Article:
Xu, Z., Li, B., Hu, Y. et al. (2026) Self-supervised Monocular Depth and Pose Estimation for Endoscopy with Latent Priors. IEEE Transactions on Medical Imaging. ISSN: 0278-0062

https://doi.org/10.1109/tmi.2026.3671423

Self-supervised Monocular Depth and Pose Estimation for Endoscopy with Latent Priors

Ziang Xu, Bin Li, Yang Hu, Chenyu Zhang, James East, Sharib Ali*, *Member, IEEE* and Jens Rittscher

***Abstract*— Accurate 3D reconstruction in endoscopy enables quantitative and holistic lesion characterization within the gastrointestinal (GI) tract. To achieve this, reliable depth and pose estimation is required. However, endoscopy systems are monocular, and existing methods relying on synthetic datasets or complex models often lack generalizability in challenging endoscopic conditions. We propose a robust self-supervised monocular depth and pose estimation framework that incorporates a StyleGAN-based generator and a Variational Autoencoder (VAE). The StyleGAN generator leverages extensive depth scenes from natural images to condition the depth network, enhancing realism and robustness of depth predictions through latent feature priors. For pose estimation, we reformulate it within a VAE framework, treating pose transitions as latent variables to regularize scale, stabilize z-axis prominence, and improve x-y sensitivity. To further enhance pose stability and generalizability, we introduce a prior transfer module that distills motion knowledge from natural scene SLAM systems. Specifically, pose priors from a pretrained SLAM model—supervised on large-scale natural scene datasets—are used to guide the latent distribution of pose through a KL-divergence reparameterization. This mechanism effectively transfers structural motion priors into the endoscopic domain, improving trajectory consistency under challenging conditions. This dual refinement pipeline enables accurate depth and pose predictions, effectively addressing the GI tract's complex textures and lighting. Extensive evaluations on SimCol, C3VD, and EndoSLAM datasets confirm our framework's superior performance over published self-supervised methods in endoscopic depth and pose estimation. All data descriptions and code are available at https://github.com/EricXuziang/Self-supervised-with-Latent-Priors.git.**

***Index Terms*— Self-supervised learning, deep learning, endoscopy, monocular depth and pose estimation**

I. Introduction

Colorectal cancer (CRC) is a major global health concern, ranking as the third most common cancer worldwide and accounting for approximately 35% of cancer-related deaths [1]. Accurate lesion mapping in colonoscopy is essential for effective CRC diagnosis and treatment but remains challenging due to its dependency on endoscopist expertise [2]. By estimating the camera's distance to the mucosal surface (depth) and reliably computing its ego-motion (relative pose), it becomes feasible to reconstruct the scene in 3D [3]. In the context of colonoscopy, 2D video sequences can be used to recover depth and relative camera pose for 3D scene reconstruction [4], as shown in Fig. 1. This enables lesion localization, size estimation, and interactive visualization as an objective and measurable 3D scene. 3D reconstruction thus provides a powerful and consistent method for precise lesion mapping and structural assessment, thereby enhancing clinical accuracy [5].

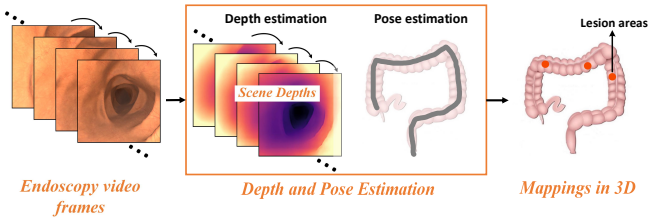

Fig. 1. Workflow for 3D lesion mapping in endoscopy: depth estimation generates scene depth maps from monocular video frames, which combined with pose trajectory, allow 3D reconstruction of the colon and precise lesion localization for improved diagnostics and surgical planning.

Effective 3D reconstruction in endoscopy relies on depth and pose estimation. Depth estimation provides spatial information, while pose estimation tracks camera orientation and movement through the gastrointestinal (GI) tract. However, reliable ground truth for depth and pose are difficult to obtain due to the dynamic, constricted nature of the GI tract. Additionally, endoscopy typically relies on monocular imaging, which limits depth perception and complicates 3D mapping in the GI tract's variable and occluded environment.

Recent research has addressed these challenges with synthetic datasets, self-supervised learning, and monocular approaches [6]–[9]. Synthetic datasets enable training by simulating endoscopic conditions, while self-supervised learning leverage temporal and spatial cues to estimate depth and pose without manual labels. While stereo endoscopes are being developed they are not being used in routine practice. Hence, self-supervised monocular depth estimation is the preferred approach, framing depth estimation as a view synthesis problem and optimizing image reconstruction to avoid traditional depth labels. Monocular systems rely on a single view and require

Ziang Xu, B. Li, C. Zhang, and J. Rittscher are with the Department of Engineering, University of Oxford, Oxford, UK; J. East is with Nuffield Department of Clinical Medicine, Experimental Medicine Division, University of Oxford, Oxford, UK. Y. Hu is with the University of Leicester, UK. S. Ali is with School of Computer Science, University of Leeds, Leeds, LS2 9JT, UK. Ziang Xu is now with the Chinese University of Hong Kong, Hong Kong (corresponding email: s.s.ali@leeds.ac.uk and ziangxu@cuhk.edu.hk).

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number UKRI914] and by the Academy of Medical Sciences through the Springboard scheme (SBF0010\1191), the Oxford Branch of Ludwig Cancer Research and National Institute for Health Research (NIHR) Oxford Biomedical Research Centre.

J. Rittscher and S. Ali: Shared senior (last) authorship

an additional pose network to infer motion, adding complexity to compensate for limited depth cues.

Despite recent advances in monocular depth and pose estimation, such as Monodepth [3], [10], Depth Anything [11], and DepthPro [12], current methods struggle in 3D colonoscopy due to the unique challenges of the endoscopic environment. This includes complex and often homogeneous textures, and large illumination variabilities. In addition, most of the techniques in colonoscopy [4], [13], [14] are trained on synthetic data relying on the established ground truth and highly curated data from an experimental setup. This limits model generalizability on real-world colonoscopy data. Also, most of these methods do not utilize sequence data in their training. To address these issues, we propose a robust approach that incorporates *generative latent priors* into a joint self-supervised framework for depth and pose estimation, conditioning predictions to overcome these challenges. Here, generative latent priors refer to the latent distribution learned by a pretrained generative adversarial network, specifically, a StyleGAN model trained on natural scene data. Our encoder produces a compact latent representation that is mapped into the StyleGAN intermediate latent space ($\mathcal{W}$ space) and used to condition the pretrained generator. By supplying encoder-derived latent codes to the generator, the model leverages the rich spatial structure and natural image statistics captured during StyleGAN's large-scale training, resulting in more stable and reliable depth prediction, particularly in low-texture, low-light, or geometrically ambiguous regions.

Our self-supervised backbone builds on Monodepth2's principles [3], where a depth network estimates the current scene's depth and a pose network predicts the relative pose transitions between frames. Then, reprojected adjacent frames are warped to predict the current frame, forming a self-supervised loop by comparing this prediction with the actual frame. However, Monodepth2's depth and pose predictions are driven solely by reprojection consistency, lacking conditioning for realistic scene depth or camera poses. In our work we introduce a variational autoencoder (VAE) technique to minimize the distribution variation between estimated pose and the pose obtained from an off-the-shelf SLAM-based model pre-trained (DROID-SLAM) [15] on large natural scene datasets.

Key contributions of our work are summarised below:

- We propose a depth estimation framework composed of an encoder, a pretrained StyleGAN-based generator, and a decoder. The StyleGAN generator is trained on depth images to learn a structured latent space that captures depth priors. Given an input image, an encoder predicts latent codes in this learned latent space, which are used to condition the StyleGAN generator. The generator produces intermediate depth-aware representations, which, together with encoder features, condition a decoder that synthesizes the final high-fidelity depth prediction.
- For pose estimation, we introduce a VAE-based formulation, where the encoder corresponds to the pose network and the decoder corresponds to the reprojection algorithm, treating poses as latent variables to regularise pose trajectories.
- We incorporate informative priors from DROID-SLAM [15] pretrained on natural scenes, transferring its latent motion understanding to the medical domain. A KL-divergence guided regularization enhances geometric consistency, particularly in textureless and high-speed endoscopic sequences.
- Extensive experiments on SimCol [13], C3VD [7], and EndoSLAM [16] datasets demonstrate superior performance compared to existing methods, with ablation studies confirming the impact of each component.

II. Related work

Most clinical endoscopy platforms are monocular, so our review focuses on key advances in monocular depth and pose estimation and their application in endoscopy. Monocular depth estimation is inherently challenging and ill-posed, as a single 2D image may correspond to multiple 3D scenes. Supervised deep learning approaches leverage accurate depth labels as supervision, enabling models to learn the relationship between RGB images and depth values. Eigen et al. [17] propose a dual-network model, where one network makes a global prediction, refined by a second network. Performance improvements come from novel loss functions, such as the Huber loss [18] and scale-invariant loss [19], as well as by framing depth estimation as a multi-class regression task.

Monocular video sequences can be used as self-supervision signals, but require the network to learn both depth and camera pose. To optimise model effects others have introduced additional constraints such uncertainty [20], normal consistency [21], semantic segmentation [22], and visual odometry [23]. Remarkably, Monodepth2 [3] enhances model convergence speed and accuracy without additional constraints by optimizing minimum reprojection loss and introducing automatic masking to handle static objects. Several subsequent methods build on Monodepth2, including DualRefine [24], MonoViT [25], and Lite-Mono [26].

Real depth maps are challenging to obtain in endoscopy due to the need for specialized depth sensors and the variability in measurement quality [27]. Synthetic data can address this challenge by generating realistic simulated scenes along with corresponding depth maps, thereby overcoming the limitations of acquiring real-world depth maps and providing high-quality training data for deep learning models. To improve depth and pose estimation in endoscopy, Mahmood et al. [4] proposed a joint CNN-CRF framework trained on synthetic datasets with ground truth and adapted through adversarial training. Generative approaches, such as those by Rau et al. [13] and Mahmood et al. [28], employed GANs to generate depth maps, though direct GAN-generated maps often lack accuracy, and supervised learning is impractical given the need for extensive ground truth data in endoscopy.

Self-supervised learning methods, which reframe monocular depth estimation as a view synthesis problem, have shown promise by eliminating the need for ground truth depth maps. Hwang et al. [29] proposed a depth feedback network using self-supervised neighboring frame depth prediction with reconstruction error computation. Ozyoruk et al. [16] combined

a residual network with spatial attention and a luminance-aware loss to improve robustness under varying illumination and introduced the EndoSLAM dataset. Shao et al. [30] introduced a unified self-supervised framework for monocular depth and ego-motion estimation in endoscopic scenes, which incorporates a novel concept called appearance flow to handle brightness inconsistencies caused by illumination changes. Liu et al. [9] presented a self-supervised monocular depth estimation model with dual attention, using multi-scale structural similarity and L_1 losses to maintain luminance and color invariance. Yang et al. [31] proposed a lightweight network that tightly couples CNN and Transformer modules within a hierarchical encoder for monocular depth estimation. Instead of fusing features at the deepest layer, their method extracts multi-scale features using CNNs for local texture and Transformers for global shape, and enhances the pose network with multi-head attention to improve motion prediction accuracy.

A novel framework based on intrinsic image decomposition is proposed by Li et al. [32], which integrates unsupervised depth estimation with intrinsic decomposition to address challenges in endoscopic imaging. The framework includes an image intrinsic decomposition module and a synthesis reconstruction module, designed to work seamlessly for accurate depth prediction. By leveraging the albedo map from intrinsic decomposition, it avoids the need for image pre-processing. Rodriguez et al. [33] presented LightDepth, which utilized the inverse-square relationship between pixel brightness and surface distance, leveraging the co-located camera and light source in endoscopic devices. The method demonstrates competitive performance. EndoDAC, proposed by Cui et al. [34], introduced a self-supervised adaptation framework for depth estimation in endoscopic surgery, which effectively transfers foundation models to the medical domain. This method leverages a Dynamic Low-Rank Adaptation module and Convolutional Neck blocks to achieve depth estimation tailored to endoscopic scenes with a remarkably small number of trainable parameters.

In summary, existing methods for endoscopic depth and pose estimation remain limited in the context of colonoscopy. Supervised approaches [4], [13], [28] rely on synthetic or GAN-generated depth data due to the lack of real ground truth [27], but they suffer from a significant domain gap when applied to real colonoscopy videos. Self-supervised methods [9], [29]–[34] mitigate this dependency by leveraging photometric consistency, intrinsic decomposition, or foundation model adaptation. Nevertheless, these strategies often fail in textureless, specular, or poorly illuminated regions of colon, where image-based cues are unreliable. Furthermore, none of these methods explicitly enforce domain-invariant depth representations or provide robust pose regularization tailored to the quasi-static and deformable characteristics of colonoscopic scenes. To overcome these limitations, we propose a novel depth estimation framework that integrates an encoder–decoder architecture with a pretrained StyleGAN prior, enabling domain-invariant and structurally consistent depth recovery. Our framework is coupled with a VAE-based pose estimation formulation that regularizes trajectories by treating poses as latent variables. We further transfer geometric priors from DROID-SLAM enhanced with a KL-divergence regularization that ensures the consistency in textureless scenes.

III. Methodology

Our model is designed to estimate depth and camera pose for reconstructing endoscopic movement traces and colon tract in a monocular setting. Briefly, our model comprises two main branches, DepthNet and PoseNet, which jointly predict depth and pose in a self-supervised framework, following the principles of Monodepth2 [3]. Our architecture follows an encoding-decoding process (Fig 2). Specifically, DepthNet encodes the current frame's depth, while PoseNet encodes the camera poses of adjacent frames (frames +1 and -1). The encoded outputs are used in a reprojection step that warps adjacent RGB frames to reconstruct the current frame (frame 0). A reconstruction loss is then calculated between the warped output and the ground truth frame, completing the self-supervised learning loop.

To enhance the encoding process, we inject StyleGAN-derived generative latent priors into both depth and pose branches. For depth encoding, we leverage a pretrained StyleGAN generator [35] that captures structured depth priors, guiding the network to produce realistic depth maps. For pose encoding, we incorporate a variational autoencoder (VAE) where the pose network serves as the encoder, with predicted poses as latent variables and the reprojection process as the decoder. To regularize the latent space, we enforce a KL divergence term that constrains the predicted pose distribution to a prior. Instead of using a standard Gaussian prior, we guide the latent space with pose estimates from a pretrained DROID-SLAM [15], transferring structural motion priors from natural scene datasets. This prior regularization enforces smooth, scale-consistent pose transitions across x-, y-, and z-axis. The following sections detail each model component.

A. DepthNet with StyleGAN-based generator

1) Encoder and Decoder

Our DepthNet (as illustrated in Fig. 2(a)) consists of an encoder $\mathbf{E_{dep}}$, a depth prior generated by StyleGAN-based generator $\mathbf{L_{depth}}$, and a decoder $\mathbf{D_{dep}}$. The input RGB image, $\mathbf{x}$, is processed by the encoder, which comprises a series of downsampling and convolutional layers, producing a set of output feature maps at decreasing resolutions, $\mathbf{h} = \{h^{n-1}, \ldots, h^0\}$, where n represents the number of resolutions.

Following the design and implementation of a StyleGAN generator [36], feature generation is modulated by a set of resolution-specific latent (style) vectors. Starting from an initial latent code, progressive upsampling is performed through a sequence of generator blocks. At each resolution, the corresponding style vector modulates the feature maps via Adaptive Instance Normalization (AdaIN) [37].

The feature maps produced by the DepthNet encoder, i.e., $\mathbf{h} = \{h^{n-1}, \ldots, h^0\}$, are treated as resolution-specific latent vectors and fused into the progressive upscaling cascade at their corresponding resolutions (Fig 2, top panel). This process generates a set of output feature maps, $\mathbf{a} =$

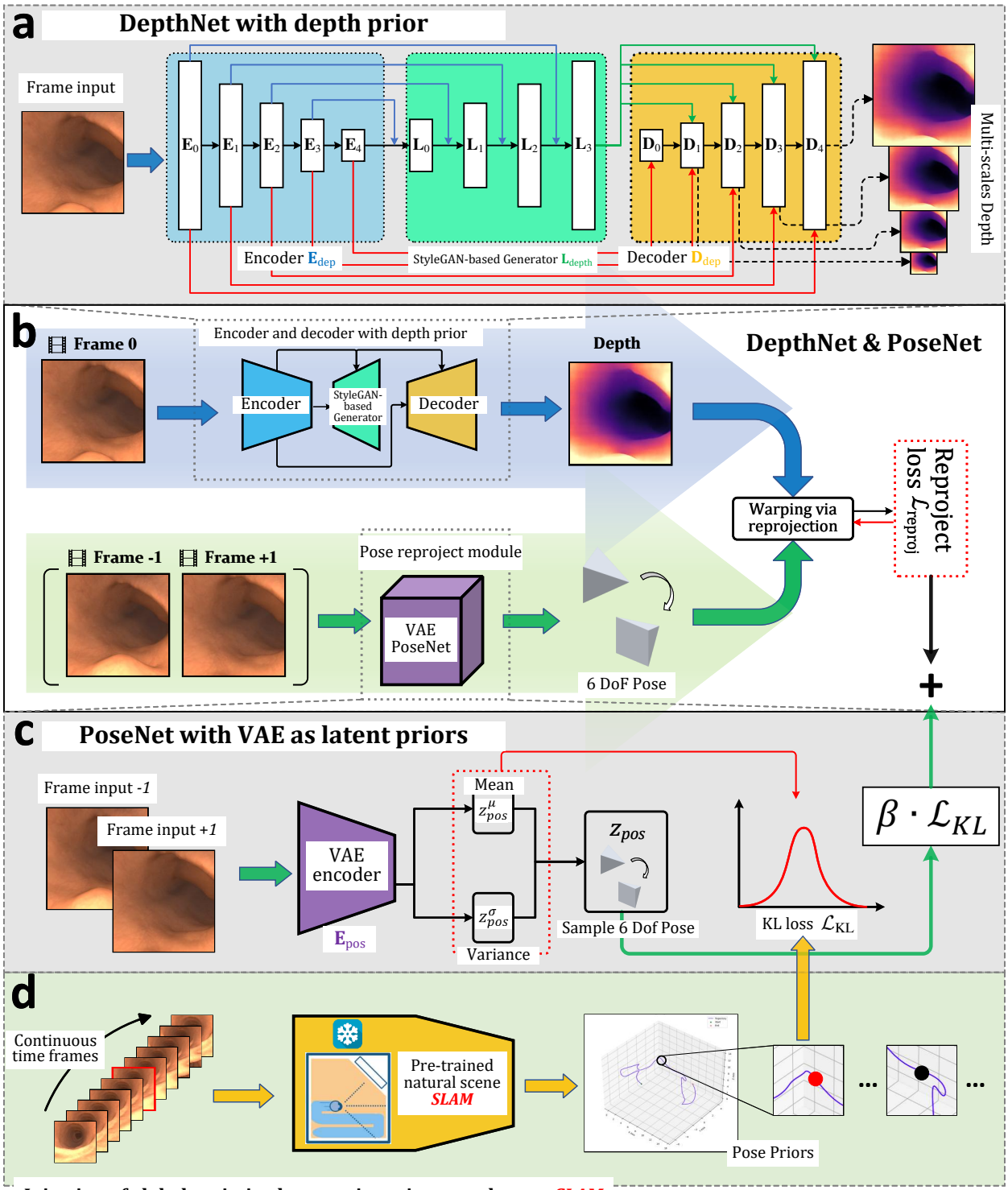

Fig. 2. **Overview of the proposed method.** The method consists of a depth estimation network with a pre-trained StyleGAN-based generator and a VAE-constrained pose estimation network. The entire method is self-supervised training through subsequent reprojection as a supervision signal. (a) Our DepthNet integrates a pretrained StyleGAN-based generator $\mathbf{L_{depth}}$ encoding depth priors from natural scenes. Features from encoder $\mathbf{E_{dep}}$ are injected into the StyleGAN-based generator via AdaIN at multiple scales, enabling depth-aware modulation. The decoder $\mathbf{D_{dep}}$ fuses prior-informed features with encoder outputs to predict depth at multiple scales. (b) Overview of the end-to-end monocular framework for joint depth and pose estimation. DepthNet and PoseNet predict depth and relative motion, enabling self-supervised learning via reprojection-based reconstruction of the current frame. (c) The VAE-constrained PoseNet $\mathbf{E_{pos}}$ encodes relative poses as probabilistic latent variables. It predicts both mean $\mathbf{z^{\mu}_{pos}}$ and variance $\mathbf{z^{\sigma}_{pos}}$, enabling sampling via reparameterization. A KL term enforces smooth, scale-consistent pose distributions for improved temporal stability. (d) This module injects global pose priors from pretrained DROID-SLAM [15]. Instead of using a standard Gaussian prior, the predicted poses are regularized toward SLAM-informed distributions, enhancing temporal consistency and robustness in texture-limited endoscopic settings.The overall framework is jointly optimized by the KL divergence loss $\mathcal{L}_{\text{KL}}$ and the reprojection loss $\mathcal{L}_{\text{reproj}}$.

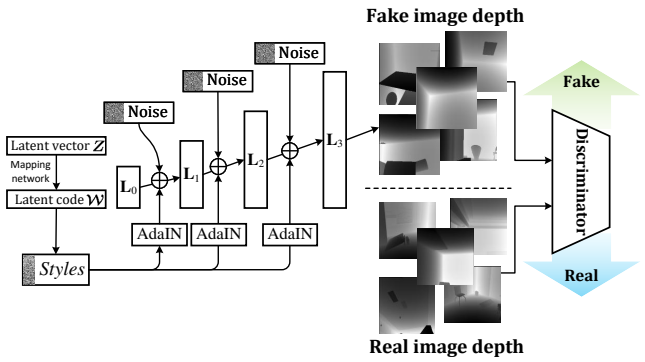

Fig. 3. The pre-training process of the StyleGAN prior in our framework. A Gaussian latent vector ($\boldsymbol{Z}$) is first passed through a mapping network to produce a latent code ($\boldsymbol{\mathcal{W}}$), which is then used to modulate features at each resolution ($\mathsf{L_0}, \mathsf{L_1}, \mathsf{L_2}, \mathsf{L_3}$)) via adaptive instance normalization (AdaIN). In this context, [36]. styles refer to the resolution-specific affine transformation parameters (scale and bias) derived from the latent code, which control visual characteristics of the generated depth maps. Coarse layers influence global geometric structure, middle layers affect medium-scale features such as depth contrast and organ structure, and fine layers control surface texture and local structural details. Trained within a GAN framework, the StyleGAN prior produces realistic depth maps, while a discriminator refines generation quality by distinguishing real from synthetic maps.

$\{a^{n-1}, \ldots, a^0\}$, at each resolution, which are then concatenated with the decoder $\mathbf{D_{dep}}$ at the corresponding resolutions.

The decoder $\mathbf{D_{dep}}$ receives both the feature maps from the encoder ($\mathbf{h} = h^k$), and the feature maps modulated by the latent vectors, $\mathbf{a} = a^k$, at the corresponding resolutions. It produces depth predictions, $\mathbf{d} = \mathbf{D_{dep}}([h^k, a^k], k \in \{n-1, 0\}) = \{d^0, d^1, d^2, d^3\}$, at the four largest resolutions. These correspond to depth predictions at four different scales, as illustrated in Fig 2 top panel. The key idea is to leverage the structured depth priors learned by the pretrained StyleGAN generator and reformulate depth prediction as a latent code conditioning process. The depth encoder $\mathbf{E_{dep}}$ extracts features from the RGB images and maps them into the generator's latent space, producing latent codes that condition the pretrained StyleGAN prior, $\mathbf{L_{depth}}$, to form a depth-aware representation. The latent codes retrieved from $\mathbf{L_{depth}}$ are then used in the decoding process to reconstruct depth maps.

By pretraining StyleGAN-based generator $\mathbf{L_{depth}}$ with a large amount of scene depth data from other domains with accessible ground truth depth data, we condition the depth generation process. This encourages the decoder to produce outputs that resemble realistic depth maps.

2) StyleGAN-based generator

Generative pretraining mode: The pretrained network $\mathbf{L_{depth}}$ follows a StyleGAN-like architecture [36]. The core idea is to enhance the depth prediction process by injecting prior knowledge of scene depths, obtained from domains where depth data is abundant (e.g., natural scenes), which is often easier to acquire than in specialized domains such as endoscopy.

During training, a random Gaussian latent vector is generated and passed through a series of upsampling and modulated convolution blocks. At each resolution level, a new Gaussian noise vector is combined with the features via Adaptive Instance Normalization (AdaIN), adding variability to the outputs. After upscaling to the target resolution, a final convolutional block reduces the channels to 1, producing the output depth map (Fig. 3). The discriminator is a standard residual CNN trained to classify real vs. synthetic depth maps generated by the pretrained generator ($\mathbf{L_{depth}}$). The discriminator takes in a depth map and outputs a binary classification score, where a score of 1 indicates a real depth map, and a score of 0 indicates a synthetic one.

The goal is to optimize the pretrained StyleGAN-based generator ($\mathbf{L_{depth}}$) to generate realistic depth maps, while the discriminator is trained to accurately distinguish real from generated depth maps. Following the original StyeGAN implementation, we include a gradient penalty term to stabilize the training [38]. The training process alternates between minimizing the generator's loss and maximizing the discriminator's classification accuracy.

StyleGAN-based Generator: After the StyleGAN-based generator $\mathbf{L_{depth}}$ is trained to produce realistic scene depth outputs, its weights are frozen, and it is integrated into the encoder-decoder architecture discussed in section III-A.1. Note that a new set of trainable Adaptive Instance Normalization (AdaIN) blocks is initialized and optimized during the downstream training of the depth encoder $\mathbf{E_{dep}}$ and the decoder $\mathbf{D_{dep}}$. The output of the decoder $\mathbf{d} = \mathbf{D_{dep}}[\mathbf{h} = \mathbf{E_{dep}}(x), \mathbf{a} = \mathbf{L_{depth}}(h)]$ will be fed to the reprojection algorithm to reconstruct the input RGB frame and the training process of $\mathbf{D_{dep}}$ and $\mathbf{E_{dep}}$ will be discussed in section III-C.

B. VAE-constrained PoseNet

Regularizing the scale of pose estimation via VAE. For generating pose predictions, we utilize a pose encoder $\mathbf{E_{pos}}$ (as illustrated in Fig. 2(c)) that takes an RGB input and generates a pose estimation, represented as a vector of shape $\mathbb{R}^{6\times1}$, consisting of six pose parameters. The pose estimation is computed separately for both the previous frame (-1) and the subsequent frame ($+1$). After obtaining the pose estimations from $\mathbf{E_{pos}}$, these estimated poses are input to the reprojection algorithm along with the depth output d_i from DepthNet. The reprojection algorithm, serving as the decoder in this context, takes the estimated poses and depth maps and warps the RGB images of the -1 and $+1$ frames to produce reprojected RGB images at the current frame. The reprojected images are then compared with the ground truth RGB image of the current frame using MSE loss, ensuring spatial temporal consistency and completing the self-supervision [3].

The key difference in this approach is that we treat the output of PoseNet, which represents the relative pose differences between adjacent frames (-1 and +1) to the current frame, as latent variables in a VAE [39], [40]. Specifically, the VAE encoder $\mathbf{E_{pos}}$ predicts both the mean $\mathbf{z}^{\mu}_{\mathbf{pos}}$ and variance $\mathbf{z}^{\sigma}_{\mathbf{pos}}$ of the latent pose distribution, from which we sample pose representations via the reparameterization. The reprojection algorithm acts as the decoder in this setup, constraining these relative pose parameters by enforcing a KL divergence between the predicted pose differences and a

Gaussian prior. This regularization suppresses the prominence in z-axis movements while improves the relative sensitivity of x-y pose changes between adjacent frames [41], reflecting prior knowledge that endoscopic movements are typically smooth to minimize potential damage to the GI tract.

Transferring latent priors of pose estimation from natural scene SLAM. To further improve the quality and stability of pose estimation in our self-supervised framework, we incorporate external priors learned from a large-scale supervised system, DROID-SLAM [15], trained on natural scene datasets (as shown in Fig. 2(d)). Specifically, we use a pretrained DROID-SLAM to obtain initial pose estimates between adjacent frames in the endoscopic video. DROID-SLAM was trained with ground truth supervision on natural scenes datasets and optimized globally using bundle adjustment [42]. This kind of training is not feasible in endoscopic scenarios due to the lack of ground truth and we leverage the predictions made based on a natural scenes trained system as informative priors.

We regularize the predicted latent poses to match a Gaussian centered at the pose estimates from the petrained DROID-SLAM, i.e., $\mathcal{N}(\mu_{\text{SLAM}}, \mathbf{I})$. This reparameterized prior effectively transfers structural motion knowledge from natural scenes to the endoscopy domain. The VAE formulation enables our model to flexibly adapt and refine this prior through gradient-based learning, balancing between the external SLAM-informed prior and the photometric consistency constraints from self-supervision. The overall training objective thus consists of two principal components: (1) a photometric reconstruction loss to enforce image-level consistency across frames, and (2) a KL divergence loss to align the predicted pose distribution with informative SLAM-transferred priors. The overall optimization objective consists of two main components:

Reprojection Loss: This is computed as the mean squared error (MSE) between the ground truth RGB image $\mathbf{x}_{\text{gt}}$ at the current frame and the reprojected RGB image $\mathbf{x}_{\text{reproj}}$, obtained by warping adjacent frames using the estimated poses and depth (see Section III-C). The reprojected image is written as:

$$\mathbf{x}_{reproj} = Reproj\big(\mathbf{x}_{\text{src}}, \mathbf{D}_{\text{tgt}}, \mathbf{T}_{\text{src}\to\text{tgt}}\big), \tag{1}$$

where $\mathbf{x}_{\text{src}}$ denotes the source (adjacent) RGB image, $\mathbf{D}_{\text{tgt}}$ is the predicted depth map of the target (current) frame, and $\mathbf{T}_{\text{src}\to\text{tgt}} \in \text{SE}(3)$ represents the relative pose that transforms 3D points from the source camera coordinate system to the target one. The reprojection function $\text{Reproj}(\cdot)$ is implemented via differentiable inverse warping: for each pixel (u, v) in the target image, the corresponding 3D point is first reconstructed by back-projecting (u, v) using the inverse camera intrinsics $\mathbf{K}^{-1}$ and the depth $\mathbf{D}_{\text{tgt}}(u, v)$, then transformed into the source camera frame via $\mathbf{T}_{\text{src}\to\text{tgt}}$, and finally reprojected onto the source image plane using the camera projection $\pi(X, Y, Z) = (X/Z, Y/Z)$; the resulting non-integer pixel coordinates are used to bilinearly sample the RGB value from $\mathbf{x}_{\text{src}}$. The reprojection loss is then given by:

$$\mathcal{L}_{reproj} = \left\|\mathbf{x}_{\text{gt}} - \mathbf{x}_{reproj}\right\|_2^2. \tag{2}$$

KL Divergence Regularization: The KL divergence between the predicted pose parameters and a Gaussian prior, which smooths the changes in pose estimates:

$$\mathcal{L}_{\text{KL}} = \text{KL}\left(q(\mathbf{z_{pos}}) \| \mathcal{N}(\mu_{\text{SLAM}}, \mathbf{I})\right), \tag{3}$$

where $q(\mathbf{z_{pos}}) : \mathbf{z_{pos}} \sim \mathcal{N}(\mathbf{z}^{\mu}_{\mathbf{pos}}, \mathbf{z}^{\boldsymbol{\sigma^2}}_{\mathbf{pos}})$ represents the distribution of pose parameters estimated by PoseNet, and $\mathcal{N}(\mu_{\text{SLAM}}, \mathbf{I})$ denotes a standard Gaussian prior. Total Loss for the VAE-constrained PoseNet: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reproj}} + \beta \cdot \mathcal{L}_{\text{KL}}$, where $\beta = 0.1$ controls the weight for KL divergence.

C. Self-supervised Reprojection

The reprojection algorithm takes the depth map d_i (output from DepthNet for the current frame), the RGB images of the adjacent frames ($\mathbf{x}_{-1}$ and $\mathbf{x}_{+1}$), and the estimated pose differences $\mathbf{z_{pos,-1}}$ and $\mathbf{z_{pos,+1}}$ (via PoseNet) that specify the relative pose differences from the current frame to the adjacent frames. The algorithm uses these inputs to warp and interpolate the RGB images of -1 and $+1$ frames, generating the reprojected RGB image of the current frame, $\mathbf{I}_{\text{reproj}}$, same as Monodepth2 [3], where the reprojection per-pixel loss is computed as the minimum photometric error over the adjacent source images to handle disocclusion. The overall loss function for training all sub-networks: DepthNet (with $\mathbf{E_{dep}}$ and $\mathbf{D_{dep}}$) and PoseNet (with $\mathbf{E_{pos}}$) is given by:

$$\begin{aligned}\mathcal{L}_{\text{total}}(\theta_{\text{dep}}, \theta_{\text{pos}}) = \ & \min\Big(\left\|\mathbf{x}_0 - \text{Reproj}(\mathbf{x}_{-1}, \mathbf{z_{pos,-1}}, \mathbf{d}_i)\right\|_2^2, \\ & \left\|\mathbf{x}_0 - \text{Reproj}(\mathbf{x}_{+1}, \mathbf{z_{pos,+1}}, \mathbf{d}_i)\right\|_2^2\Big) \\ & + \beta \text{KL}\big(q(\mathbf{z_{pos}}) \| \mathcal{N}(\mu_{\text{SLAM}}, \mathbf{I})\big),\end{aligned} \tag{4}$$

where $\mathbf{x}_0$ denotes the current RGB frame. The estimated pose differences between the adjacent frames and the current frame are represented as $\mathbf{z_{pos,-1}}$ and $\mathbf{z_{pos,+1}}$ for the previous and next frames, respectively. The reprojection operation, $\text{Reproj}(\cdot)$, utilizes these pose differences along with the depth estimates to reconstruct the current view. The distribution of the estimated poses is modeled as $q(\mathbf{z_{pos}})$, and the KL divergence between this distribution and the prior is weighted by a coefficient β, which is set to 0.1 in our experiments.

IV. Experiments

A. Datasets

Existing research often uses synthetic data generated from 3D models, as these provide RGB images with corresponding ground truth depth maps to evaluate endoscopic depth estimation models. Three main datasets are included: the SimCol [43] and C3VD [7] and EndoSLAM datasets [16].

The SimCol dataset, created by Rau et al. [43] using CT scans of the human colon and rendered in Unity, includes 33 scenes with a total of 37,833 RGB frames, each paired with ground truth depth maps and camera poses. Depth values are scaled to [0, 1] representing [0, 20] cm. We utilized the SimCol dataset, specifically SyntheticColon I and SyntheticColon II, for the training, validation, and testing sets. Trajectory S1–S3, S6–S8, S11–S13, B1–B3, B6–B8, and B11–B13 were used

TABLE I

COMPARISON OF DEPTH ESTIMATION METHODS ACROSS VARIOUS METRICS ON THE SIMCOL DATASET.

Dataset	Method	Year	Depth Error (↓) (in cm)					Depth Accuracy (↑)		
			Abs Rel	Sq Rel	RMSE	RMSE log	L_1 error	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SimCol-I & II	Monodepth2 [3]	2019	0.151	0.209	0.709	0.200	0.443	0.837	0.946	0.979
	MonoViT [25]	2022	0.138	0.234	0.715	0.157	0.423	0.863	0.952	0.952
	DualRefine [24]	2023	0.133	0.289	0.717	0.152	0.425	0.871	0.944	0.974
	Lite-Mono [26]	2023	0.126	0.184	0.640	0.144	0.358	0.886	0.969	0.990
	Depth Pro† [12]	2024	0.174	0.137	0.708	0.211	0.479	0.734	0.949	0.988
	Endo-SFMLearner* [16]	2021	0.137	0.202	0.689	0.149	0.407	0.891	0.955	0.981
	AF-SFMLearner* [30]	2023	0.129	0.181	0.656	0.141	0.359	0.879	0.975	0.991
	IID-SfmLearner* [32]	2024	0.131	0.199	0.669	0.143	0.372	0.899	0.961	0.980
	EndoDAC* [34]	2024	0.127	0.165	0.647	0.135	0.361	0.911	0.980	0.991
	Ours	2025	**0.121**	**0.110**	**0.608**	**0.123**	**0.339**	**0.940**	**0.983**	**0.996**
SimCol-III	Monodepth2 [3]	2019	0.135	0.137	0.624	0.187	0.412	0.840	0.937	0.957
	MonoViT [25]	2022	0.128	0.125	0.595	0.169	0.400	0.858	0.957	0.981
	DualRefine [24]	2023	0.115	0.177	0.577	0.159	0.392	0.889	0.947	0.966
	Lite-Mono [26]	2023	0.121	0.091	0.539	0.158	0.368	0.872	0.966	0.987
	Depth Pro† [12]	2024	0.178	0.150	0.757	0.216	0.547	0.704	0.948	0.990
	Endo-SFMLearner* [16]	2021	0.133	0.112	0.566	0.172	0.396	0.884	0.951	0.979
	AF-SFMLearner* [30]	2023	0.119	0.105	0.537	0.151	0.361	0.891	0.959	0.989
	IID-SfmLearner* [32]	2024	0.124	0.109	0.542	0.157	0.373	0.889	0.957	0.982
	EndoDAC* [34]	2024	0.116	0.099	0.521	0.149	0.350	0.901	0.969	0.989
	Ours	2025	**0.110**	**0.079**	**0.501**	**0.119**	**0.320**	**0.931**	**0.983**	**0.992**

† Foundation model trained on large-scale natural images; * Endoscopy-specific self-supervised method

TABLE II

COMPARISON OF POSE ESTIMATION METHODS ACROSS VARIOUS METRICS ON SIMCOL DATASET(SYNTHETIC).

Dataset	Method	Pose Error (↓)		
		ATE (in cm)	Trans. RPE (in cm)	Rot. RPE (deg)
SimCol-I & II	Monodepth2 [3]	0.5502	0.1371	0.3632
	DualRefine [24]	0.5224	0.1216	0.2912
	MonoViT [25]	0.5331	0.1279	0.3260
	Lite-Mono [26]	0.5154	0.1208	0.2775
	DROID-SLAM [15]	0.4577	0.1093	0.1769
	Endo-SFMLearner* [16]	0.5069	0.1176	0.1977
	AF-SFMLearner* [30]	0.5016	0.1088	0.1943
	IID-SfmLearner* [32]	0.5139	0.1141	0.209
	EndoDAC* [34]	0.5027	0.1091	0.1877
	Ours	**0.4209**	**0.0833**	**0.1077**
SimCol-III	Monodepth2 [3]	0.5363	0.1099	0.3235
	DualRefine [24]	0.4644	0.1043	0.2959
	MonoViT [25]	0.4818	0.1106	0.3072
	Lite-Mono [26]	0.4541	0.0864	0.2833
	DROID-SLAM [15]	0.4429	0.0933	0.2636
	Endo-SFMLearner* [16]	0.4511	0.1046	0.2951
	AF-SFMLearner* [30]	0.4502	0.0847	0.2715
	IID-SfmLearner* [32]	0.4677	0.0936	0.2815
	EndoDAC* [34]	0.4519	0.0922	0.2706
	Ours	**0.4246**	**0.0736**	**0.2273**

as the training set. Trajectory S4, S9, S14, B4, B9, and B14 were designated as the validation set. Trajectory S5, S10, S15, B5, B10, and B15 were assigned to the testing set. Additionally, all frames from SyntheticColon III (O1–O3) were used exclusively as the testing set.

The C3VD dataset [7] introduces a novel multimodal 2D-3D registration method that aligns clinical video sequences with ground truth 3D models by transforming images into depth maps via a GAN and optimizing edge alignment using an evolutionary algorithm. The dataset includes 22 registered video sequences with paired ground truth depth maps. To assess the model's generalization capability, the dataset is partitioned into four anatomical subsets: Cecum, Descending Colon, Sigmoid Colon, and Transverse Colon.

The EndoSLAM dataset [16] includes both synthetic videos generated using the VRCaps simulation environment and real endoscopy videos from ex-vivo porcine GI organs. The synthetic dataset comprises 21,887 colon frames and 12,558 small bowel frames, each with ground truth depth maps and camera poses. These were used as test sets to evaluate model generalization across different GI regions. The real endoscopy videos in EndoSLAM dataset consist of only pose ground truth as it is difficult to use depth camera in endoscope. EndoSLAM and C3VD dataset have been used as the unseen test (not used in training) set only.

To pre-train StyleGAN-based generator, we use SceneNet RGB-D dataset [44], consisting of more than 15,000 frames of synthesized indoor RGB images and ground truth depth maps. In addition, we evaluate the performance of multiple methods on pose estimation. For the synthetic dataset [7], [43], each video frame is associated with ground-truth camera pose, enabling quantitative analysis. Furthermore, we assess the method on real in vivo surgical data from the EndoSLAM dataset [16], which includes multiple trajectories captured with three different endoscopic cameras: HighCam Colon-IV, LowCam Colon-IV, and MiroCam Colon-III.

B. Implementation details

Hyperparameters: For self-supervised depth and pose estimation, we implemented our method in PyTorch and trained on NVIDIA Quadro RTX 6000 with a batch size of 12, Adam as the optimizer with the weight decay of $1e^{-3}$, an initial learning rate of $1e^{-4}$, input resolution of 480×480 and training epoch of 20. To pre-train the StyleGAN-based generator, we used a progressive training approach with a fixed learning rate of 10^{-4}, gradually increasing the output resolution over the course of 100 epochs. To ensure the fairness of the experiments, all comparative models were trained using the same training strategies and the same training and testing datasets. Only SimCol dataset is used for training and held-out testing while other two datasets (C3VD and EndoSLAM) are used only for testing.

Evaluation metrics: We report eight commonly used metrics proposed in [17], [43] for evaluating the depth estimation accuracy, which are Abs Rel, Sq Rel, RMSE, RMSE log,

TABLE III

COMPARISON OF DEPTH ESTIMATION METHODS ACROSS VARIOUS METRICS ON THE C3VD AND ENDOSLAM DATASET (UNSEEN TEST).

Dataset	Method	Year	Depth Error (↓)					Depth Accuracy (↑)		
			Abs Rel	Sq Rel	RMSE	RMSE log	L_1 error	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
C3VD Cecum	Monodepth2 [3]	2019	0.282	3.506	10.315	0.323	8.186	0.577	0.864	0.947
	MonoViT [25]	2022	0.263	3.274	9.633	0.301	7.644	0.581	0.866	0.947
	DualRefine [24]	2023	0.262	3.260	9.592	0.300	7.612	0.585	0.869	0.949
	Lite-Mono [26]	2023	0.251	3.132	9.217	0.288	7.314	0.582	0.899	0.950
	Depth Pro† [12]	2024	0.288	3.588	10.559	0.330	8.379	0.562	0.860	0.941
	Endo-SFMLearner* [16]	2021	0.245	3.051	8.979	0.281	7.125	0.590	0.889	0.956
	AF-SFMLearner* [30]	2023	0.233	2.901	8.536	0.267	6.774	0.601	0.898	0.959
	IID-SfmLearner* [32]	2024	0.239	2.947	8.641	0.272	6.991	0.615	0.879	0.956
	EndoDAC* [34]	2024	0.230	2.900	8.502	0.266	6.717	0.618	0.899	0.961
	Ours	2025	**0.227**	**2.824**	**8.309**	**0.260**	**6.594**	**0.625**	**0.907**	**0.967**
C3VD Descending Colon	Monodepth2 [3]	2019	0.256	1.800	5.442	0.285	4.470	0.558	0.861	0.953
	MonoViT [25]	2022	0.244	1.718	5.197	0.272	4.268	0.568	0.870	0.957
	DualRefine [24]	2023	0.236	1.659	5.018	0.262	4.121	0.591	0.879	0.969
	Lite-Mono [26]	2023	0.232	1.637	4.951	0.259	4.066	0.633	0.899	0.975
	Depth Pro† [12]	2024	0.276	1.945	5.882	0.308	4.831	0.551	0.860	0.951
	Endo-SFMLearner* [16]	2021	0.233	1.640	4.959	0.259	4.073	0.631	0.902	0.975
	AF-SFMLearner* [30]	2023	0.229	1.613	4.877	0.255	4.005	0.662	0.911	0.974
	IID-SfmLearner* [32]	2024	0.231	1.629	4.922	0.259	4.051	0.658	0.909	0.966
	EndoDAC* [34]	2024	0.229	1.607	4.871	0.255	4.001	0.651	0.911	0.969
	Ours	2025	**0.223**	**1.569**	**4.746**	**0.248**	**3.898**	**0.653**	**0.922**	**0.979**
C3VD Sigmoid Colon	Monodepth2 [3]	2019	0.281	2.210	6.752	0.320	5.113	0.587	0.859	0.946
	MonoViT [25]	2022	0.267	2.103	6.426	0.304	4.866	0.609	0.861	0.946
	DualRefine [24]	2023	0.263	2.074	6.337	0.300	4.798	0.607	0.877	0.951
	Lite-Mono [26]	2023	0.250	1.967	6.011	0.284	4.551	0.616	0.883	0.955
	Depth Pro† [12]	2024	0.291	2.289	6.993	0.331	5.295	0.577	0.850	0.937
	Endo-SFMLearner* [16]	2021	0.251	1.979	6.048	0.286	4.579	0.611	0.886	0.957
	AF-SFMLearner* [30]	2023	0.245	1.928	5.892	0.279	4.461	0.653	0.901	0.955
	IID-SfmLearner* [32]	2024	0.249	1.957	5.905	0.280	4.511	0.639	0.901	0.949
	EndoDAC* [34]	2024	0.241	1.908	5.814	0.277	4.427	0.659	0.902	0.959
	Ours	2025	**0.233**	**1.842**	**5.629**	**0.266**	**4.262**	**0.672**	**0.904**	**0.973**
C3VD Transcending Colon	Monodepth2 [3]	2019	0.328	2.826	8.731	0.367	6.474	0.527	0.804	0.935
	MonoViT [25]	2022	0.313	2.698	8.337	0.350	6.181	0.551	0.829	0.950
	DualRefine [24]	2023	0.311	2.684	8.292	0.348	6.148	0.556	0.831	0.955
	Lite-Mono [26]	2023	0.298	2.564	7.921	0.332	5.873	0.579	0.860	0.959
	Depth Pro† [12]	2024	0.338	2.911	8.995	0.378	6.669	0.521	0.811	0.922
	Endo-SFMLearner* [16]	2021	0.299	2.568	7.936	0.333	5.884	0.601	0.839	0.952
	AF-SFMLearner* [30]	2023	0.291	2.514	7.767	0.326	5.759	0.617	0.875	0.962
	IID-SfmLearner* [32]	2024	0.295	2.551	7.899	0.331	5.831	0.609	0.846	0.957
	EndoDAC* [34]	2024	0.287	2.469	7.390	0.319	5.605	0.637	0.879	0.963
	Ours	2025	**0.261**	**2.248**	**6.946**	**0.291**	**5.150**	**0.669**	**0.882**	**0.970**
EndoSLAM Colon	Monodepth2 [3]	2019	0.447	0.672	0.701	0.449	0.459	0.411	0.725	0.860
	MonoViT [25]	2022	0.385	0.633	0.682	0.413	0.328	0.579	0.858	0.929
	DualRefine [24]	2023	0.423	0.562	0.643	0.439	0.404	0.479	0.749	0.865
	Lite-Mono [26]	2023	0.404	0.559	0.596	0.426	0.389	0.421	0.796	0.887
	Depth Pro† [12]	2024	0.492	0.713	0.779	0.461	0.477	0.403	0.719	0.833
	Endo-SFMLearner* [16]	2021	0.391	0.607	0.571	0.409	0.299	0.522	0.832	0.901
	AF-SfMLearner* [30]	2023	0.370	0.532	0.552	0.389	0.260	0.559	0.846	0.935
	IID-SfmLearner* [32]	2024	0.368	0.530	0.563	0.396	0.266	0.557	0.843	0.927
	EndoDAC* [34]	2024	0.362	0.529	0.550	0.385	0.257	0.570	0.854	0.939
	Ours	2025	**0.348**	**0.511**	**0.533**	**0.380**	**0.235**	**0.573**	**0.869**	**0.956**
EndoSLAM Small Intestine	Monodepth2 [3]	2019	0.444	0.602	0.653	0.453	0.417	0.410	0.701	0.866
	MonoViT [25]	2022	0.423	0.569	0.635	0.441	0.406	0.417	0.722	0.879
	DualRefine [24]	2023	0.422	0.536	0.617	0.442	0.403	0.418	0.704	0.875
	Lite-Mono [26]	2023	0.369	0.519	0.533	0.393	0.324	0.501	0.810	0.913
	Depth Pro† [12]	2024	0.471	0.633	0.679	0.459	0.430	0.412	0.689	0.856
	Endo-SFMLearner* [16]	2021	0.377	0.582	0.547	0.399	0.353	0.531	0.789	0.892
	AF-SfMLearner* [30]	2023	0.352	0.521	0.520	0.385	0.267	0.611	0.819	0.952
	IID-SfmLearner* [32]	2024	0.359	0.537	0.525	0.390	0.319	0.602	0.799	0.898
	EndoDAC* [34]	2024	0.347	0.520	0.515	0.381	0.248	0.632	0.833	0.959
	Ours	2025	**0.329**	**0.504**	**0.501**	**0.371**	**0.201**	**0.643**	**0.865**	**0.965**

† Foundation model trained on large-scale natural images; * Endoscopy-specific self-supervised method

Note: Depth error for EndoSLAM dataset are reported in centimeters (cm), while those for C3VD are in millimeters (mm).

$\delta < 1.25, \delta < 1.25^2, \delta < 1.25^3$ and L_1 error. Absolute Trajectory Error (ATE) evaluates the global consistency of a predicted trajectory with respect to the ground truth. Translational Relative Pose Error (Trans. RPE) and Rotational Relative Pose Error (Rot. RPE) quantifies the local accuracy of pose estimation by measuring the difference in translational motion and the angular difference between two consecutive poses over a fixed time interval. We follow the established practice in works such as SimCol [43] and EndoSLAM [16], which utilize a five-frame interval for Trans. RPE computation. We use these three metrics for evaluating the pose estimation. Similar to other monocular depth estimation approaches [3], [16], [25], [26], [30], which estimates trajectory up to an unknown scale factor, we perform scale-aware alignment before calculating the Absolute Trajectory Error (ATE). Specifically, the estimated trajectory is aligned to the ground truth via a single optimal similarity transformation that minimizes the least-squares error.

TABLE IV
COMPARISON OF POSE ESTIMATION METHODS ACROSS VARIOUS METRICS ON C3VD DATASET (SYNTHETIC, UNSEEN).

Dataset	Method	Pose Error (↓) ATE (in mm)	Trans. RPE (in mm)	Rot. RPE (deg)
C3VD Cecum	Monodepth2 [3]	3.5397	0.2583	0.4665
	DualRefine [24]	3.1318	0.2344	0.3876
	MonoViT [25]	3.4566	0.2394	0.4018
	Lite-Mono [26]	2.8878	0.2362	0.3571
	DROID-SLAM [15]	2.4961	0.2297	0.2448
	Endo-SFMLearner* [16]	2.5513	0.2588	0.3453
	AF-SFMLearner* [30]	2.4697	0.2334	0.3066
	IID-SfmLearner* [32]	2.5177	0.2508	0.3161
	EndoDAC* [34]	2.4719	0.2301	0.2895
	Ours	**2.2331**	**0.2125**	**0.1579**
C3VD Descending Colon	Monodepth2 [3]	11.6014	0.4926	0.4322
	DualRefine [24]	9.9003	0.4178	0.3652
	MonoViT [25]	10.3427	0.4226	0.3869
	Lite-Mono [26]	9.4211	0.3734	0.3007
	DROID-SLAM [15]	9.3594	0.3571	0.2182
	Endo-SFMLearner* [16]	9.4610	0.3699	0.2477
	AF-SFMLearner* [30]	9.3321	0.3608	0.2493
	IID-SfmLearner* [32]	9.4427	0.3653	0.2501
	EndoDAC* [34]	9.3075	0.3601	0.2375
	Ours	**9.0766**	**0.3413**	**0.1535**
C3VD Sigmoid Colon	Monodepth2 [3]	2.1315	0.2758	0.2581
	DualRefine [24]	2.1094	0.2440	0.2066
	MonoViT [25]	2.1115	0.2697	0.2246
	Lite-Mono [26]	1.9833	0.2125	0.1791
	DROID-SLAM [15]	1.7317	0.1881	0.1699
	Endo-SFMLearner* [16]	1.8297	0.2176	**0.1470**
	AF-SFMLearner* [30]	1.7345	0.1844	0.1853
	IID-SfmLearner* [32]	1.7941	0.1996	0.1728
	EndoDAC* [34]	1.7573	0.1731	0.1411
	Ours	**1.6535**	**0.1408**	0.1706
C3VD Transcending Colon	Monodepth2 [3]	3.8970	0.2271	0.1510
	DualRefine [24]	3.3377	0.1892	0.1273
	MonoViT [25]	3.5271	0.1997	0.1462
	Lite-Mono [26]	3.0016	0.1808	0.1116
	DROID-SLAM [15]	2.4977	0.1602	0.1041
	Endo-SFMLearner* [16]	2.9108	0.1834	0.1079
	AF-SFMLearner* [30]	2.5011	**0.1578**	0.1053
	IID-SfmLearner* [32]	2.6785	0.1619	0.1077
	EndoDAC* [34]	2.5548	0.1592	0.1062
	Ours	**2.1556**	0.1598	**0.0901**

TABLE V
COMPARISON OF POSE ESTIMATION METHODS ACROSS VARIOUS METRICS ON ENDOSLAM DATASET (REAL EX-VIVO, UNSEEN).

Dataset	Method	Pose Error (↓) ATE (in m)	Trans. RPE (in m)	Rot. RPE (deg)
EndoSLAM HighCam Colon-IV	Monodepth2 [3]	0.1231	0.0038	0.9298
	DualRefine [24]	0.1192	0.0039	0.9064
	MonoViT [25]	0.1155	0.0033	0.8349
	Lite-Mono [26]	0.0924	0.0030	0.6973
	DROID-SLAM [15]	0.0833	0.003	0.5941
	Endo-SFMLearner* [16]	0.1055	0.0032	0.7025
	AF-SFMLearner* [30]	0.0861	0.0024	0.6278
	IID-SfmLearner* [32]	0.0997	0.0033	0.6995
	EndoDAC* [34]	0.0872	0.0029	0.6211
	Ours	**0.0654**	**0.0021**	**0.4713**
EndoSLAM LowCam Colon-IV	Monodepth2 [3]	0.1355	0.0024	1.4661
	DualRefine [24]	0.1121	0.0021	1.1792
	MonoViT [25]	0.1319	0.0022	1.1297
	Lite-Mono [26]	0.1120	0.0017	1.0524
	DROID-SLAM [15]	0.0939	0.0016	0.8133
	Endo-SFMLearner* [16]	0.1027	0.0017	1.0079
	AF-SFMLearner* [30]	0.0974	0.0015	0.8948
	IID-SfmLearner* [32]	0.1038	0.0018	0.9465
	EndoDAC* [34]	0.0946	0.0015	0.8277
	Ours	**0.0881**	**0.0011**	**0.6110**
EndoSLAM MiroCam Colon-III	Monodepth2 [3]	0.2515	0.0053	0.7775
	DualRefine [24]	0.2502	0.0051	0.7164
	MonoViT [25]	0.2419	0.0049	0.7670
	Lite-Mono [26]	0.2448	0.0050	0.6636
	DROID-SLAM [15]	0.2197	0.0043	0.5901
	Endo-SFMLearner* [16]	0.2306	0.0042	0.5908
	AF-SFMLearner* [30]	0.2411	0.0049	0.6265
	IID-SfmLearner* [32]	0.2471	0.0048	0.6065
	EndoDAC* [34]	0.2299	0.0044	0.5971
	Ours	**0.1963**	**0.0039**	**0.5739**

C. Results

1) Results in the SimCol dataset

Methods included for comparison are Monodepth2 [3], DualRefine [24], MonoVit [25], Lite-Mono [26], foundation model Depth Pro [12], endoscopy-specific model Endo-SFMLearner [16] and AF-SFMLearner [30]. The depth estimation results on the SimCol dataset are presented in Table I. Our method outperforms Monodepth2 across metrics and demonstrates superior performance compared to other recent methods. Specifically, on the Synthetic Colon III test set, our method achieves an RMSE of 0.501, an Abs Rel of 0.110, and an L_1 error of 0.320—reducing RMSE by 0.123 and 0.038 compared to Monodepth2 and Lite-Mono, respectively. Compared to the endoscopy-specific state-of-the-art model AF-SFMLearner, our method achieves improvements of 0.036 in RMSE and 0.041 in L_1 error. Similarly, our method surpasses other methods on the Synthetic Colon I/II test set. Depth Pro was also evaluated based on transfer learning (i.e., the natural scene depth pretrained foundation model is directly used to perform inference on endoscopy datasets), but the results on both test sets were unsatisfactory. As shown in Table II, our method demonstrates solid improvement over recent methods in pose estimation, achieving ATE values of 0.4209 and 0.4246, Trans. RPE values of 0.0833 and 0.0736 and Rot. RPE values of 0.1077 and 0.2273 on the Simcol-I/II and Simcol-III test sets, respectively.

2) Results in the C3VD dataset

The four subsets of C3VD are used to evaluate the generalization ability of the proposed method, as summarized in Table III. Compared to the baseline Monodepth2, our method achieves over 15% improvement in RMSE across all subsets, along with substantial gains in all other evaluation metrics. Furthermore, our approach consistently outperforms the state-of-the-art AF-SFMLearner, demonstrating its superior effectiveness on endoscopic data.

As shown in Table IV, our proposed method achieves superior performance in both ATE and RPE on the C3VD Cecum and Descending Colon subsets, consistently outperforming the compared methods. However, Endo-SFMLearner and AF-SFMLearner achieve lower rotational and translational RPE on the Sigmoid Colon and Transverse Colon subsets, respectively. Nevertheless, our method maintains the lowest ATE across all subsets, indicating better overall global pose consistency.

3) Results in the EndoSLAM dataset

The proposed method is evaluated on the EndoSLAM dataset to show its generalization ability. Table III shows the results of compared method on EndoSLAM colon and small intestine datasets. Our proposed method achieves lower RMSE values of 0.533 and 0.501 in colon and small intestine datasets, a relative improvement of 0.168/0.063 and 0.152/0.032 over Monodepth2 and Lite-Mono, respectively. To assess real-world applicability, we validate the compared methods on a real porcine colon dataset that closely mimics human anatomy, captured using three distinct endoscopic cameras: HighCam, LowCam, and MiroCam. In Table V, we can see that our

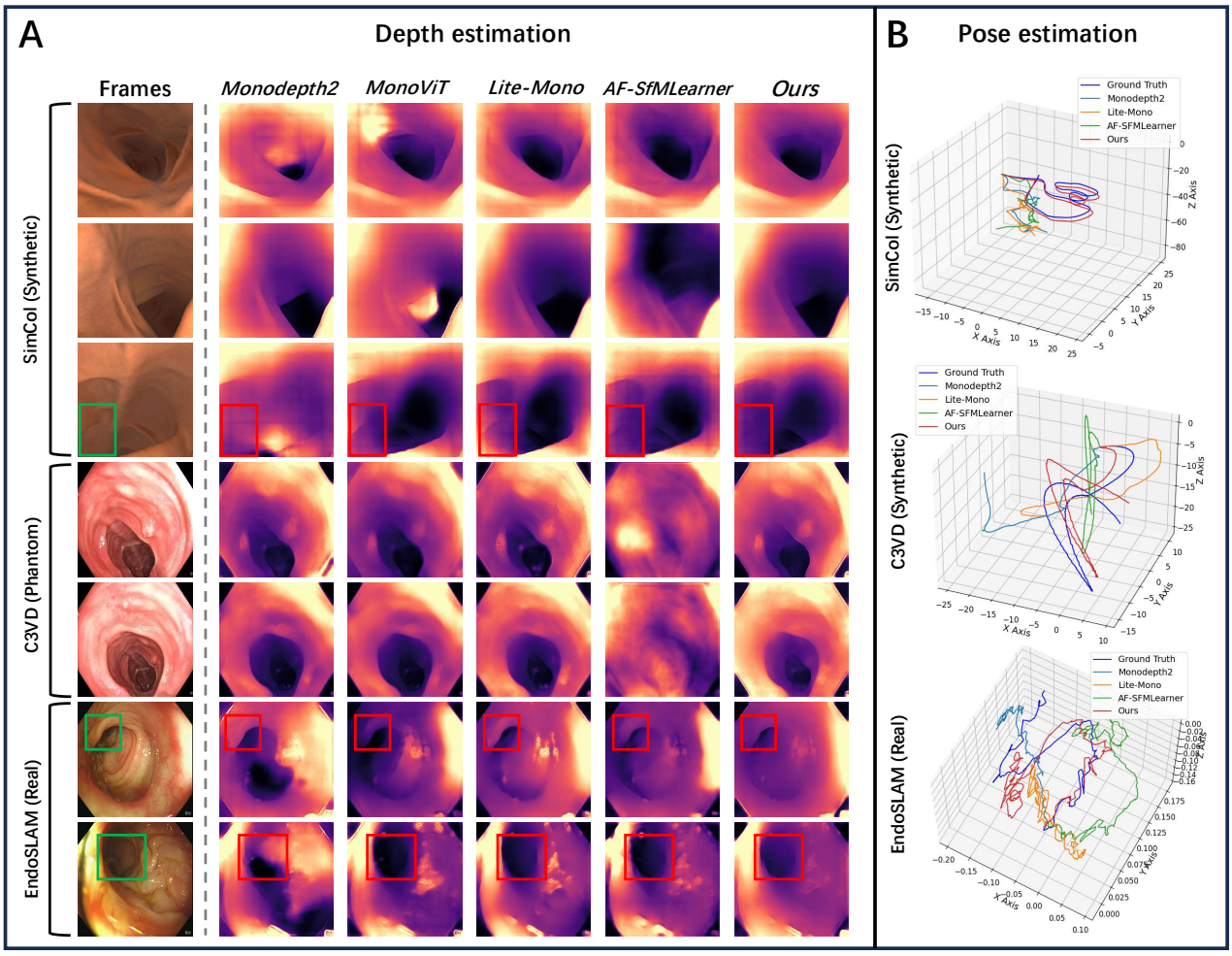

Fig. 4. Qualitative results of depth and pose estimation. The left part in Fig. (A) shows depth maps generated by Monodepth2 [3], MonoVit [25], Lite-Mono [26], AF-SFMLearner [30] and Ours. Our method demonstrates superior performance, particularly on challenging phantom and real frames, where complex textures and lighting variations pose significant challenges for depth and pose estimation. The right part in Fig. (B) shows qualitative results of pose estimation. Our method outperforms existing approaches by achieving more consistent scale across x-, y-, z-axis and better alignment of rotation angles, especially in sequences involving complex camera motion and anatomical deformation.

method maintains superior performance even in non-virtual datasets.

4) Qualitative results for depth and pose estimation

Fig. 4(A) (top three rows) presents a qualitative comparison on the synthetic SimCol dataset. It can be observed that our method not only ensures more accurate depth estimation overall but also maintains accurate depth in challenging areas such as colonic folds. A representative pose trajectory visualization is shown in Fig. 4 (B-top), where our method produces a trajectory better aligned with the ground truth compared to other methods. In contrast, we observe that many existing methods exaggerate the z-axis scale compared to the x- and y- axes, resulting in near-linear trajectories that misalign with the colon's natural curvature. Colonoscopic movement is challenging to model due to pronounced z-axis progression with gradual curvature at key anatomical points. Our approach addresses this by treating pose outputs as latent variables in a VAE-like framework, regularizing pose transition scales across all axes. This mitigates z-axis dominance, improving sensitivity to x and y axis changes and preserving the natural curvature and spatial alignment of the endoscopic trajectory. As shown in rows four and five of Fig. 4(A) on the Phantom C3VD dataset, our method exhibits superior robustness in handling reflective areas. The visualization in Fig. 4 (B-middle) shows pose estimation results on the C3VD dataset. It can be observed that our method aligns more closely with the ground truth in both trajectory shape and spatial positioning.

Given the absence of public real-world depth datasets for colonoscopy due to inherent clinical constraints, we address this by presenting qualitative results on real data EndoSLAM (real) in Fig. 4(A), last row. Our results demonstrate that our method maintains superior performance even on unseen real endoscopic images compared to other approaches. Fig. 4 (B-bottom) shows the pose estimation results on a human colonoscopy clip, where our method not only predicts the correct orientation, but also produces a more accurate trajectory. Our method's strength lies in incorporating depth-specific priors via a pretrained generative latent generator and VAE-based pose regularization, enhancing robustness in complex endoscopic scenes. This conditioning enables accurate estima-

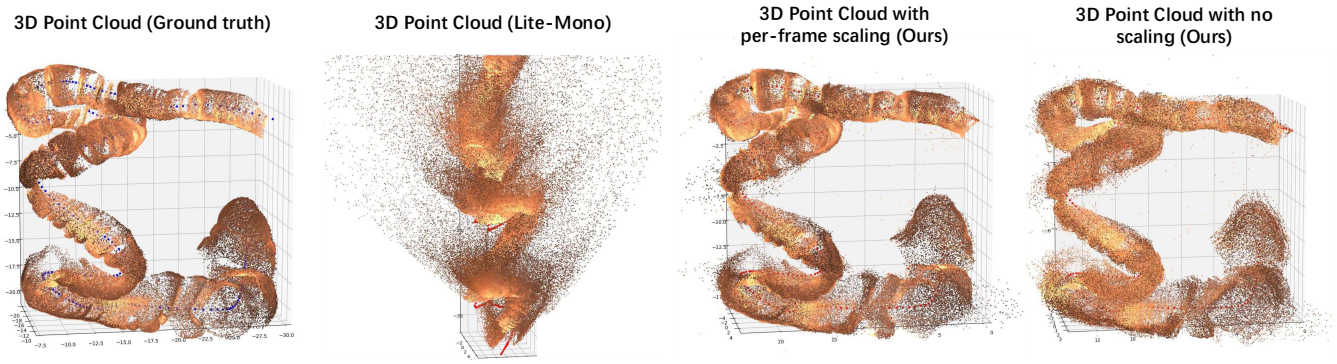

Fig. 5. 3D point cloud reconstruction by back-projecting the predicted depth maps using the corresponding camera intrinsics and estimated poses for SimCol dataset. The first column shows the 3D point cloud reconstruction using ground truth pose and depth. The second column shows the 3D point cloud reconstruction from Lite-Mono without applyging depth scaling. The third column presents the 3D point cloud reconstruction results from our method after per-frame alignment with ground truth depth. Similarly, the fourth column presents our model's raw, unscaled output, reconstructed using only rotation-aligned camera poses and without any depth scaling. Despite the complete absence of ground truth scale supervision, our method recovers a globally consistent 3D structure highly similar to the ground truth, with only minor local deviations and artifacts.

TABLE VI
ABLATION STUDY ON PROPOSED COMPONENTS WITHIN OUR MODEL ARCHITECTURE.

Dataset	Architecture			Depth Error(↓)					Pose Error (↓)		
	StyleGAN generator	VAE	DROID-SLAM prior	Abs Rel	Sq Rel	RMSE	RMSE log	L_1 error	ATE	Trans. RPE	Rot. RPE
SimCol I & II	✓	✓	✓	**0.121**	**0.110**	**0.608**	**0.123**	**0.339**	**0.4209**	**0.0833**	**0.1077**
	✓	✓		0.128	0.111	0.622	0.128	0.358	0.4552	0.0901	0.1418
		✓	✓	0.142	0.138	0.636	0.134	0.369	0.4605	0.1009	0.1277
	✓			0.129	0.112	0.627	0.128	0.365	0.4671	0.0906	0.1583
		✓		0.149	0.147	0.646	0.138	0.372	0.4601	0.1012	0.1329
				0.151	0.167	0.660	0.142	0.381	0.4981	0.1071	0.1719
SimCol III	✓	✓	✓	**0.110**	**0.079**	**0.501**	**0.119**	**0.320**	**0.4246**	**0.0736**	**0.2273**
	✓	✓		0.114	0.109	0.520	0.133	0.327	0.4409	0.0851	0.2511
		✓	✓	0.119	0.148	0.547	0.151	0.347	0.4335	0.0897	0.2390
	✓			0.114	0.112	0.522	0.135	0.329	0.4453	0.0886	0.2539
		✓		0.123	0.129	0.550	0.144	0.364	0.4418	0.0925	0.2521
				0.126	0.130	0.562	0.149	0.375	0.5056	0.0962	0.2733

tion despite challenges like reflective surfaces, colonic folds, and variable lighting, allowing effective generalization across diverse environments without ground truth labels.

5) 3D reconstruction quality assessment in SimCol dataset

To further highlight the qualitative effectiveness of our approach, we reconstruct a 3D point cloud by back-projecting the predicted depth maps using camera intrinsics and predicted poses. As shown in Fig. 5, the second column visualizes the 3D point cloud reconstructed using Lite-Mono predictions. While the overall structure is recognizable, this reconstruction exhibits notable geometric inaccuracies — including drifted camera trajectories (evident from misaligned red trajectory lines). These artifacts stem primarily from accumulated pose estimation errors and imperfect depth predictions under challenging endoscopic conditions, highlighting the limitations of monocular methods without robust motion priors or dense supervision. The right columns (3rd and 4th), in contrast, presents the reconstruction by our method without any depth scaling — using only rotation-aligned poses to preserve the model's native output scale. Remarkably, even in this unscaled setting, our method still recovers a globally consistent structure highly similar to the ground truth, with only minor local deviations and artifacts. This confirms the robustness of our framework in preserving scene layout and relative depth relationships without relying on ground-truth scale supervision — a critical property for real-world deployment in self-supervised or metric-agnostic scenarios, where per-frame scaling is neither available nor desirable. Together, these visualizations showcase the framework's potential in recovering realistic 3D geometry under both ideal and practical conditions.

D. Ablation Studies

By selectively removing modules, we evaluated the resulting performance on the SimCol dataset, as shown in Table VI. In the architecture design, we have added StyleGAN generator, VAE and DROID-SLAM prior.

Impact of the StyleGAN generator Module: Removing the StyleGAN-based generator module led to a marked decrease in all depth estimation metrics, showing its critical role. Here, we removed only the StyleGAN-based generator module while keeping the encoder-decoder configuration retaining the skip connections pointed by red arrown in Fig. 2.

Impact of the VAE Module: The removal of the VAE module caused a significant drop in pose estimation accuracy, also highlighting its importance. Additionally, removing the VAE module decreases depth estimation metrics, and vice versa. This effect reflects the interdependence of the depth and pose branches in reprojection, where both branches contribute to self-supervised learning by warping adjacent frames to predict the current frame, underscoring the interconnected nature of our framework.

V. Conclusion and Future Directions

In this work we proposed a novel architecture for depth and pose estimation in endoscopy, leveraging generative latent priors for robust, self-supervised optimization of both tasks. Our approach achieved state-of-the-art performance on the SimCol, C3VD and EndoSLAM datasets, demonstrating strength in handling complex camera pose variations that challenge existing methods. Although our results are encouraging, a comprehensive quantitative validation in clinical setting remains constrained by the absence of ground truth data acquired during colonoscopy in patients. In future work, we plan to integrate real-world colonoscopy data for 3D reconstruction and potentially match it with the CT scans from the same patient to validate the 3D reconstruction. Additionally, we will aim to incorporate 3D reconstruction of the other hollow organs in the GI tract, which will further substantiate our approach and expand its applicability within clinical workflows.

References

[1] L. H. Biller and D. Schrag, "Diagnosis and treatment of metastatic colorectal cancer: a review," *Jama*, vol. 325, no. 7, pp. 669–685, 2021.

[2] M. Bretthauer, M. Løberg, P. Wieszczy, M. Kalager, L. Emilsson, K. Garborg, M. Rupinski, E. Dekker, M. Spaander, M. Bugajski *et al.*, "Effect of colonoscopy screening on risks of colorectal cancer and related death," *New England Journal of Medicine*, vol. 387, no. 17, pp. 1547–1556, 2022.

[3] 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*, 2019, pp. 3828–3838.

[4] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical image analysis*, vol. 48, pp. 230–243, 2018.

[5] R. Ma, R. Wang, S. Pizer, J. Rosenman, S. K. McGill, and J.-M. Frahm, "Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*. Springer, 2019, pp. 573–582.

[6] D. Yoon, H.-J. Kong, B. S. Kim, W. S. Cho, J. C. Lee, M. Cho, M. H. Lim, S. Y. Yang, S. H. Lim, J. Lee *et al.*, "Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network," *Scientific reports*, vol. 12, no. 1, p. 261, 2022.

[7] 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*, vol. 90, p. 102956, 2023.

[8] K. Cheng, Y. Ma, B. Sun, Y. Li, and X. Chen, "Depth estimation for colonoscopy images with self-supervised learning from videos," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer, 2021, pp. 119–128.

[9] Y. Liu and S. Zuo, "Self-supervised monocular depth estimation for gastrointestinal endoscopy," *Computer Methods and Programs in Biomedicine*, vol. 238, p. 107619, 2023.

[10] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[11] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[12] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2410.02073

[13] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1167–1176, 2019.

[14] P. E. C. Solano, A. Bulpitt, V. Subramanian, and S. Ali, "Multi-task learning with cross-task consistency for improved depth estimation in colonoscopy," *Medical Image Analysis*, vol. 99, p. 103379, 2025.

[15] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in neural information processing systems*, 2021.

[16] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical image analysis*, vol. 71, p. 102058, 2021.

[17] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[18] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[19] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[20] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3227–3237.

[21] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 225–234.

[22] V. R. Kumar, M. Klingner, S. Yogamani, S. Milz, T. Fingscheidt, and P. Mader, "Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 61–71.

[23] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2022–2030.

[24] A. Bangunharcana, A. Magd, and K.-S. Kim, "Dualrefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 726–738.

[25] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 international conference on 3D vision (3DV)*. IEEE, 2022, pp. 668–678.

[26] 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*, 2023, pp. 18 537–18 546.

[27] S. Matthias, M. Kästner, and E. Reithmeier, "A 3d measuring endoscope for hand-guided operation," *Measurement Science and Technology*, vol. 29, no. 9, p. 094001, 2018.

[28] F. Mahmood, Z. Yang, R. Chen, D. Borders, W. Xu, and N. J. Durr, "Polyp segmentation and classification using predicted depth from monocular endoscopy," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. SPIE, 2019, pp. 268–272.

[29] S.-J. Hwang, S.-J. Park, G.-M. Kim, and J.-H. Baek, "Unsupervised monocular depth estimation for colonoscope system using feedback network," *Sensors*, vol. 21, no. 8, p. 2691, 2021.

[30] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Medical image analysis*, vol. 77, p. 102338, 2022.

[31] Z. Yang, J. Pan, J. Dai, Z. Sun, and Y. Xiao, "Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1934–1944, 2024.

[32] B. Li, B. Liu, M. Zhu, X. Luo, and F. Zhou, "Image intrinsic-based unsupervised monocular depth estimation in endoscopy," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[33] J. Rodríguez-Puigvert, V. M. Batlle, J. M. M. Montiel, R. Martinez-Cantin, P. Fua, J. D. Tardós, and J. Civera, "Lightdepth: Single-view depth self-supervision from illumination decline," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 21 273–21 283.

[34] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 208–218.

[35] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 245–14 254.

[36] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[37] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[39] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." *ICLR (Poster)*, vol. 3, 2017.

[40] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[41] D. Dutta, C. Amballa, Z. Xu, Y.-L. Wei, and R. R. Choudhury, "Learning energy-based variational latent prior for vaes," *arXiv preprint arXiv:2510.00260*, 2025.

[42] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*. Springer, 2000, pp. 298–372.

[43] A. Rau, S. Bano, Y. Jin, P. Azagra, J. Morlana, R. Kader, E. Sanderson, B. J. Matuszewski, J. Y. Lee, D.-J. Lee *et al.*, "Simcol3d—3d reconstruction during colonoscopy challenge," *Medical Image Analysis*, vol. 96, p. 103195, 2024.

[44] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth," *arXiv preprint arXiv:1612.05079*, 2016.