videos only. To address this problem, we propose a self-supervised training scheme that enables training on datasets without full labels.

Dynamics-aware Flow Loss. Previous monocular depth estimation methods [5,11,29] enforce temporal consistency by minimizing the difference between the flow-warped depth $\hat{\mathbf{D}}_i$ and $\hat{\mathbf{D}}_{i+1}$, assuming that the depths of corresponding points remain stationary. However, this assumption does not hold in real-world surgical scenes, which feature dynamic instruments and deformable tissues.

To address this limitation, as shown in Fig. 3, we propose a dynamics-aware flow loss that eliminates the stationary assumption by decoupling optical flow into pose-induced motion and pointmap-derived scene flow. Specifically, given the input image sequences, we first compute the forward optical flow as $\mathbf{O}_{i\to j}$ with an off-the-shelf model [37]. Optical flow captures 2D motion of pixels between frames, encompassing both camera motion and scene flow. To calculate the scene flow between frame i and j, we leverage optical flow to find correspondences between pointmap $\hat{\mathbf{X}}_{i,1}$ and $\hat{\mathbf{X}}_{j,1}$, the scene flow is calculated as:

$$\hat{\mathbf{S}}_{i\to j}(\mathbf{u}) = \hat{\mathbf{X}}_{j,1}(\mathbf{u} + \mathbf{O}_{i\to j}(\mathbf{u})) - \hat{\mathbf{X}}_{i,1}(\mathbf{u}), \tag{3}$$

where $\mathbf{u}$ is the homogeneous 2D coordinate. We restrict the computation to the valid region and define $\mathbf{u'} = \{\mathbf{u} | 0 < \mathbf{u} + \mathbf{O}_{i\to j}(\mathbf{u}) < (H, W)\}$. Then the estimated optical flow $\hat{\mathbf{f}}_{i\to j}$ could be calculated by combining the scene flow

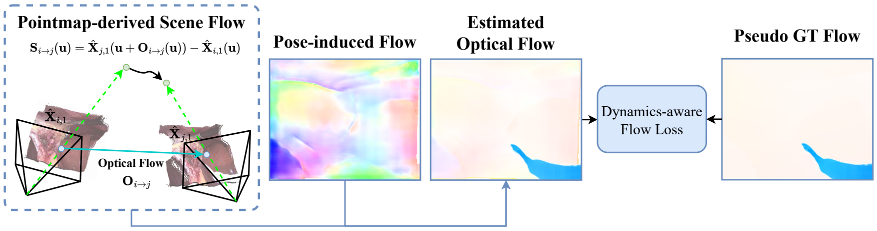
Pointmap-derived Scene Flow
$\mathbf{S}_{i\to j}(\mathbf{u}) = \hat{\mathbf{X}}_{j,1}(\mathbf{u}+\mathbf{O}_{i\to j}(\mathbf{u})) - \hat{\mathbf{X}}_{i,1}(\mathbf{u})$
$\hat{\mathbf{X}}_{i,1}$
$\hat{\mathbf{X}}_{j,1}$
Optical Flow
$\mathbf{O}_{i\to j}$
Pose-induced Flow
Estimated Optical Flow
Dynamics-aware Flow Loss
Pseudo GT Flow

Fig. 3. Illustration of dynamics-aware flow loss for self-supervised training to achieve enhanced temporal consistency.

with pose-induced flow as:

$$\hat{\mathbf{f}}_{i\to j}(\mathbf{u}') = \hat{\mathbf{K}}\hat{\mathbf{T}}_{j,1}(\hat{\mathbf{X}}_{i,1}(\mathbf{u}') + \hat{\mathbf{S}}_{i\to j}(\mathbf{u}')) - \mathbf{u}', \tag{4}$$

where $\hat{\mathbf{K}}$ denotes the estimated intrinsic by solving a simple optimization following [31]. The dynamics-aware flow loss can be written as:

$$\mathcal{L}_{\text{Dflow}}^{i\to j} = \| \hat{\mathbf{f}}_{i\to j}(\mathbf{u}') - \mathbf{O}_{i\to j}(\mathbf{u}') \|_1 \ . \tag{5}$$

Based on $\mathcal{L}_{\text{Dflow}}^{i\to j}$, we avoid the need for camera pose and depth for training.

Monocular Depth Loss. For datasets without either GT depth or pose, we use an off-the-shelf video depth model [29] to obtain the monocular depth and adopt a scale-invariant depth loss in Midas [25] to supervise the predicted depth $\hat{\mathbf{D}}$. We first calculate the shift and scale by least square to align $\mathbf{D}$ to $\hat{\mathbf{D}}$ and obtain $\tilde{\mathbf{D}}$, then minimize the $\mathcal{L}_2$ loss and gradient loss as follows:

$$\mathcal{L}_{\text{dep}} = \mathcal{L}_2 + \mathcal{L}_{\text{smooth}} = \frac{1}{M} \parallel \tilde{\mathbf{D}} - \mathbf{D} \parallel_2^2 + \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}(| \ \nabla_x \mathbf{R}_i^k + \nabla_y \mathbf{R}_i^k \ |), \tag{6}$$

where $\mathbf{R}_i$ denotes the difference between $\tilde{\mathbf{D}}$ and $\hat{\mathbf{D}}$ with scale level $K = 4$, M denotes the total pixels of image.

2.4 Training and Inference

Total Loss. Our total loss for training Endo3R is as follows:

$$\mathcal{L}_{all} = \lambda_1\mathcal{L}_{\text{Dflow}} + \lambda_2\mathcal{L}_{\text{dep}} + \lambda_3\mathcal{L}_{\text{conf}}, \tag{7}$$

where $\mathcal{L}_{\text{conf}}$ denotes the confidence-aware regression loss to supervise the pointmaps following [31], $\lambda_1, \lambda_2, \lambda_3$ denote the weights for losses.

3 Experiments

3.1 Implementation Details

Training Datasets. We train our Endo3R with a mixture of datasets, with four datasets containing GT/Stereo depth and pose (SCARED [21], StereoMIS [22],

Table 1. Quantitative comparison with depth estimation methods.

	Methods	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	FPS↑
SCARED	Monodepth2 [7]	0.432	3.548	4.704	0.431	0.425	22.05
	Endo-SfM [17]	0.241	0.865	2.286	0.267	0.585	7.33
	AF-SfM [30]	0.257	0.960	2.162	0.291	0.573	3.17
	EndoDAC [3]	0.242	0.934	2.014	0.275	0.584	31.79
	Transfer [5]	0.297	1.207	2.561	0.319	0.561	9.37
	DA-V2 [18]	0.313	1.425	2.839	0.453	0.508	4.18
	VDA [29]	0.291	1.186	2.447	0.296	0.647	6.86
	Endo DM [8]	0.203	0.651	2.063	0.245	0.612	14.58
	Monst3R [16]	0.198	0.539	1.965	0.234	0.626	18.68
	Endo3R(Ours)	**0.124**	**0.227**	**1.209**	**0.135**	**0.839**	19.17
Hamlyn	Monodepth2 [7]	0.379	9.318	20.472	0.403	0.439	22.05
	Endo-SfM [17]	0.252	4.335	14.430	0.268	0.628	7.33
	AF-SfM [30]	0.286	5.715	15.895	0.301	0.508	3.17
	EndoDAC [3]	0.275	5.557	15.669	0.288	0.519	31.79
	Transfer [5]	0.281	5.790	15.936	0.312	0.504	9.37
	DA-V2 [18]	0.334	7.713	19.548	0.362	0.461	4.18
	VDA [29]	0.315	7.492	19.231	0.347	0.476	6.86
	Endo DM [8]	0.216	4.639	14.799	0.273	0.619	14.58
	Monst3R [16]	0.198	4.193	15.221	0.241	0.645	18.68
	Endo3R(Ours)	**0.170**	**3.139**	**11.569**	**0.196**	**0.707**	19.17

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

Method	ATE ↓	RPE_r ↓	RPE_t ↓
Endo-SfM [17]	0.157	0.252	0.259
AF-SfM [30]	0.125	0.235	0.241
EndoDAC [3]	0.124	0.223	0.233
Robust [22]	0.131	0.241	0.245
Endo3R(Ours)	**0.112**	**0.201**	**0.228**

Table 3. Ablation study of Endo3R for different components.

Setting	Abs Rel↓	RMSE↓	$\delta < 1.25$ ↑
Baseline	0.198	1.965	0.626
w/ Uncertain.	0.165	1.654	0.720
w/ $\mathcal{L}_{\text{dep}}$	0.153	1.486	0.772
w/ $\mathcal{L}_{\text{Dflow}}$	**0.124**	**1.209**	**0.839**

C3VD [32], and Endomapper [23]), four datasets without GT data (AutoLaparo [39], Cholec80 [2], EndoVis17 [19], and EndoVis18 [20]). Specifically, we conduct stereo rectification for SCARED [21] and StereoMIS [22], using StereoAnything [34] to calculate the stereo depth of the left view for training. To evaluate the depth estimation, we evaluate our method on 320×256 resolution and follow the train and test split in SCARED [21]. To evaluate the generalization ability, we test on all 22 videos of the unseen Hamlyn Dataset for cross-dataset zero-shot validation.

Evaluation Metrics. We compare Endo3R with state-of-the-art depth estimation methods. We follow [3] to use five metrics commonly used in monocular depth estimation: Abs Rel, Sq Rel, RMSE, RMSE log, $\delta < 1.25$. We also compare the inference FPS to compare the efficiency. To evaluate the pose accuracy, we perform a 5-frame pose evaluation and adopt Absolute Trajectory Error (ATE) and Relative Pose Error (RPE), including rotation RPE_r and translation RPE_t. Note that the unit for RPE_t and ATE is mm, and the unit for RPE_r is degree.

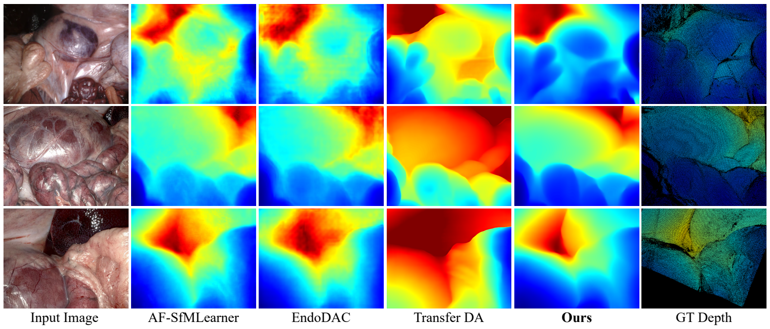

Fig. 4. Qualitative results of monocular depth estimation.

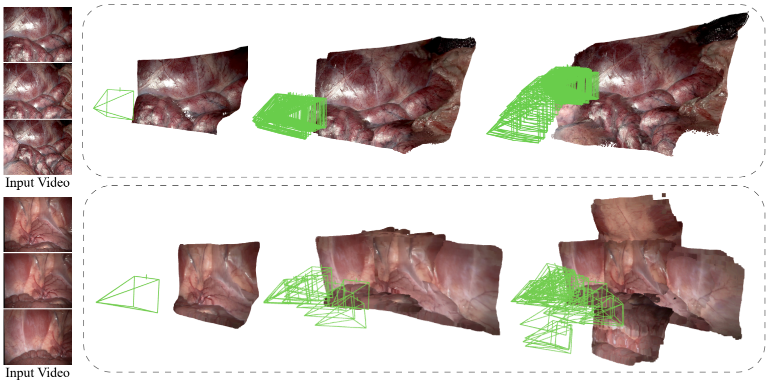

Fig. 5. Qualitative results of Online 3D Reconstruction.

3.2 Experimental Results

Quantitative Comparison. We evaluate our method and SOTA depth estimation methods on SCARED and Hamlyn datasets. The results in Tab. 1 reveal that our approach achieves a substantial improvement in depth estimation accuracy compared to existing methods, even without training on the Hamlyn Dataset. Notably, while delivering superior accuracy, our method maintains a competitive FPS rate to support online applications. We also report the pose estimation results on SCARED in Tab. 2. The results demonstrate that our method achieves the highest pose estimation accuracy.

Qualitative Comparison. The qualitative evaluation of our depth estimation is illustrated in Fig. 4, demonstrating that Endo3R produces more precise depth maps with improved relative scale. Furthermore, Fig. 5 presents our online 3D reconstruction results with pose estimation. The high-quality 3D reconstructions can be attributed to the superior depth and pose estimation accuracy. Please find more visualization results in the supplementary video.

Ablation Study. We set Monst3R [16] as baseline and conduct ablation studies on the different components of Endo3R. As reported in Tab. 3, it shows the effectiveness of each component with increasing performance.

4 Conclusion

We present Endo3R, a unified framework for online 3D reconstruction from uncalibrated surgical videos. By jointly estimating depth, pose, and scene geometry in a single stage, our method eliminates the need for multi-stage pipelines or offline optimization. This work provides a foundation in real-time surgical scene understanding and computer-assisted intervention. For the limitation, Endo3R may suffer from performance degradation over extra-long sequences due to the lack of global bundle adjustment, leading to the drifting issue. Future work will focus on enhancing the robustness and expanding to more surgical modalities.

Acknowledgments. This work is supported in part by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part of the HK RGC AoE under AoE/E-407/24-N, Grant 14218322, and Grant 14207320, in part by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government via the Hong Kong Centre for Logistics Robotics, in part by the Multi-Scale Medical Robotics Centre, and in part by the CUHK T Stone Robotics Institute.

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

References

1. A. Dosovitskiy *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. A. Twinanda *et al.*: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
3. B. Cui *et al.*: Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: MICCAI. pp. 208–218. Springer (2024)
4. B. Mildenhall *et al.*: Nerf: Representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
5. C. Budd *et al.*: Transferring relative monocular depth to surgical vision with temporal consistency. In: MICCAI. pp. 692–702. Springer (2024)
6. C. Campos *et al.*: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE transactions on robotics **37**(6), 1874–1890 (2021)
7. C. Godard *et al.*: Digging into self-supervised monocular depth estimation. In: ICCV. pp. 3828–3838 (2019)
8. D. Recasens *et al.*: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE RA-L **6**(4), 7225–7232 (2021)

9. E. Arnold *et al.*: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV. pp. 690–708. Springer (2022)
10. H. Wang *et al.*: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024)
11. H. Yang *et al.*: Depth any video with scalable synthetic data. arXiv preprint arXiv:2410.10815 (2024)
12. J. Guo *et al.*: A visual navigation perspective for category-level object pose estimation. In: ECCV. pp. 123–141. Springer (2022)
13. J. Guo *et al.*: Free-surgs: Sfm-free 3d gaussian splatting for surgical scene reconstruction. In: MICCAI. pp. 350–360. Springer (2024)
14. J. Guo *et al.*: Uc-nerf: Uncertainty-aware conditional neural radiance fields from endoscopic sparse views. IEEE Transactions on Medical Imaging (2024)
15. J. Schonberger *et al.*: Structure-from-motion revisited. In: ICCV. pp. 4104–4113 (2016)
16. J. Zhang *et al.*: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
17. K. Ozyoruk *et al.*: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical image analysis **71**, 102058 (2021)
18. L. Yang *et al.*: Depth anything v2. NeurIPS **37**, 21875–21911 (2025)
19. M. Allan *et al.*: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
20. M. Allan *et al.*: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
21. M. Allan *et al.*: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
22. M. Hayoz *et al.*: Learning how to robustly estimate camera pose in endoscopic videos. International journal of computer assisted radiology and surgery **18**(7), 1185–1192 (2023)
23. P. Azagra *et al.*: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1), 671 (2023)
24. P. Zhang *et al.*: Real-time navigation for laparoscopic hepatectomy using image fusion of preoperative 3d surgical plan and intraoperative indocyanine green fluorescence imaging. Surgical endoscopy **34**, 3449–3459 (2020)
25. R. Ranftl *et al.*: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1623–1637 (2020)
26. R. Ranftl *et al.*: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
27. R. Wei *et al.*: Enhanced scale-aware depth estimation for monocular endoscopic scenes with geometric modeling. In: MICCAI. pp. 263–273. Springer (2024)
28. R. Wei *et al.*: Scale-aware monocular reconstruction via robot kinematics and visual data in neural radiance fields. Artificial Intelligence Surgery **4**(3), 187–198 (2024)
29. S. Chen *et al.*: Video depth anything: Consistent depth estimation for super-long videos. arXiv preprint arXiv:2501.12375 (2025)
30. S. Shao *et al.*: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical image analysis **77**, 102338 (2022)
31. S. Wang *et al.*: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
32. T. Bobrow *et al.*: Colonoscopy 3d video dataset with paired depth from 2d-3d registration. Medical Image Analysis p. 102956 (2023)

33. X. Gu *et al.*: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR. pp. 2495–2504 (2020)
34. X. Guo *et al.*: Stereo anything: Unifying stereo matching with large-scale mixed data. arXiv preprint arXiv:2411.14053 (2024)
35. Y. Lu *et al.*: Autonomous intelligent navigation for flexible endoscopy using monocular depth guidance and 3-d shape planning. In: ICRA. pp. 1–7. IEEE (2023)
36. Y. Yao *et al.*: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV. pp. 767–783 (2018)
37. Z. Teed *et al.*: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
38. Z. Teed *et al.*: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems **34**, 16558–16569 (2021)
39. Z. Wang *et al.*: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: MICCAI. pp. 486–496 (2022)