IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

ColonAdapter: Geometry Estimation Through Foundation Model Adaptation for Colonoscopy

Zhiyi Jiang[1], Yifu Wang[2], Xuelian Cheng[3*], and Zongyuan Ge[4]

***Abstract*—Estimating 3D geometry from monocular colonoscopy images is challenging due to non-Lambertian surfaces, moving light sources, and large textureless regions. While recent 3D geometric foundation models eliminate the need for multi-stage pipelines, their performance deteriorates in clinical scenes. These models are primarily trained on natural scene datasets and struggle with specularity and homogeneous textures typical in colonoscopy, leading to inaccurate geometry estimation. In this paper, we present ColonAdapter, a self-supervised fine-tuning framework that adapts geometric foundation models for colonoscopy geometry estimation. Our method leverages pretrained geometric priors while tailoring them to clinical data. To improve performance in low-texture regions and ensure scale consistency, we introduce a Detail Restoration Module (DRM) and a geometry consistency loss. Furthermore, a confidence-weighted photometric loss enhances training stability in clinical environments. Experiments on both synthetic and real datasets demonstrate that our approach achieves state-of-the-art performance in camera pose estimation, monocular depth prediction, and dense 3D point map reconstruction, without requiring ground-truth intrinsic parameters.**

***Index Terms*—Deep Learning for Visual Perception, Computer Vision for Medical Robotics, Localization**

I. Introduction

COLORECTAL cancer (CRC) is among the third most common type of cancer in the world, imposing a healthcare burden globally [1]. In the screening and treating of CRC, colonoscopy has been widely utilized as a gold-standard procedure [2]. Despite its effectiveness, the colonoscopic procedure is subject to the experience of the clinician, as they have to screen a complex anatomical environment through monocular videos, which has a limited field of view and a lack of spatial information. To overcome these challenges, one promising approach is to estimate 3D organ geometry by dense reconstruction from monocular images [3]. Traditional feature-based reconstruction methods struggle in clinical scenarios because of the sparse and unevenly distributed key points in endoscopic images [5], resulting in poor reconstructions. To address these limitations, deep learning approaches that bypass handcrafted feature correspondences have gained traction for mapping and reconstruction in clinical settings. Some methods achieve dense reconstructions by incorporating learning-based depth and pose estimation within SLAM [6] or SfM [7] frameworks. However, they typically assume known ground-truth intrinsic parameters, which are often unavailable or vary across devices. Additionally, their reliance on multi-stage reconstruction pipelines makes them prone to noise and error propagation [4].

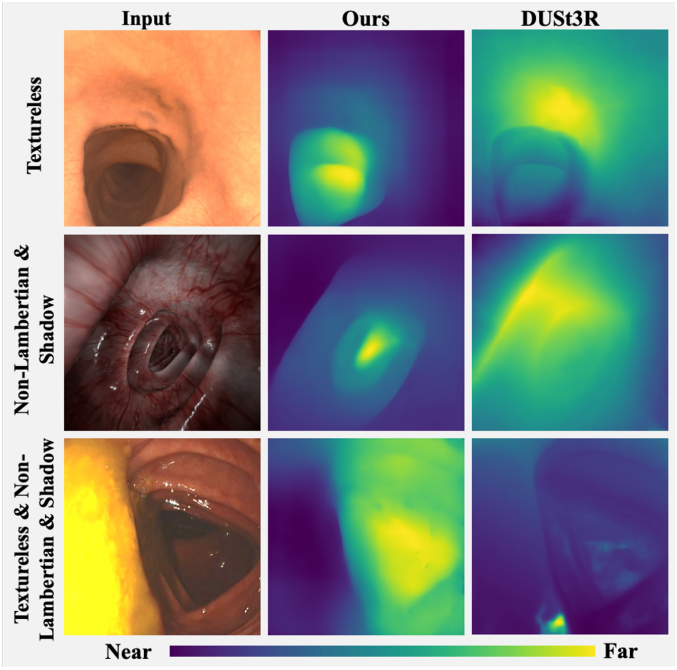

Fig. 1: Comparison of depth map estimations on colonoscopy images between a 3D geometric foundation model (DUSt3R [4]) and our proposed method. The top row shows how extensive textureless regions cause the model to misinterpret distant areas as close (and vice versa). The middle row highlights how shadows caused by moving light sources and complex anatomical structures lead DUSt3R to erroneously label them as distant. In contrast, the bottom row illustrates that our method accurately reconstructs scenes containing textureless surfaces, dynamic shadows, and non-Lambertian regions, where DUSt3R fails.

Recent advancements in 3D geometric foundation models [4] [8] offer a promising solution by directly generating dense 3D point maps with 3D geometric information, eliminating the need for sub-modules and reliance on ground-truth intrinsic parameters. However, these models often struggle in challenging scenarios such as low-overlap or textureless regions, leading to degraded performance [9] [10]. LoRA3D

Manuscript received: June 16, 2025; Revised: September 16, 2025; Accepted: October 3, 2025.

This paper was recommended for publication by Editor Pascal Vasseur upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Jiangsu Department of Technology Natural Science Fund (Grants No: BK20250441) and Center of Excellence for Antimicrobial Therapeutics Discovery and Innovation (CEATDI) (Grants No: MSRI8002003).

[1]Z. Jiang is with School of Computer Science and Engineering, Southeast University, China zyjiang97@outlook.com

[2]Y. Wang is Vetex Lab, Shanghai, China 1fwang927@gmail.com

[3]X. Cheng is with Southeast University - Monash University Joint Graduate School, Suzhou 215123, China, Monash Suzhou Research Institute, Monash University, Suzhou 215000, China, and also with Department of Data Science & Artificial Intelligence (DSAI), Monash University, Clayton, VIC 3800, Australia xuelian.cheng@monash.edu

[3]Z. Ge is with Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia zongyuan.ge@monash.edu

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[9] proposed a self-calibration technique to address these challenges, but it relies on the assumption that the overall scene geometry is relatively accurate. In natural scenes, this assumption is generally valid, as challenging regions typically occupy only a relatively small portion of the image, resulting in reliable scene reconstructions with only minor local errors. In clinical domains like colonoscopy, however, the situation is different: large textureless areas, complex anatomical structures, moving light sources, and non-Lambertian surfaces lead to inaccurate reconstruction of the entire scene, as shown in Fig. 1. Consequently, existing methods are not suitable for addressing these substantial challenges for colonoscopy.

To adapt 3D geometric foundation models for colonoscopy, we propose a self-supervised fine-tuning strategy using only monocular videos, without camera information or ground-truth depth. Instead of relying on pixel-wise 3D point map losses, we demonstrate that photometric loss supervision suffices by leveraging the geometric priors of the foundation model. To enhance fine-detail recovery, we introduce a Detail Restoration Module (DRM) that fuses fine details from an auxiliary convolutional encoder into the foundation model. This straightforward fusion strategy also ensures that the DRM can be seamlessly extended to other ViT-based geometric foundation models.

Our main contributions are as follows: (1) Development of a self-supervised fine-tuning framework that adapts 3D geometric foundation models to colonoscopy scene through the supervision of photometric constraints. The confidence map is integrated into the photometric loss for training stability enhancement, and a geometric consistency loss is introduced to ensure coherent geometric predictions across frames. (2) Design of a Detail Restoration Module for ViT-based geometric foundation models, featuring an auxiliary CNN encoder and a feature fusion adapter for enhanced multi-level detail extraction and reconstruction accuracy in colonoscopy scenes. (3) Comprehensive evaluation on three colonoscopy datasets, demonstrating strong performance across multiple 3D vision tasks, including camera pose estimation, monocular depth estimation, and dense point map estimation.

II. Related Work

A. Self-Supervised Learning

Self-supervised learning has achieved notable success in a range of geometric perception tasks, including monocular depth prediction [11], multi-view monocular depth estimation [12], and structure-from-motion [13]. These methods predominantly rely on photometric constraints, which perform well in natural scenes but are often violated in challenging environments, such as clinical scenarios. To address this, AF-SfMLearner [5] introduced appearance flow to handle brightness inconsistencies, while MonoPCC [14] proposed a photometric-invariant cycle consistency constraint. In parallel, geometric constraints have proven effective in addressing domain-specific challenges, as demonstrated by Wang et al. [15], who leverage inherent geometrical properties of satellite structures in their self-training approach for satellite pose estimation. Despite these advances, existing methods still rely on ground-truth camera intrinsics and predefined depth ranges during training. In contrast, our method processes consecutive frames as pair-view inputs and directly outputs scene geometry in the form of 3D point maps, from which both intrinsic parameters and camera poses can be derived.

B. Vision Foundation Model Specialization

Specializing vision foundation models through fine-tuning has become the standard approach for customizing pre-trained models to specific domains. For 3D geometric foundation models, several extensions have been proposed to enhance an existing architecture DUSt3R [4]. LoRA3D [9] applies low-rank adaptation (LoRA) [16] for efficient self-calibration. Align3R [17] integrates monocular depth priors to enhance video depth estimation. However, these methods are not specifically designed for endoscopic or colonoscopic environments, which contain unique challenges such as large textureless regions, non-Lambertian surfaces, and moving light sources.

In endoscopy scenario, few works explored adapting depth anything model (DAM) with self-supervised framework. DARES [18] applies LoRA to the DAM and joint training with a pose net. EndoDAC [19] introduced DV-LoRA and an extra intrinsic head for intrinsic estimation. With the intrinsic head, EndoDAC could be trained on any endoscopic videos. However, the separate models utilized for different sub-tasks making it vulnerable to noise and errors in each individual component.

C. Concurrent Work

Endo3R [20] is a concurrent work that proposes a reconstruction framework for clinical scenarios and similarly incorporates optical flow into its loss formulation. A key distinction, however, lies in the supervision strategy: Endo3R relies on datasets with ground-truth depth or pseudo-labels generated by external video depth anything models, while our method is fully self-supervised and does not require ground-truth annotations or auxiliary depth models. Furthermore, our approach introduces an additional module to enhance the extraction of fine details, an aspect not addressed by Endo3R.

III. Method

We achieve geometry estimation using foundation model adaptation and further formulate such adaptation as 2D image reconstruction using a self-supervised framework. The overall training pipeline of the proposed framework is depicted in Fig. 2, where DUSt3R is employed as the backbone foundation model and LoRA is used for fine-tuning. Section A describes the details of overall framework. Section B introduces the Detail Restoration Module (DRM) for fine details enhancement. Section C outlines the training objectives, including the confidence-weighted photometric loss and the geometry consistency loss.

A. Overall Framework

The framework main architecture, as illustrated in Fig. 2, consists of adapted foundation module and affiliation module.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

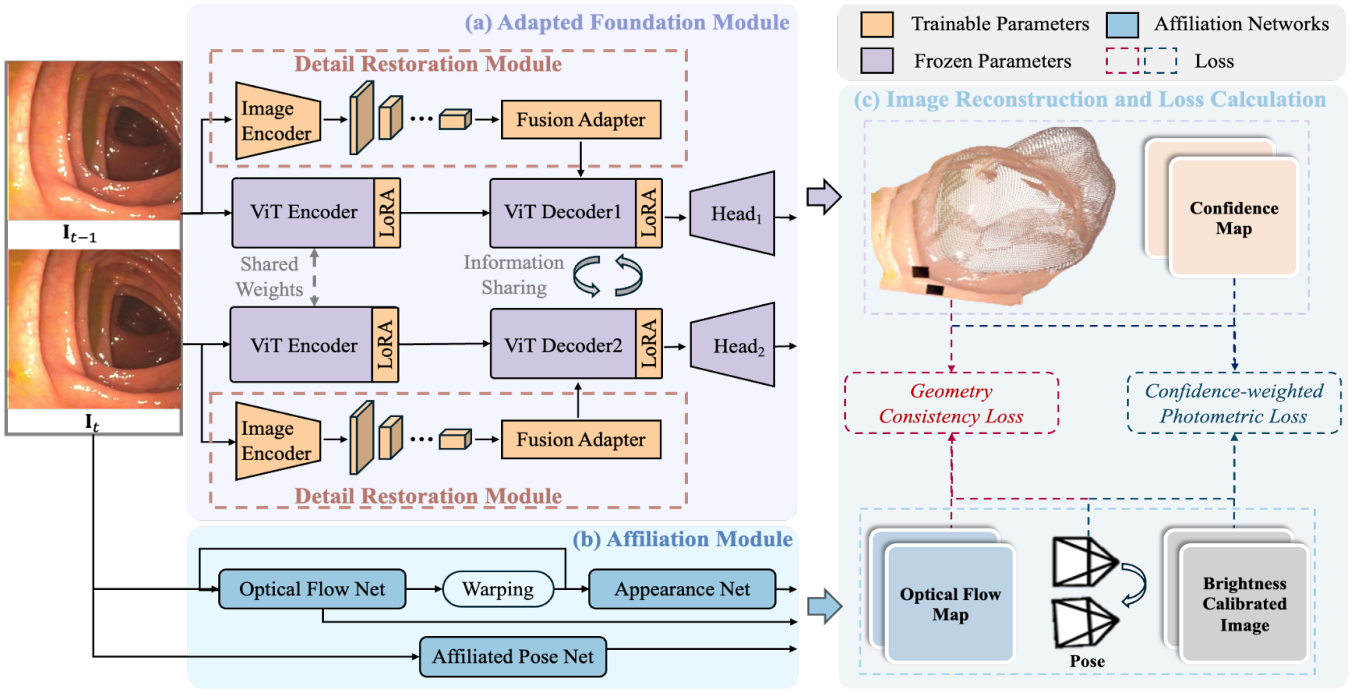

Fig. 2: The training pipeline of our proposed framework, consisting of an adapted foundation module (a), an affiliation module (b), and loss calculation (c). The adapted foundation module takes two input images and generates corresponding point maps along with their confidence maps. The affiliation module, used only during training, provides image brightness calibration, optical flow, and affiliated pose information. With the generated components from these two modules, we reconstruct image and calculate the losses. During evaluation, we rely solely on the adapted foundation module to generate point maps, which are then used to derive camera information, including poses and intrinsic parameters.

The adapted foundation module generates the point map and calculates the intrinsic parameters using the point map. The affiliation module estimates the camera pose, calibrates brightness, and provides geometric consistency. The results of the two modules are subsequently input for image reconstruction and loss calculation.

(a) Adapted Foundation Module

The adapted foundation model primarily consists of a backbone foundation model with LoRA for fine-tuning and a Detail Restoration Module (DRM) for enhancing fine details, with further discussion of the DRM provided in Section B. The adapted foundation model takes two RGB images $\mathbf{I}^t$, $\mathbf{I}^{t-1} \in \mathbb{R}^{W\times H\times 3}$ as input and predicts corresponding point maps $\mathbf{X}^{t;t}$, $\mathbf{X}^{t-1;t} \in \mathbb{R}^{W\times H\times 3}$, along with their associated confidence maps $\mathbf{C}^{t;t}$, $\mathbf{C}^{t-1;t} \in \mathbb{R}^{W\times H}$. Here, $\mathbf{X}^{t-1;t}$ represents the point map of frame $t-1$ expressed in the coordinate of frame t. During inference, the model directly generates dense point maps that can be utilized to derive camera poses and intrinsic parameters. During training, we only utilize the point map $\mathbf{X}^{t;t}$ and the derived intrinsic parameters to achieve the image reconstruction.

From the predicted point map, intrinsic parameters are estimated. This process assumes square pixels and centered principal points, thereby simplifying the task to estimating the focal length. Following the approach in [4], the Weiszfeld algorithm is used to iteratively minimize the reprojection error:

$$f^* = \arg\min_f \sum_i \left\| \mathbf{u}_i^t - f \cdot \mathbf{q}_i^t \right\|^2, \tag{1}$$

where $\mathbf{u}_i^t$ are the observed image points in image $\mathbf{I}^t$ and $\mathbf{q}_i$ is the corresponding normalized direction vector derived from point map $\mathbf{X}^{t;t}$. This optimization typically converges within 10 iterations.

(b) Affiliation Module

The affiliation module consists of an optical flow network, an appearance network, and an affiliated pose network, all of which are used exclusively during the training phase. Following the approach in [5], the optical flow and appearance networks are employed to address photometric inconsistencies caused by moving light sources and complex environments. Specifically, the optical flow network warps $\mathbf{I}^{t-1}$ toward $\mathbf{I}^t$. The resulting warped image is then input to the appearance network, together with $\mathbf{I}^t$, to calibrate the brightness of $\mathbf{I}^t$ and align it with $\mathbf{I}^{t-1}$. This process calibrates the brightness in $\mathbf{I}^t$ and generates the calibrated image $\tilde{\mathbf{I}}^t$.

To provide pose information, we introduce an affiliated pose network to estimate the relative pose $\mathbf{T}^{t\to t-1} \in SE(3)$ between the two frames. This choice is motivated by the fact that PnP algorithm used by [4] could fail and the algorithm accuracy also depends on the quality of the estimated confidence map. To improve generalizability, a pose network is utilized which relies solely on the image pair as input.

(c) Image Reconstruction and Loss Calculation

Given the estimated point map $\mathbf{X}^{t;t}$ and relative pose $\mathbf{T}^{t\to t-1}$, the target image $\mathbf{I}^t$ is reconstructed by warping the source image $\mathbf{I}^{t-1}$:

$$\hat{\mathbf{I}}^t = \pi\big(K, \mathbf{T}^{t\to t-1}, \mathbf{X}^{t;t}, \mathbf{I}^{t-1}\big), \tag{2}$$

where $K \in \mathbb{R}^{3\times 3}$ is the intrinsic matrix calculated from the focal length $f \in \mathbb{R}$ and $\pi(\cdot)$ denotes the re-projection operation. Here, $\hat{\mathbf{I}}^t$ represents the reconstructed target image obtained from $\mathbf{I}^{t-1}$. The reconstructed image is further utilized in the loss calculation.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

The primary training objective is to minimize the difference between the reconstructed and target images, using photometric loss [5] that combines the structural similarity index (SSIM) [21] and pixel-wise differences. For each valid pixel i, the photometric loss of frame t is defined as:

$$l^t_{\text{photo}}(i) = \alpha \cdot \frac{1 - \text{SSIM}(\tilde{\mathbf{I}}^t(i), \hat{\mathbf{I}}^t(i))}{2} + (1-\alpha) \cdot \left\| \tilde{\mathbf{I}}^t(i) - \hat{\mathbf{I}}^t(i) \right\|_1, \quad (3)$$

where $\tilde{\mathbf{I}}^t(i)$ represent the brightness-calibrated target image at valid pixel i, which is determined using visibility masks [5]. The balancing factor α is set to 0.85. To facilitate the model training in colonoscopy, we further enhanced photometric loss with confidence and extra geometry consistency. The details for these two losses are expanded in Section C.

B. Detail Restoration Module

As demonstrated in [17], the DUSt3R generates only coarse 3D point maps. To enhance fine details, we incorporate fine-grained, high-frequency features extracted by a CNN encoder and fuse them with ViT features. This approach avoids relying on depth maps generated by monocular depth models [17], which may suffer from scale ambiguity [22]. For the CNN encoder, we adopt ResNet-18 [23], as it captures sufficient structural details in multi-level. Using heavier encoders yields only marginal improvements at the cost of significantly higher computational demands. The architecture of the Detail Restoration Module (DRM) is illustrated in Fig. 3.

In the fusion of extracted CNN features and ViT features, we explored several strategies for feature fusion adapter, including: (1) a feature exchange block inspired by ViT-Adapter [24] architectures to enable bidirectional information flow; (2) a Convolutional Block Attention Module (CBAM) [25] , which introduces channel and spatial attention between two kinds of features; and (3) a zero convolution [26] with two blocks to match the channel and spatial dimensions of CNN and ViT features. Although the third method employs a simpler architecture with significantly fewer parameters, it achieves superior performance in our experiments, as shown in Table VI. Thus, it is selected as the final design for the feature fusion module.

In the proposed DRM, a pair of input frames is processed by the ResNet-18 encoder (shared weights) to extract multi-level features. The multi-level features pass through five Fusion Adapters, each containing a channel projection block, a spatial projection block, and a zero convolution. The channel projection block utilizes 1×1 convolution for channel alignment. The spatial projection block uses 3×3 convolution and average pooling to refine features and match the spatial resolution of the ViT features. The aligned CNN features are subsequently fused with the ViT encoder outputs using the zero-convolution. Unlike [17], which injects monocular priors into all transformer layers, we restrict our fusion to the first five layers of the ViT decoder. Following [27], deeper transformer layers capture high-level semantics, while shallow layers retain low-level details. Thus, we integrate low-level CNN features only into these shallow layers, enhancing detail reconstruction while preserving the high-level transformer representations.

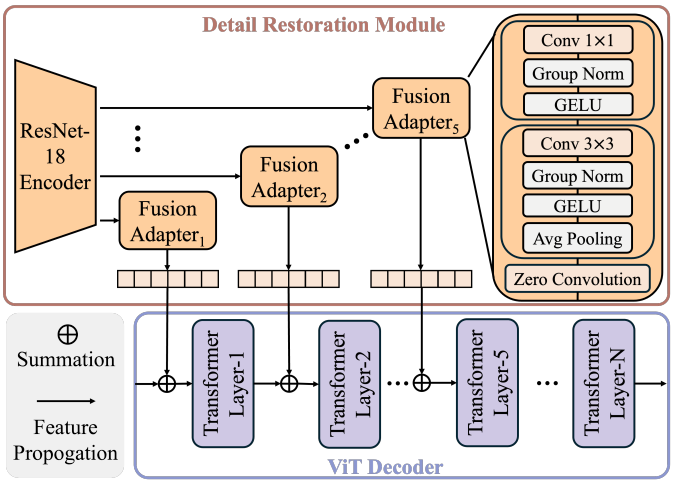

Fig. 3: Architecture of the proposed Detail Restoration Module. The ResNet-18 extracts multi-level features from the input images, which are then fused with ViT features through fusion adapters. The fused features are injected only into the first five layers of the ViT decoder to provide low-level information.

C. Training Objectives

(a) Confidence-Weighted Photometric Loss

While this photometric loss supervises the estimated point maps, we observed that relying on it alone can slow down convergence. In some cases, it even degrades performance compared to the pre-trained foundation model when introducing the DRM, as shown in the fifth row of Table VII. To address such issue, we introduce a confidence-weighted photometric loss that leverages per-pixel confidence maps predicted by the backbone model.

Confidence maps estimate the reliability of predictions at each pixel, allowing models to focus on reliable regions. This technique is used in 3D reconstruction frameworks like DUSt3R [4] and VGGT [8], enhancing robustness in occluded or low-texture areas. A direct approach is to adopt the confidence-aware loss from DUSt3R and replace the 3D regression loss with the photometric loss. However, using only the confidence-aware loss neglects low-confidence regions. To address this, we combine the photometric loss of frame $t-1$ with the confidence-aware loss of frame t to form the confidence-weighted photometric loss. This design ensures that low-confidence regions in frame t are supervised when warped to frame $t-1$ for photometric loss computation, without being discarded by the confidence map of frame $t-1$. In addition, such design complements the confidence-aware loss by providing stable gradient coverage, enhancing training robustness and optimization stability.

For the loss calculation, the model processes adjacent frames $\mathbf{I}^t$ and $\mathbf{I}^{t-1}$ bidirectionally, producing four point maps: $\mathbf{X}^{t;t}$, $\mathbf{X}^{t-1;t}$, $\mathbf{X}^{t;t-1}$, and $\mathbf{X}^{t-1;t-1}$. We use $\mathbf{X}^{t;t}$ and its confidence map $\mathbf{C}^{t;t}$ for the confidence-aware loss, and $\mathbf{X}^{t-1;t-1}$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

for the photometric loss:

$$\mathcal{L}_{\text{conf-photo}} = \lambda_{conf} \cdot \left(\frac{1}{N^t} \sum_{i=1}^{N^t} \mathbf{C}_i^t \cdot l_{\text{photo}}^t(i) - \beta \cdot \log(\mathbf{C}_i^t) \right) + \lambda_{photo} \cdot \left(\frac{1}{N^{t-1}} \sum_{i=1}^{N^{t-1}} l_{\text{photo}}^{t-1}(i) \right), \quad (4)$$

where (N^t, N^{t-1}) denote the numbers of valid pixels, and $(\lambda_{conf}, \lambda_{photo})$ are weights for the confidence-aware and photometric losses. Incorporating the confidence map not only calibrates itself but also enhances downstream tasks that relies on confidence map such as PnP-based pose estimation and camera intrinsic estimation.

(b) Geometry Consistency Loss

The point maps from the adapted foundation module and the poses estimated by the affiliated pose network are inherently scale-invariant, which may lead to scale inconsistency. Furthermore, confidence-weighted photometric loss does not directly supervise $\mathbf{X}^{t-1;t}$ and $\mathbf{X}^{t;t-1}$. To address both the scale inconsistencies and the lack of direct supervision for these point maps, we introduce a geometry consistency loss. This loss has two main components: one enforces scale consistency and geometric coherence across frames, while the other aligns the predicted camera poses scale from the affiliation module with the point maps scale. This loss leverages the four point maps generated in the $\mathcal{L}_{\text{conf-photo}}$ calculation.

For the first component, we align the scales of the point maps using the optical flow network within the affiliation module, which predicts the optical flows $\mathbf{F}^{t\leftarrow t-1}$, $\mathbf{F}^{t-1\leftarrow t} \in \mathbb{R}^{W\times H\times 2}$ and the occlusion masks $\mathbf{M}^{t\leftarrow t-1}$, $\mathbf{M}^{t-1\leftarrow t} \in \mathbb{R}^{W\times H}$. The optical flows warp the point maps $\mathbf{X}^{t;t}$ and $\mathbf{X}^{t-1;t-1}$ to produce $\hat{\mathbf{X}}^{t-1;t}$ and $\hat{\mathbf{X}}^{t;t-1}$. With the warped point maps and occlusion masks, the corresponding alignment term is defined as:

$$\mathcal{T}_{\text{flow}} = \mathbf{M}^{t\leftarrow t-1} \left\| \mathbf{X}^{t-1;t} - \hat{\mathbf{X}}^{t-1;t} \right\|_1 + \mathbf{M}^{t-1\leftarrow t} \left\| \mathbf{X}^{t;t-1} - \hat{\mathbf{X}}^{t;t-1} \right\|_1. \quad (5)$$

For the second component, to align the estimated poses scale from the affiliation module with the point maps scale, we use the predicted poses $\mathbf{T}^{t\to t-1}$ and $\mathbf{T}^{t-1\to t}$ to transform $\mathbf{X}^{t-1;t}$ into the coordinate frames of $t-1$ and t, resulting in $\hat{\mathbf{X}}^{t-1;t-1}$ and $\hat{\mathbf{X}}^{t;t}$. The corresponding transformation term is defined as:

$$\mathcal{T}_{\text{pose}} = \left\| \mathbf{X}^{t-1;t-1} - \hat{\mathbf{X}}^{t-1;t-1} \right\|_1 + \left\| \mathbf{X}^{t;t} - \hat{\mathbf{X}}^{t;t} \right\|_1. \quad (6)$$

The final geometry consistency loss combines the two components, each weighted by separate factors λ_{flow} and λ_{pose}:

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{flow}} \cdot \mathcal{T}_{\text{flow}} + \lambda_{\text{pose}} \cdot \mathcal{T}_{\text{pose}}, \quad (7)$$

where λ_{flow} and λ_{pose} are the weights of the two components, respectively.

IV. EXPERIMENTS, RESULTS, AND DISCUSSION

A. Experimental Setup

Simcol3D: Provided by the MICCAI 2022 EndoVis Challenge, SimCol3D [28] is a synthetic colonoscopy dataset containing ground truth depth and pose. Virtual light sources are attached to the endoscope to simulate realistic illumination. For our experiments, we use data SyntheticColon I from SimCol3D, selecting 10 sequences for training and 3 for testing.

Simulated Colonscopy Dataset (CSD): The CSD dataset is collected from a colonoscopy simulator [3] offering a more complex environment with rich vascular textures. Three predefined paths are included; the first is used for testing, while the remaining two serve for training. This results in 8,246 training images and 2,034 testing images.

EndoMapper: EndoMapper [29] is a large-scale clinical dataset comprising 96 high-definition colonoscopy sequences, totaling over 24 hours of real-world video. Due to the absence of ground-truth depth, we extract representative frames from colonoscopy videos for qualitative evaluation only.

C3VD: C3VD [30] is a phantom dataset acquired using an Olympus CF-HQ190L endoscope, containing 22 video sequences (10,015 images). For the quantitative evaluation of generalizability, We selected one representative sequence from each anatomical region for testing: Cecum (`cecum_t2_b`), Descending Colon (`desc_t4_a`), Sigmoid Colon (`sigmoid_t3_a`), and Transcending Colon (`trans_t4_a`).

Implementation Details: All input images are resized to 224×224 for training efficiency. Experiments are conducted using a single NVIDIA H100 GPU, with batch size 8. The learning rate is set to 1e-4. The CNN encoder (ResNet-18) is initialized with ImageNet-pretrained weights, while DUSt3R is initialized from its official pre-trained model.

B. Camera Pose Estimation

Camera pose estimation is evaluated on the SimCol3D dataset through two experiments. Following the protocol in [31], we randomly sample 10 frames from distinct segments of

TABLE I: Quantitative camera pose estimation on various sequences of SimCol3D, evaluated using 10 random frames in accordance with [31]. "G.I." indicates the use of ground-truth intrinsic parameters.

Method	G.I.	ATE ↓	RPE Trans ↓	RPE Rot ↓
AF-SfMLearner	✓	<u>0.0067</u>	0.1479	0.8518
Lite-Mono	✓	0.0169	0.1438	0.5962
Monodepth2	✓	0.0103	0.1459	<u>0.5679</u>
DARES	✓	0.0109	0.1473	0.9073
EndoDAC	×	0.0146	<u>0.1411</u>	0.5787
Ours	×	**0.0062**	**0.0570**	**0.5573**

TABLE II: Quantitative ego-motion comparison on the SimCol3D dataset. The ATE is averaged over all 5-frame snippets following [5].

Method	G.I.	ATE_{s1} ↓	ATE_{s2} ↓	ATE_{s3} ↓
AF-SfMLearner	✓	0.2704	0.2710	<u>0.1591</u>
Lite-Mono	✓	0.3064	0.3114	0.1620
Monodepth2	✓	0.2951	0.3143	0.1702
DARES	✓	0.2678	**0.2620**	0.1594
EndoDAC	×	0.2711	0.2937	0.1596
Ours	×	**0.2654**	<u>0.2660</u>	**0.1580**

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III: Monocular depth estimation on SimCol3D and CSD datasets. Best results are in **bold**, second-best are underlined.

Method	G.I.	SimCol3D					CSD				
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
AF-SfMLearner	✓	0.077	0.474	4.124	0.111	0.951	0.071	0.208	2.565	0.110	0.950
Lite-Mono	✓	0.092	0.708	3.976	0.123	0.937	0.080	0.207	2.428	0.117	0.943
Monodepth2	✓	0.083	0.554	3.861	0.112	0.950	0.100	0.668	3.208	0.129	0.915
DARES	✓	0.077	0.439	4.458	0.121	0.944	0.116	0.616	3.339	0.157	0.881
ManyDepth	✓	0.084	0.526	4.291	0.118	0.944	0.072	0.207	**2.347**	0.104	0.959
EndoDAC	×	0.170	4.333	9.464	0.226	0.807	0.215	1.908	6.989	0.284	0.682
Dust3R	×	0.262	3.07	10.744	0.338	0.555	0.270	2.111	7.457	0.394	0.572
Ours	×	**0.062**	**0.340**	**3.732**	**0.094**	**0.969**	**0.063**	**0.183**	2.563	**0.099**	**0.962**

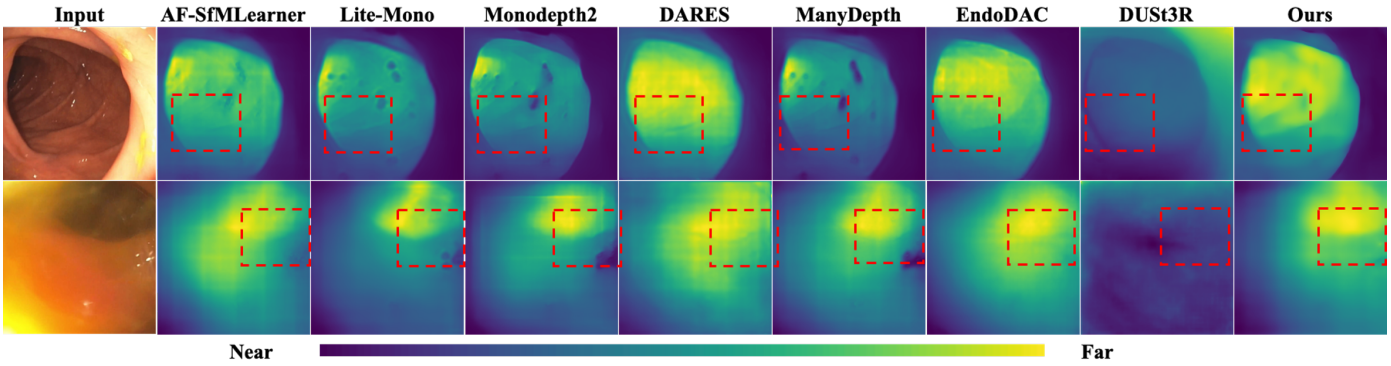

Fig. 4: Qualitative depth estimation results on the EndoMapper dataset. The top row highlights our method's ability, enhanced by the integration of DRM, to capture fine structural details (highlighted with red box) that are missed by other approaches. The bottom row shows that even in the presence of unseen artifacts such as bubble, our method still predicts artifact-free geometry.

the colon and assess performance using Absolute Translation Error (ATE), Relative Translation Error (RPE-trans), and Relative Rotation Error (RPE-rot). All predicted poses are aligned to ground truth using Sim(3) Umeyama alignment.

We compare our method with other self-supervised joint depth-pose estimation method AF-SfMLeaner [5], Lite-Mono [32], Monodepth2 [11], DARES [18], ManyDepth [12], and EndoDAC [19]. To ensure a fair comparison, all methods are trained on the same datasets as ours using the released code of authors. For EndoDAC, camera intrinsics learning is enabled during training. As shown in Table I, the proposed method outperforms the other methods across all metrics without known camera intrinsics in training. In particular, our method achieves significantly lower RPE values, benefiting from the simultaneous prediction of two point maps which enhances temporal consistency.

A second experiment is conducted using 3 full trajectories from the dataset in Table II. We follow the 5-frame evaluation protocol with ATE as the primary metric, consistent with [5]. Three sequences are selected from different synthetic colons in SimCol3D. Results demonstrate that our approach achieves competitive performance compared to both supervised and self-supervised methods, highlighting its robustness across varied trajectories.

C. Monocular Depth Estimation

We provide quantitative evaluations on the SimCol3D and CSD datasets (Table III), comparing our method to single-view method [5] [32] [11] [19] [18] and multi-view depth estimation method [12]. Among them, EndoDAC [19] and DARES [18] are depth foundation model adaptation approaches. Similar to camera pose estimation, the comparison methods are trained on the same datasets as ours using the released code of authors. We evaluated performance using five metrics: absolute relative error, square relative error, root-mean-square error, root-mean-square log error, and threshold accuracy. Notably, only EndoDAC does not require known camera intrinsics, while the other methods rely on ground-truth intrinsics and fixed depth ranges during training. Our method also avoids these dependencies, making the task more challenging, particularly on datasets like SimCol3D, which contain large textureless regions. Despite these challenges, our method achieves superior accuracy across most metrics.

For the CSD dataset, our RMSE is slightly higher than Lite-Mono and ManyDepth. This is because our model doesn't use predefined maximum depth ranges during training. Conventional monocular depth models benefit from these known limits, especially in tubular structures like the colon where inferring distant surfaces can be challenging. Since the CSD dataset mainly composed of tubular structure images with relatively large depth ranges, the performance of our model is impacted by the absence of depth range.

To evaluate generalizability, we conducted a quantitative analysis on the C3VD dataset and a qualitative analysis on the EndoMapper dataset using models trained on the SimCol3D dataset. As shown in Table IV, our method exhibits a slightly higher absolute relative error than DARES but outperforms DARES and other methods across the remaining metrics, indicating more stable and accurate predictions. Qualitative evaluation on the EndoMapper dataset shows improvements

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV: Monocular depth estimation on C3VD dataset. Best results are in **bold**, second-best are underlined.

Method	G.I.	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
AF-SfMLearner	✓	0.152	1.153	6.321	0.190	0.805
Lite-Mono	✓	0.146	1.406	7.210	0.191	0.812
Monodepth2	✓	0.143	1.248	7.241	0.188	0.820
DARES	✓	**0.134**	1.096	6.593	0.175	0.817
ManyDepth	✓	0.190	1.960	8.641	0.260	0.716
EndoDAC	×	0.153	1.380	7.280	0.185	0.812
Dust3R	×	0.390	6.787	13.738	0.411	0.392
Ours	×	0.139	**0.956**	**5.592**	**0.175**	**0.832**

over DUSt3R (Fig. 4), especially in challenging conditions like specularity and textureless regions. In the top row, methods that do not leverage foundation models (e.g., Monodepth2) exhibit noticeable artifacts. Although depth foundation model adaptation methods like DARES and EndoDAC reduce these artifacts, they still struggle to capture fine local structures (highlighted in red boxes). In the second row, the images include challenging conditions such as specularity, textureless region, and bubble. Our method accurately predicts the geometry without introducing artifacts, whereas other methods either introduce artifacts or fail to preserve fine details.

D. Point Map Estimation

We further evaluate our point map estimation results On SimCol3D. We add an extra completeness metric in addition to the three metrics: accuracy, square relative error, and RMSE log used in [5]. Since the self-supervised approaches produce only depth maps, point clouds must be reconstructed by projecting depth predictions into 3D space using camera intrinsics. To ensure a fair comparison, we adopt the same scale normalization strategy used in depth evaluation to mitigate scale ambiguity.

For evaluation on the SimCol3D dataset, we select one method that requires ground-truth intrinsics (AF-SfMLearner), due to its strong performance in Table III, and two methods that do not require ground-truth intrinsics (EndoDAC and DUSt3R). Quantitative results in Table V show that our method outperforms other three methods, demonstrating superior point map estimation accuracy even without access to ground-truth intrinsic parameters.

In addition to quantitative evaluation, we present two qualitative comparisons in Fig. 5 to highlight the improvements of our model over DUSt3R on real colonoscopy images from the EndoMapper dataset. For reconstruction from more than two views, we follow DUSt3R to implement the global alignment. Our method produces high-quality reconstructions and generalizes well to challenging scenarios, including non-Lambertian surfaces and textureless regions.

E. Ablation Study

To assess the effectiveness of key components in our model design, we conduct ablation studies on the SimCol3D dataset, focusing on monocular depth estimation.

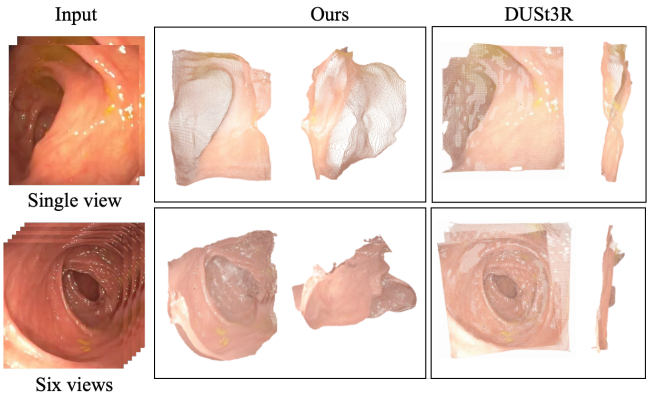

Fig. 5: Qualitative comparison of our predicted 3D point maps and the baseline DUSt3R on real colonoscopy images. In the top row, our method successfully recovers a 3D scene from two images containing textureless and non-Lambertian surfaces, while DUSt3R produces a distorted plane. In the bottom row, our method reconstructs the scene geometry, whereas DUSt3R predicts most of the region as a plane.

(a) Different Feature Fusion

Different strategies for fusing CNN and transformer features yield varying results. The best performance is achieved using a lightweight sequential fusion strategy, as detailed in Table VI. In contrast, more complex attention-based modules, such as feature exchange and CBAM, do not yield improvements. We hypothesize that while attention-based modules can refine representations in some scenarios, they may introduce noise or misaligned information that interferes with the pretrained transformer features.

(b) Impact of Loss Terms and Detail Restoration Module

We evaluate the effects of the two proposed loss terms, confidence-weighted photometric loss and geometry consistency loss, alongside DRM. The first two rows in the table demonstrate that two losses enhance accuracy even without

TABLE V: Point map estimation on SimCol3D

Method	G.I.	Acc. ↓	Comp. ↓	SqRel ↓	RMSE log ↓
AF-SfMLearner	✓	1.779	1.284	0.032	0.116
EndoDAC	×	19.610	8.566	2.463	0.368
DUSt3R	×	41.535	27.028	1.109	0.486
Ours	×	**1.742**	**1.182**	**0.028**	**0.111**

TABLE VI: Ablation study results on fusion strategy

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	σ ↑
Fusion(1)	0.075	0.528	4.666	0.114	0.944
Fusion(2)	0.072	0.459	4.318	0.107	0.952
Fusion(3)	0.062	0.340	3.732	0.094	0.969

TABLE VII: Ablation study results

$\mathcal{L}_{\text{conf-photo}}$	$\mathcal{L}_{\text{geo}}$	DRM	Abs Rel ↓	RMSE ↓	σ ↑
×	×	×	0.204	8.384	0.699
✓	×	×	0.121	5.800	0.881
✓	✓	×	0.116	5.505	0.892
✓	×	✓	0.104	5.300	0.904
×	✓	✓	0.334	11.703	0.437
✓	✓	✓	0.062	3.732	0.969

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

DRM. While the geometry consistency loss is not directly designed for accuracy improvement, it still contributes by ensuring both decoders (through cross-attention) are actively fine-tuned and complementary.

After introducing DRM, these loss terms continue to improve performance. Notably, as shown in the fifth row, removing the confidence-weighted photometric loss leads to performance degradation, even dropping below the baseline foundation model. This underscores the stabilizing role of confidence weighting during training, as it guides the model to prioritize reliable regions.

V. Conclusion

In this paper, we introduce a self-supervised framework for fine-tuning 3D geometric foundation models in the colonoscopy domain. Following most foundation models, our method leverages point map representation to provide geometry information. The introduction of the Detail Restoration Module (DRM) enhances the extraction of fine details, which could also be applied to other ViT-based foundation models. Evaluations on both synthetic and real-world colonoscopy datasets demonstrate strong performance in pose estimation, monocular depth prediction, and dense point map reconstruction. However, utilizing DUSt3R as backbone introduces limitations, including high computational cost for long sequences, which restricts real-time scalability. In future work, we plan to integrate the adapted geometric foundation model into a SLAM framework to enable efficient and consistent 3D colon reconstruction.

References

[1] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: a cancer journal for clinicians*, vol. 73, no. 3, pp. 233–254, 2023.

[2] D. K. Rex *et al.*, "Quality indicators for colonoscopy," *Gastrointestinal Endoscopy*, vol. 81, no. 1, pp. 31–53, 2015.

[3] S. Zhang, L. Zhao, S. Huang, M. Ye, and Q. Hao, "A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 85–95, 2021.

[4] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 20 697–20 709.

[5] S. Shao *et al.*, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue." *Medical Image Analysis*, vol. 77, p. 102338, 2022.

[6] R. Ma *et al.*, "Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy," *Medical Image Analysis*, vol. 72, p. 102100, 2021.

[7] D. Recasens, J. Lamarca, J. M. Fácil, J. M. M. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[8] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 5294–5306.

[9] Z. Lu *et al.*, "Lora3d: Low-rank self-calibration of 3d geometric foundation models," in *International Conference on Learning Representations*, 2025.

[10] W. Li, S. Liu, P. Qiao, and Y. Dou, "Mono3r: Exploiting monocular cues for geometric 3d reconstruction," *arXiv preprint arXiv:2504.13419*, 2025.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[12] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1164–1174.

[13] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6612–6619.

[14] Z. Wang, Y. Zhou, S. He, T. Li, F. Huang, Q. Ding, X. Feng, M. Liu, and Q. Li, "MonoPCC: Photometric-invariant cycle constraint for monocular depth estimation of endoscopic images," *Medical Image Analysis*, vol. 102, p. 103534, 2025.

[15] Z. Wang, M. Chen, Y. Guo, Z. Li, and Q. Yu, "Bridging the domain gap in satellite pose estimation: A self-training approach based on geometrical constraints," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 3, pp. 2500–2514, 2024.

[16] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[17] J. Lu *et al.*, "Align3r: Aligned monocular depth estimation for dynamic videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 22 820–22 830.

[18] M. S. Zeinoddin *et al.*, "Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model," *arXiv preprint arXiv:2408.17433*, 2024.

[19] B. Cui, M. Islam, L. Bai, A. Wang, and H. Ren, " EndoDAC: Efficient Adapting Foundation Model for Self-Supervised Depth Estimation from Any Endoscopic Camera ," in *proceedings of Medical Image Computing and Computer Assisted Intervention*, vol. LNCS 15006, October 2024.

[20] J. Guo *et al.*, "Endo3r: Unified online reconstruction from dynamic monocular endoscopic video," *arXiv preprint arXiv:2504.03198*, 2025.

[21] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5249–5260.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVP Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] Z. Chen *et al.*, "Vision transformer adapter for dense predictions," *International Conference on Learning Representations*, 2023.

[25] M. Yin, Z. Chen, and C. Zhang, "A CNN-Transformer Network Combining CBAM for Change Detection in High-Resolution Remote Sensing Images," *Remote Sensing*, vol. 15, no. 9, p. 2406, 2023.

[26] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models." in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3813–3824.

[27] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 159–12 168.

[28] A. Rau *et al.*, "Simcol3d — 3d reconstruction during colonoscopy challenge," *Medical Image Analysis*, vol. 96, p. 103195, 2024.

[29] P. Azagra *et al.*, "Endomapper dataset of complete calibrated endoscopy procedures," *Scientific Data*, vol. 10, no. 1, Oct. 2023.

[30] T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr, "Colonoscopy 3d video dataset with paired depth from 2d-3d registration," *Medical Image Analysis*, p. 102956, 2023.

[31] J. Zhang *et al.*, "Monst3r: A simple approach for estimating geometry in the presence of motion," in *International Conference on Learning Representations*, 2025.

[32] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "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 (CVPR)*, June 2023, pp. 18 537–18 546.