# ColonCrafter: A Depth Estimation Model for Colonoscopy Videos Using Diffusion Priors

Romain Hardy[1], Tyler M. Berzin MD[2], Pranav Rajpurkar PhD[1]
*1. Department of Biomedical Informatics, Harvard Medical School*
*2. Center for Advanced Endoscopy, Beth Israel Deaconess Medical Center*

Three-dimensional (3D) scene understanding in colonoscopy presents significant challenges that necessitate automated methods for accurate depth estimation. However, existing depth estimation models for endoscopy struggle with temporal consistency across video sequences, limiting their applicability for 3D reconstruction. We present ColonCrafter, a diffusion-based depth estimation model that generates temporally consistent depth maps from monocular colonoscopy videos. Our approach learns robust geometric priors from synthetic colonoscopy sequences, enabling reliable depth estimation across frames. We also introduce a style transfer technique that preserves geometric structure while adapting realistic clinical videos to match our synthetic training domain. ColonCrafter achieves state-of-the-art zero-shot performance on the C3VD dataset, outperforming both general-purpose and endoscopy-specific approaches. Although full trajectory 3D reconstruction remains a challenge, we demonstrate clinically relevant applications of ColonCrafter, including 3D point cloud generation and surface coverage assessment. Our code will be made publicly available at `https://github.com/rajpurkarlab/ColonCrafter`.

*Keywords*: colonoscopy, depth estimation, 3D reconstruction, diffusion models

## 1. Introduction

Colorectal cancer (CRC) remains a leading cause of cancer-related mortality, with 52,900 projected deaths in 2025 in the United States alone.[1] Colonoscopy serves as the gold standard for CRC screening and is designated as a first-tier test by the American College of Gastroenterology.[2] During this procedure, gastroenterologists navigate a flexible endoscope through the colon to identify and remove precancerous polyps and other lesions. However, the clinical effectiveness of colonoscopy is constrained by fundamental limitations in human visual perception and spatial reasoning within the complex three-dimensional colonic environment.

These limitations manifest in several critical ways that directly impact patient outcomes. Incomplete examinations occur due to poor visualization behind haustral folds, leading to miss rates of up to 26% for adenomas.[3–5] Clinicians struggle to relocate previously identified lesions during the same procedure or across multiple sessions, complicating treatment planning and follow-up care.[6] Perhaps most importantly, accurate measurement of polyp size—a critical factor in determining removal strategy and surveillance intervals—remains challenging using current two-dimensional visualization methods.[7,8] These challenges underscore a

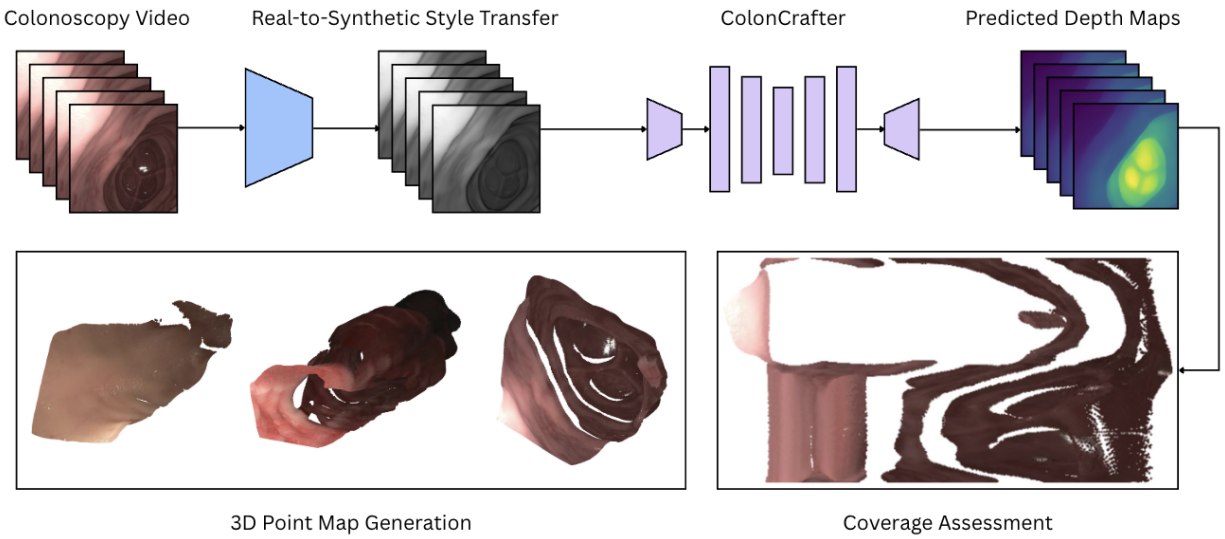


Fig. 1. Our approach incorporates two key components: (1) ColonCrafter, a diffusion-based depth estimation model trained on large-scale synthetic colonoscopy sequences, and (2) a domain adaptation technique that maps real colonoscopy frames into the synthetic training domain of ColonCrafter while preserving geometric structure. ColonCrafter outputs accurate, temporally consistent depth maps suitable for downstream 3D reconstruction and surface coverage assessment.

fundamental mismatch between the inherently three-dimensional nature of the colon and the two-dimensional visual information available to clinicians.

Bridging this gap between human expertise and the geometric complexity of colonoscopy requires computational tools that can augment clinical decision-making through precise three-dimensional scene understanding. 3D reconstruction of the colon could transform clinical practice by enabling complete surface coverage assessment,[9,10] accurate lesion localization and size measurement,[11] and robust lesion registration across multiple viewpoints and examination sessions. Such capabilities would complement rather than replace human expertise, providing clinicians with enhanced spatial awareness while preserving their critical role in clinical interpretation and decision-making.

However, achieving reliable 3D reconstruction from colonoscopy videos presents formidable technical challenges that render conventional computer vision approaches ineffective. The colonic environment systematically violates fundamental assumptions underlying traditional Simultaneous Localization and Mapping (SLAM) algorithms.[12,13] The mucosa lacks distinctive visual features necessary for robust tracking,[12,14] exhibits non-Lambertian reflectance with severe specular highlights,[6,15] and undergoes continuous deformation due to peristalsis and insufflation.[16,17] Additionally, rapidly changing illumination from the endoscope's point light source and erratic motion patterns—including rapid rotations, forward-backward movements, and frequent occlusions—further complicate reconstruction efforts.[13,18]

Recent advances in deep learning have enabled progress toward colonoscopy-specific depth estimation and SLAM systems.[16,19–21] Self-supervised approaches have shown promise by learning from unlabeled colonoscopy videos,[19,20,22] while others have leveraged synthetic data (i.e.,

computer-simulated videos with paired depth maps) to overcome the scarcity of ground-truth annotations.[23–25] Nevertheless, existing methods still struggle with long-term temporal consistency and often fail to generalize across the diverse appearance variations present in clinical data.[12,26,27] The persistent domain gap between synthetic training data and real colonoscopy imagery represents a critical barrier to clinical translation,[23,28,29] with models trained on synthetic data typically exhibiting poor performance when deployed on real clinical videos.[13,30]

To address these challenges and enable clinically meaningful AI-assisted colonoscopy, we present ColonCrafter (Figure 1), a diffusion-based depth estimation framework designed to generate temporally consistent depth maps from monocular colonoscopy videos. Our approach addresses the fundamental limitations of existing methods through three key innovations. First, we formulate monocular depth estimation (MDE) as a conditional generation task within a diffusion framework, enabling the model to learn robust priors over the complex appearance and geometry of colonic scenes. Second, we train our model on a large-scale dataset of synthetic colonoscopy sequences derived from computed tomography (CT) scans, providing rich supervision for reliable video-level reconstructions. Third, we introduce a style transfer technique that adapts real colonoscopy videos to match the appearance of our synthetic training domain while preserving the geometric structure essential for accurate depth estimation.

Our main contributions advance the integration of AI and clinical expertise in colonoscopy:

- We present the first diffusion-based depth estimation framework specifically designed for colonoscopy, capable of generating temporally consistent dense depth maps that enable robust 3D scene understanding.
- We develop a novel style transfer technique that bridges the domain gap between synthetic training data and real colonoscopy videos while preserving geometric cues essential for clinical applications.
- We demonstrate state-of-the-art zero-shot performance on the C3VD[9] benchmark, showing significant improvements in depth estimation accuracy compared to existing methods.

## 2. Methods

We introduce ColonCrafter, a diffusion-based depth estimation model tailored to colonoscopy. Given a monocular colonoscopy video sequence, ColonCrafter estimates temporally consistent dense depth maps, enabling 3D reconstruction of the colon surface. Our approach is based on a video diffusion model trained on a large-scale dataset of synthetic colonoscopy videos derived from CT scans. To overcome the domain gap between our synthetic training data and real-world clinical videos, we further propose a training-free style injection technique that converts any colonoscopy video to the style of our synthetic videos.

### 2.1. *Depth Estimation*

ColonCrafter is a conditional video diffusion model adapted from the DepthCrafter[31] architecture. Given an input RGB video $\mathbf{x} \in \mathbb{R}^{F\times H\times W\times 3}$, the primary objective of ColonCrafter is to predict a temporally consistent depth sequence $\mathbf{d} \in \mathbb{R}^{F\times H\times W}$. An overview of ColonCrafter is shown in Figure 2.

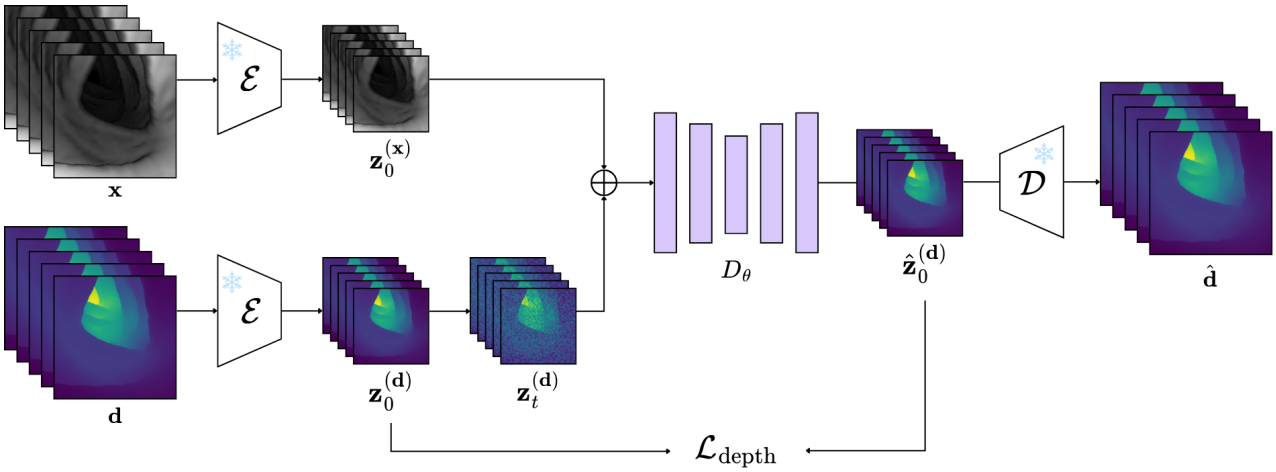


Fig. 2. Overview of the ColonCrafter architecture. The model uses a conditional diffusion framework in which paired colonoscopy videos and depth maps are projected into a latent space, and a spatio-temporal U-Net denoiser learns to recover clean depth latents from their noisy counterparts. Training is performed on synthetic colonoscopy sequences derived from CT scans to learn robust geometric priors.

As with DepthCrafter, we formulate this as a conditional generation task within the Elucidating Diffusion Models[32] framework. During training, the ground-truth depth sequence $\mathbf{d}$ is encoded to a lower-dimensional latent representation $\mathbf{z}_0^{(\mathbf{d})} = \mathcal{E}(\mathbf{d})$ using the encoder $\mathcal{E}$ of a pre-trained variational autoencoder (VAE). This latent code is then subjected to a forward diffusion process, which progressively adds noise over a continuous time variable $t$: $\mathbf{z}_t^{(\mathbf{d})} = \mathbf{z}_0^{(\mathbf{d})} + \sigma_t^2 \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is a Gaussian noise sample and $\sigma_t^2$ is the noise variance at time $t$. Finally, a spatio-temporal U-Net denoiser $D_\theta$ tries to predict the clean depth latent from the noisy latent: $\hat{\mathbf{z}}_0^{(\mathbf{d})} = D_\theta(\mathbf{z}_t^{(\mathbf{d})}; \sigma_t; \mathbf{z}_0^{(\mathbf{x})})$. Here, $\mathbf{z}_0^{(\mathbf{x})} = \mathcal{E}(\mathbf{x})$ is the encoded latent of the input video $\mathbf{x}$, and serves to condition the denoising process. The denoising objective is given by

$$\mathcal{L}_{\text{depth}} = \lambda_t \| D_\theta(\mathbf{z}_t^{(\mathbf{d})}; \sigma_t; \mathbf{z}_0^{(\mathbf{x})}) - \mathbf{z}_0^{(\mathbf{d})} \|_2^2 = \lambda_t \| \hat{\mathbf{z}}_0^{(\mathbf{d})} - \mathbf{z}_0^{(\mathbf{d})} \|_2^2, \tag{1}$$

where $\lambda_t$ is the loss weight at time $t$. To obtain the predicted depth map $\hat{\mathbf{d}}$, we project $\hat{\mathbf{z}}_0^{(\mathbf{d})}$ back to pixel space using the VAE's frozen decoder $\mathcal{D}$, i.e. $\hat{\mathbf{d}} = \mathcal{D}(\hat{\mathbf{z}}_0^{(\mathbf{d})})$.

## 2.2. *Training Details*

### 2.2.1. *Synthetic Dataset Construction*

ColonCrafter is trained on 109,329 colonoscopy images that we synthetically generated from 5 CT scans[33] using standard segmentation and virtual fly-through rendering software.[34,35] An overview of our dataset construction methodology is shown in Figure 3.

### 2.2.2. *Model Fine-Tuning*

Rather than training ColonCrafter from scratch, we fine-tune the publicly available checkpoint of DepthCrafter using Low-Rank Adaptation (LoRA).[36] This approach allows us to reuse the high-level features learned during its pre-training phase, while simultaneously reducing the

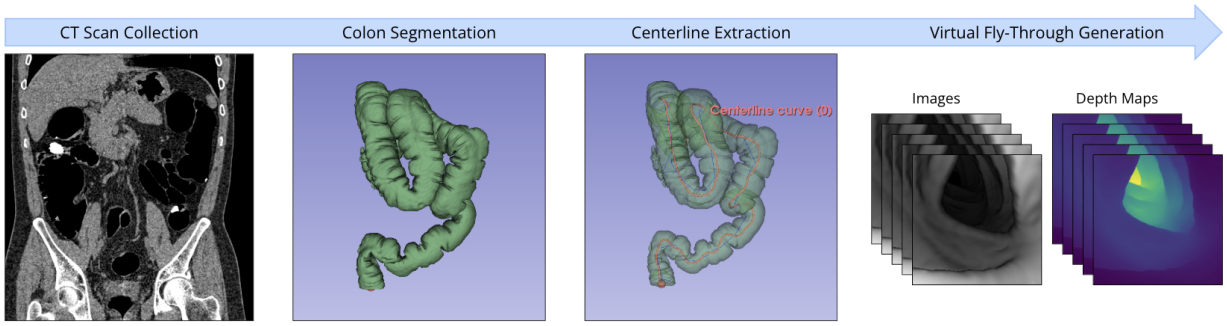


Fig. 3. Synthetic dataset construction pipeline. We segment the colonic volume from 5 CT scans, extract centerline paths, and render virtual fly-throughs to generate synthetic colonoscopy sequences with paired ground-truth depth maps.

computational cost of adapting the model to colonoscopy videos. In our implementation, we set the LoRA rank to 16 to give ColonCrafter sufficient expressive power and target the attention modules of the U-Net denoiser.

We fine-tune ColonCrafter on an NVIDIA A100 GPU for 50,000 steps using the AdamW[37] optimizer with a learning rate of $1.0 \times 10^{-5}$ on a cosine schedule with warm-up. We set the batch size to 8 and the sequence length to 16 (similar to DepthCrafter), allowing ColonCrafter to effectively exploit temporal information. To improve generalization to complex colonoscopy trajectories, we introduce several data augmentation strategies. First, we randomly sample sequences so that camera translations between successive frames are not constant. Second, we randomly flip video segments to reflect the fact that colonoscopy trajectories are rarely straight paths in practice. Third, we randomly jitter the camera intrinsics to simulate variations in endoscopic equipment. Finally, we apply a random attenuation factor to vary input video brightness, creating lighting conditions representative of real colonoscopy procedures.

### 2.3. *Real-to-Synthetic Style Transfer*

Since ColonCrafter is trained on synthetic videos with different appearances from real colonoscopy videos, we propose a style transfer approach to bridge this gap. Previous works train dedicated neural networks using cycle consistency losses to convert between image domains.[38–40] However, such models must be trained on large corpora containing sufficiently varied synthetic and real colonoscopy examples, and are prone to destroying structural and lighting cues that are crucial for depth estimation and SLAM. To address these issues, we propose a dynamic style transfer approach that enforces content preservation when shifting from the real domain to the synthetic domain, as shown in Figure 4.

Specifically, we cast real-to-synthetic style transfer as a modulation of the denoising process within a pre-trained Stable Diffusion (SD) model, inspired by the work of Chung et al.[41] on artistic style transfer. Let $x^{(\mathbf{c})}$ be a colonoscopy video frame that we want to convert to the style of a synthetic video frame $x^{(\mathbf{s})}$. First, we project $(x^{(\mathbf{c})}, x^{(\mathbf{s})})$ to latent codes $(z_0^{(\mathbf{c})}, z_0^{(\mathbf{s})})$ using the VAE encoder of the SD model. Second, we invert the clean latents $(z_0^{(\mathbf{c})}, z_0^{(\mathbf{s})})$ to noisy latents $(z_T^{(\mathbf{c})}, z_T^{(\mathbf{s})})$ over $T$ time steps. At each step $t \in [1, T]$, we store the intermediate queries $Q_t^{(\mathbf{c})}$ of

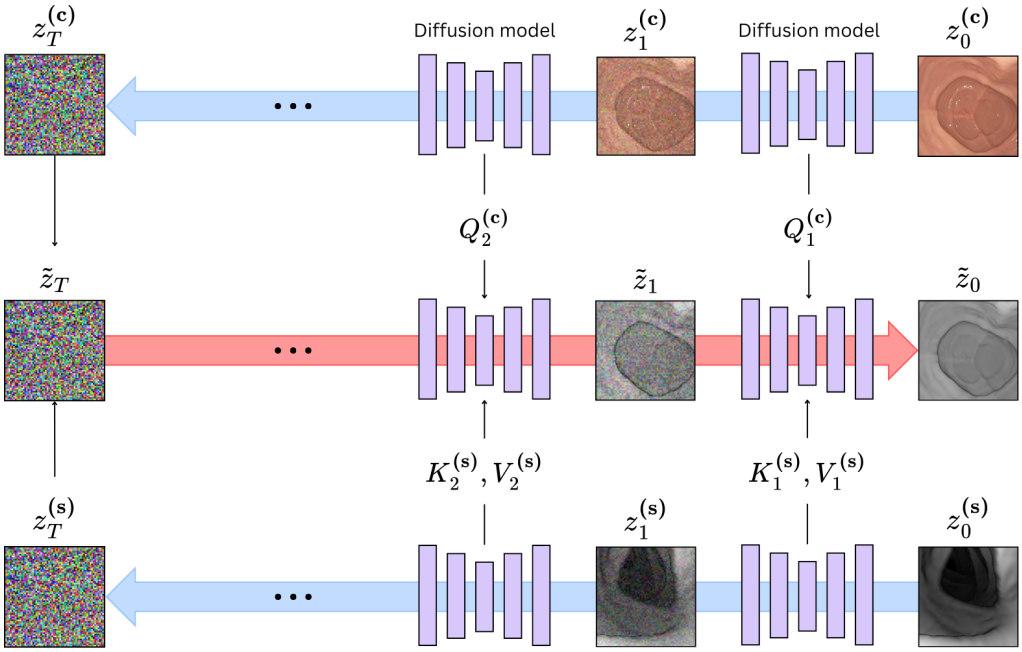


Fig. 4. Real-to-synthetic style transfer for colonoscopy videos. Given a real colonoscopy image to be converted into the style of a synthetic image, we first invert their latent representations over a sequence of time steps, storing the intermediate key, query, and value vectors at each step. Reverse diffusion with attention substitution produces an output that retains the real image's structure while adopting the synthetic style.

the content latent and the keys and values $(K_t^{(s)}, V_t^{(s)})$ of the style latent. We then carry out the reverse diffusion process with an initial latent input of $\tilde{z}_T = \text{AdaIN}(z_T^{(c)}, z_T^{(s)})$. For every step $t$ of the reverse process and for selected self-attention layers in the frozen SD model, we substitute $(K_t^{(c)}, V_t^{(c)})$ with $(K_t^{(s)}, V_t^{(s)})$. By keeping the original content queries $Q_t^{(c)}$ fixed, we preserve the content cues of $x^{(c)}$ during denoising. After completing the reverse diffusion process, we obtain the style-transferred latent code $\tilde{z}^{(c)}$, which we decode into a full-resolution style-transferred image $\tilde{x}^{(c)}$ using the SD model's VAE decoder.

We also introduce three adjustments to facilitate specularity removal and enforce the preservation of depth cues. First, we mask pixels whose intensities exceed 3 standard deviations from the mean intensity of pixels in a $16 \times 16$ patch. During the reverse diffusion process, the SD model inpaints the masked patches to produce smooth style-transferred surfaces. Second, we apply a local histogram matching approach to align the intensities of the style-transferred videos with those of the original content videos. Third, we truncate the inversion process after $T' = \alpha T$ steps, where $\alpha = 0.4$. This approach allows us to reduce the amount of noise added during inversion, thus limiting content drift during the denoising process.

## 3. Experiments and Results

**Datasets** We evaluate ColonCrafter on 10 colonoscopy sequences from C3VD,[9] a dataset featuring realistic colon phantom models with corresponding ground-truth depth annotations. We preprocess all frames following the methodology of Huang et al.[42]

**Baselines** Our evaluation encompasses both general-purpose and endoscopy-specific depth estimation models. For general-purpose baselines, we include the publicly available checkpoints of DepthCrafter[31] and Depth-Anything.[43,44] For domain-specific comparisons, we evaluate against three endoscopy-tailored models: EndoDAC,[45] EndoSfM-Learner,[21] and EndoOmni.[46] To ablate our real-to-synthetic style transfer approach, we also evaluate ColonCrafter directly on the colonoscopy sequences without style transfer.

**Evaluation Metrics** We assess model performance using standard depth estimation metrics: absolute relative error (AbsRel), squared relative error (SqRel), root mean square error (RMSE), and the $\delta_1$ accuracy measure. Rather than applying per-frame alignment, we compute global scale and shift parameters across each complete video sequence by solving:

$$\min_{\alpha,\beta\in\mathbb{R}} \sum_{p\in\Omega} \left[\alpha\hat{d}(p) + \beta - d(p)\right]^2, \tag{2}$$

where $\hat{d}(p)$ and $d(p)$ represent predicted and ground-truth depths at pixel $p$, respectively, with the summation spanning all pixels $\Omega$ in the evaluated sequence. To ensure fair comparison, we perform alignment in each model's native training domain—for example, we apply scale-shift alignment in the disparity domain for models trained on disparity data (such as DepthCrafter and ColonCrafter) before converting to the depth domain for final metric computation.

### 3.1. *Depth Estimation*

Table 1 presents the average zero-shot depth estimation performance of ColonCrafter on C3VD. ColonCrafter demonstrates excellent performance, outperforming both general-purpose and endoscopy-specific depth estimation models. Figure 5 shows qualitative examples of colonoscopy images from C3VD, their ground-truth depth maps, and the depth maps predicted zero-shot by ColonCrafter and other baseline methods.

**Comparison with General-Purpose Models** General-purpose depth estimation models show limited performance on colonoscopy sequences, reflecting the discrepancy between open-world and colonoscopy videos. This degradation stems from the unique visual characteristics of colonoscopy images: constrained lighting conditions, specular reflections, texture-poor surfaces, and limited field of view present fundamentally different challenges compared to natural outdoor scenes. Our colonoscopy-specific fine-tuning approach effectively addresses this discrepancy, improving $\delta_1$ accuracy by more than 17% compared to the base DepthCrafter model. Notably, ColonCrafter outperforms most baseline methods even without style transfer, demonstrating that the model learns robust colonoscopy-specific geometric priors through synthetic data training alone.

**Style Transfer Analysis** Figure 6 qualitatively demonstrates our real-to-synthetic style transfer approach, showing original C3VD images (left), their style-transferred counterparts (middle), and the corresponding photometric intensity difference maps (right). The transformation successfully removes specular highlights and adjusts tissue appearance to match the synthetic training distribution while preserving essential anatomical structure and depth

Table 1. Zero-shot depth estimation performance on C3VD. We compare ColonCrafter with general-purpose and endoscopy-specific monocular depth estimation (MDE) models. ColonCrafter outperforms most baselines even without style transfer. Parentheses indicate 95% confidence intervals computed using 1,000-sample bootstrap resampling.

| **Model** | $\delta_1 \uparrow$ | **AbsRel** $\downarrow$ | **SqRel** $\downarrow$ | **RMSE (mm)** $\downarrow$ |
|---|---|---|---|---|
| Depth-Anything-V1[43] | 0.55 (0.45, 0.66) | 0.28 (0.21, 0.34) | 3.58 (1.83, 5.50) | 10.08 (7.08, 13.24) |
| Depth-Anything-V2[44] | 0.61 (0.52, 0.71) | 0.24 (0.19, 0.28) | 2.34 (1.55, 3.09) | 8.20 (6.33, 9.87) |
| DepthCrafter[31] | 0.59 (0.52, 0.65) | 0.22 (0.19, 0.26) | 2.65 (2.07, 3.14) | 10.51 (9.09, 11.70) |
| EndoDAC[45] | 0.50 (0.42, 0.61) | 0.27 (0.22, 0.33) | 4.45 (2.64, 6.63) | 13.91 (10.45, 17.42) |
| EndoSfM-Learner[9] | 0.56 (0.49, 0.64) | 0.24 (0.21, 0.29) | 3.49 (2.42, 4.68) | 12.34 (9.81, 15.08) |
| EndoOmni[†46] | 0.77 (0.72, 0.81) | 0.15 (0.14, 0.16) | 1.15 (0.88, 1.43) | 6.91 (5.54, 8.20) |
| ColonCrafter | 0.77 (0.66, 0.83) | 0.16 (0.13, 0.20) | 1.17 (0.81, 1.61) | 6.42 (5.09, 7.62) |
| ColonCrafter + ST | 0.79 (0.70, 0.86) | 0.15 (0.12, 0.18) | 1.09 (0.73, 1.45) | 6.21 (5.01, 7.18) |

ST: style transfer; †: Not strictly zero-shot (partially trained on C3VD).

cues. The intensity difference maps reveal that the most significant transformations occur in regions with strong specular reflections and lighting variations, while anatomically critical features such as tissue folds and surface geometry remain largely unchanged. Applying style transfer before inference aligns ColonCrafter with its synthetic training domain, leading to consistent gains across all depth evaluation metrics.

### 3.2. *Downstream Applications*

**Point Cloud Generation** ColonCrafter integrates seamlessly with established SLAM frameworks to produce consistent 3D colon reconstructions. Following the approach of Xu et al.,[47] we track feature points across successive frames using a pre-trained SpaTracker[48] model. We then estimate camera poses through bundle adjustment by minimizing the reprojection error across all tracked points and frame pairs:

$$\min_{W_1,\ldots,W_T} \sum_{i,j\in\{1,\ldots,T\}} \left[\pi_{K_j}(W_j W_i^{-1} \pi_{K_i}^{-1}(p_i, \hat{d}_i)) - p_j\right]^2. \tag{3}$$

Here, $T$ represents the size of our tracking window, $\pi_{K_i}^{-1}$ denotes the backprojection operator that converts 2D pixel coordinates with depth $\hat{d}_i$ to 3D world coordinates using camera intrinsics $K_i$, and $W_i$ represents the camera pose (transformation matrix) for frame $i$.

Figure 7 shows examples of 3D point clouds generated using ColonCrafter for six colon segments in C3VD. These reconstructions provide detailed structural representations of the colon anatomy, clearly delineating missed surface regions (appearing as white, empty spaces) and lesions. In the bottom right point cloud, we color points belonging to a polyp in blue, demonstrating how point clouds can be used for lesion registration. This capability has significant clinical potential for improving the accuracy of longitudinal lesion assessment by facilitating precise lesion tracking and size measurement in follow-up colonoscopies.

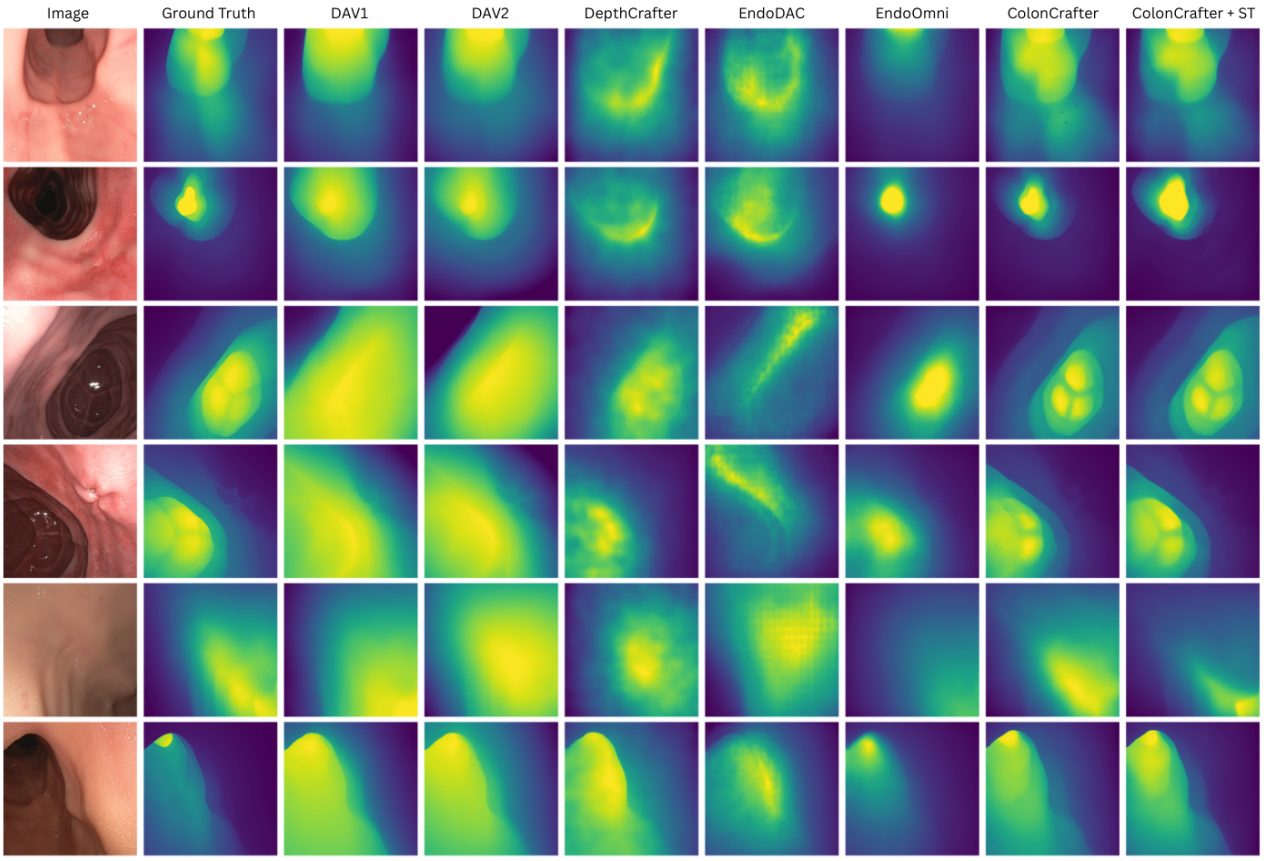


Fig. 5. Qualitative comparison of depth estimation on C3VD colonoscopy sequences. Each row shows the original colonoscopy image, the ground-truth depth map, and predictions from different methods: general-purpose models (DepthCrafter, Depth-Anything), endoscopy-specific approaches (EndoDAC, EndoOmni), and our ColonCrafter with and without style transfer (ST). ColonCrafter + ST yields the most accurate depth maps, with sharper boundaries, better preservation of fine geometric details, and improved handling of specular reflections that confuse other methods.

### 3.3. *Missing Surface Estimation*

Using the 3D reconstructions generated by ColonCrafter, we can assess unseen areas of the colon in an image sequence. For a given point cloud, we estimate its centerline using Principal Component Analysis, then "unroll" it into a 2D coverage map where the horizontal axis represents the distance along the centerline and the vertical axis represents the circumferential angle around it. Figure 8 shows coverage maps computed using ground-truth depths (left) and ColonCrafter predictions (right), demonstrating the model's ability to accurately identify missed surfaces.

## 4. Related Work

### 4.1. *Monocular Depth Estimation*

Monocular depth estimation (MDE) addresses the inherently ill-posed task of predicting dense depth maps from a single RGB image, where one image may correspond to infinitely many

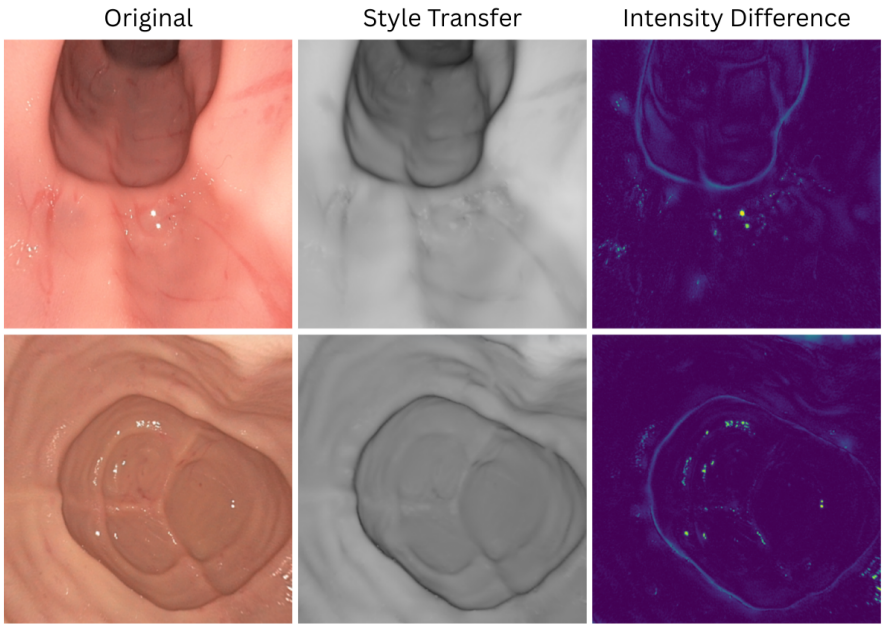


Fig. 6. Demonstration of real-to-synthetic style transfer on C3VD images. Each triplet shows: (left) the original image, (middle) the style-transferred version that mimics synthetic examples while preserving anatomical structure, and (right) the intensity difference map highlighting regions of greatest change. The transfer effectively removes specularities and adapts texture to match the synthetic domain, supporting improved depth estimation.

plausible 3D scenes. Classical geometric methods typically yield sparse reconstructions and are highly sensitive to illumination, motivating the development of learning-based approaches. Recent advances include open-world foundation models such as Depth-Anything[43,44] and DepthCrafter.[31] The latter exploits diffusion priors learned from large-scale video data, reducing the temporal inconsistencies common in single-frame models. In contrast, endoscopy-specific approaches such as EndoDAC[45] and EndoOmni[46] often rely on self-supervised or semi-supervised objectives to overcome the limited availability of ground-truth depth annotations. However, because they are primarily trained at the image level, these models still struggle to maintain temporal coherence across video sequences.

### 4.2. *Image Style Transfer*

An important challenge in endoscopic depth estimation is the mismatch between synthetic training data and real clinical images, which has motivated the use of style transfer for domain adaptation.[23,29,30,39,49] Most prior approaches adopt a synthetic-to-real strategy, training depth estimators directly on translated images to reduce distribution mismatch. In contrast, dynamic adaptation at inference time remains underexplored. Recent diffusion-based style transfer techniques offer a training-free alternative that preserves structural content while adjusting appearance for domain alignment.[41,50,51] For example, Chung et al.[41] manipulate self-attention layers of pre-trained diffusion models by substituting key-value pairs while preserving queries, thereby maintaining the original content during stylization. This property is

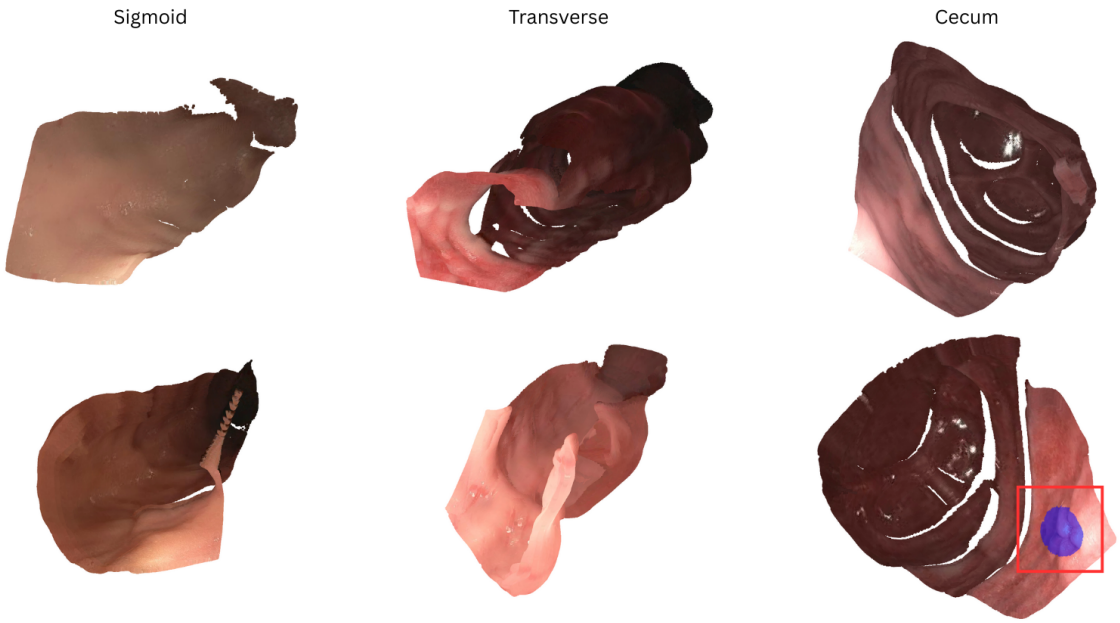


Fig. 7. 3D point cloud reconstructions of colon segments from C3VD using ColonCrafter depth predictions. Each point cloud represents the internal colon structure viewed from different perspectives, with white regions indicating areas not visible according to the reconstruction. Colored annotations highlight anatomical features and potential lesions on the point cloud.

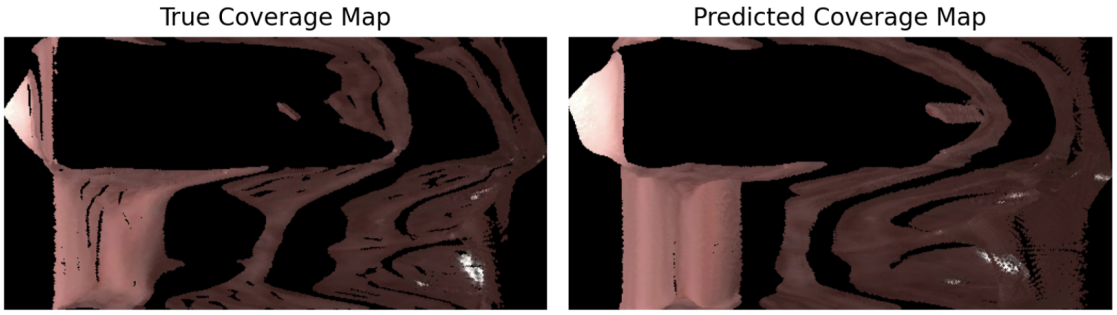


Fig. 8. Colon surface coverage analysis. Coverage maps show the surveyed areas of the colon surface "unrolled" into 2D representations, where the horizontal axis represents distance along the estimated centerline and the vertical axis represents circumferential angle. Black regions indicate areas unseen during the sequence, while colored regions represent seen surfaces. Comparison between ground-truth depth-based coverage (left) and ColonCrafter prediction-based coverage (right) demonstrates the model's relative accuracy in identifying unseen areas.

especially valuable in endoscopy, where retaining geometric cues is essential for accurate 3D reconstruction. Despite their promise, however, diffusion-based style transfer methods remain largely untested in endoscopic applications.

## 5. Discussion

**Clinical Significance** Our work addresses a fundamental challenge in colonoscopy: reconciling the three-dimensional structure of the colon with the two-dimensional visual information available to clinicians. ColonCrafter's ability to generate temporally consistent depth maps represents an important step toward augmenting clinical decision-making with computational tools. The observed improvements in depth estimation accuracy translate into practical benefits for lesion localization, size measurement, and coverage assessment during routine colonoscopic examinations.

**Technical Contributions** By adapting video diffusion models to the colonoscopy domain, we advance generative methods for medical depth estimation and achieve temporally consistent predictions that outperform both general-purpose and endoscopy-specific baselines. In parallel, our style transfer strategy offers a novel means of bridging domain gaps in medical imaging without the need for additional training data.

**Limitations and Future Work** Several limitations highlight directions for future development. First, while evaluation on C3VD phantom data provides initial validation, true clinical value requires testing on real patient procedures that include a wider range of anatomies and disease states. Second, ColonCrafter is not yet suited for full-length colonoscopy workflows, as it currently performs best on shorter video segments (with a runtime of 116 ms per frame on an A100 GPU at a resolution of $512 \times 512$), highlighting the need for further improvements in efficiency and robustness across different endoscopy systems. Finally, we plan to extend ColonCrafter into a semi-supervised framework that learns jointly from real and synthetic videos, reducing dependence on synthetic data and improving generalization to diverse clinical settings.

## 6. Conclusion

We introduced ColonCrafter, a novel diffusion-based depth estimation model for colonoscopy videos, achieving state-of-the-art zero-shot performance on a phantom dataset. By combining video diffusion modeling with a style transfer strategy, our method addresses the challenge of temporal consistency and bridges the domain gap between synthetic training data and real colonoscopy images. ColonCrafter outperforms both general-purpose and endoscopy-specific baselines, particularly in handling difficult visual conditions such as specular highlights and complex tissue geometry. Although the current framework is optimized for short video segments with smooth camera motion, it provides a strong foundation for future advances in computational colonoscopy. More broadly, it shows that synthetic data, when paired with domain adaptation, can support practical clinical applications including 3D reconstruction, lesion localization, and surface coverage analysis.

## References


1. R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung and A. Jemal, Cancer statistics, 2025, *Ca* **75**, p. 10 (2025).
2. D. K. Rex, C. R. Boland, J. A. Dominitz, F. M. Giardiello, D. A. Johnson, T. Kaltenbach, T. R. Levin, D. Lieberman and D. J. Robertson, Colorectal cancer screening: recommendations for physicians and patients from the us multi-society task force on colorectal cancer, *Gastroenterology* **153**, 307 (2017).
3. S. Zhao, S. Wang, P. Pan, T. Xia, X. Chang, X. Yang, L. Guo, Q. Meng, F. Yang, W. Qian *et al.*, Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis, *Gastroenterology* **156**, 1661 (2019).
4. A. C. Thompson, R. H. Jones, P. D. Poullos, S. Banerjee and L. K. Shin, Taller haustral folds in the proximal colon: A potential factor contributing to interval colorectal cancer, *Journal of Colon and Rectal Cancer* **1**, 45 (2016).
5. S. Mathew, S. Nadeem and A. Kaufman, Foldit: Haustral folds detection and segmentation in colonoscopy videos, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
6. B. M. Shandro, K. Emrith, G. Slabaugh, A. Poullis and M. L. Smith, Optical imaging technology in colonoscopy: Is there a role for photometric stereo?, *World Journal of Gastrointestinal Endoscopy* **12**, p. 138 (2020).
7. S. A. O'Connor, D. G. Hewett, M. O. Watson, B. J. Kendall, L. F. Hourigan and G. Holtmann, Accuracy of polyp localization at colonoscopy, *Endoscopy international open* **4**, E642 (2016).
8. M. Izzy, M. A. Virk, A. Saund, J. Tejada, F. Kargoli and S. Anand, Accuracy of endoscopists' estimate of polyp size: A continuous dilemma, *World journal of gastrointestinal endoscopy* **7**, p. 824 (2015).
9. 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* **90**, p. 102956 (2023).
10. 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.
11. D. Soriero, P. Batistotti, R. Malinaric, D. Pertile, A. Massobrio, L. Epis, B. Sperotto, V. Penza, L. S. Mattos, M. Sartini *et al.*, Efficacy of high-resolution preoperative 3d reconstructions for lesion localization in oncological colorectal surgery—first pilot study, in *Healthcare*, (5)2022.
12. R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman and J.-M. Frahm, Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy, *Medical image analysis* **72**, p. 102100 (2021).
13. M. V. Golhar, L. S. G. Fretes, L. Ayers, V. S. Akshintala, T. L. Bobrow and N. J. Durr, C3vdv2–colonoscopy 3d video dataset with enhanced realism, *arXiv preprint arXiv:2506.24074* (2025).
14. A. Schmidt, O. Mohareri, S. DiMaio, M. C. Yip and S. E. Salcudean, Tracking and mapping in medical computer vision: A review, *Medical Image Analysis* **94**, p. 103131 (2024).
15. S. Pyykölä, N. Joswig and L. Ruotsalainen, Non-lambertian surfaces and their challenges for visual slam, *IEEE Open Journal of the Computer Society* **5**, 430 (2024).
16. A. Mathew, L. Magerand, E. Trucco and L. Manfredi, Self-supervised monocular depth estimation for high field of view colonoscopy cameras, *Frontiers in Robotics and AI* **10**, p. 1212525 (2023).
17. J. Liu, K. R. Subramanian and T. S. Yoo, A robust method to track colonoscopy videos with non-informative images, *International journal of computer assisted radiology and surgery* **8**, 575 (2013).

18. H. Yao, R. W. Stidham, Z. Gao, J. Gryak and K. Najarian, Motion-based camera localization system in colonoscopy videos, *Medical image analysis* **73**, p. 102180 (2021).
19. 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*, 2021.
20. S.-J. Hwang, S.-J. Park, G.-M. Kim and J.-H. Baek, Unsupervised monocular depth estimation for colonoscope system using feedback network, *Sensors* **21**, p. 2691 (2021).
21. 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* **71**, p. 102058 (2021).
22. A. Lou and J. Noble, Ws-sfmlearner: self-supervised monocular depth and ego-motion estimation on surgical videos with unknown camera parameters, in *Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2024.
23. 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* **14**, 1167 (2019).
24. F. Mahmood and N. J. Durr, Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy, *Medical Image Analysis* **48**, 230 (2018), Early work leveraging synthetic training data to compensate for scarcity of RGB-Depth pairs in endoscopy.
25. H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S.-e. Kudo and K. Mori, Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning, in *International conference on medical image computing and computer-assisted intervention*, 2018.
26. T. Teufel, H. Shu, R. D. Soberanis-Mukul, J. E. Mangulabnan, M. Sahu, S. S. Vedula, M. Ishii, G. Hager, R. H. Taylor and M. Unberath, Oneslam to map them all: a generalized approach to slam for monocular endoscopic imaging based on tracking any point, *International Journal of Computer Assisted Radiology and Surgery* **19**, 1259 (2024).
27. J. Xu, Q. Zhang, Y. Yu, R. Zhao, X. Bian, X. Liu, J. Wang, Z. Ge and D. Qian, Deep reconstruction-recoding network for unsupervised domain adaptation and multi-center generalization in colonoscopy polyp detection, *Computer methods and programs in biomedicine* **214**, p. 106576 (2022).
28. D. He, Z. Liu, X. Yin, H. Liu, W. Gao and Y. Fu, Synthesized colonoscopy dataset from high-fidelity virtual colon with abnormal simulation, *Computers in Biology and Medicine* **186**, p. 109672 (2025).
29. A. Dinkar Jagtap, M. Heinrich and M. Himstedt, Automatic generation of synthetic colonoscopy videos for domain randomization, *Current Directions in Biomedical Engineering* **8**, 121 (2022).
30. S. Wang, A. Paruchuri, Z. Zhang, S. McGill and R. Sengupta, Structure-preserving image translation for depth estimation in colonoscopy video, *arXiv preprint arXiv:2408.10153* (2024).
31. W. Hu, X. Gao, X. Li, S. Zhao, X. Cun, Y. Zhang, L. Quan and Y. Shan, Depthcrafter: Generating consistent long depth sequences for open-world videos, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
32. T. Karras, M. Aittala, T. Aila and S. Laine, Elucidating the design space of diffusion-based generative models, *Advances in neural information processing systems* **35**, 26565 (2022).
33. K. Smith, K. Clark, W. Bennett, T. Nolan, J. Kirby, M. Wolfsberger, J. Moulton, B. Vendt and J. Freymann, Data from ct colonography. the cancer imaging archive (2015).
34. A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer,

D. Jennings, F. Fennessy, M. Sonka *et al.*, 3d slicer as an image computing platform for the quantitative imaging network, *Magnetic resonance imaging* **30**, 1323 (2012).
35. N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson and G. Gkioxari, Accelerating 3d deep learning with pytorch3d, *arXiv:2007.08501* (2020).
36. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, Lora: Low-rank adaptation of large language models., *ICLR* **1**, p. 3 (2022).
37. I. Loshchilov and F. Hutter, Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).
38. F. Mahmood, R. Chen and N. J. Durr, Unsupervised reverse domain adaptation for synthetic medical images via adversarial training, *IEEE transactions on medical imaging* **37**, 2572 (2018).
39. S. Mathew, S. Nadeem, S. Kumari and A. Kaufman, Augmenting colonoscopy using extended and directional cyclegan for lossy image translation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
40. R. J. Chen, T. L. Bobrow, T. Athey, F. Mahmood and N. J. Durr, Slam endoscopy enhanced by adversarial depth prediction, *arXiv preprint arXiv:1907.00283* (2019).
41. J. Chung, S. Hyun and J.-P. Heo, Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.
42. Y. Huang, B. Cui, L. Bai, Z. Chen, J. Wu, Z. Li, H. Liu and H. Ren, Advancing dense endoscopic reconstruction with gaussian splatting-driven surface normal-aware tracking and mapping, *arXiv preprint arXiv:2501.19319* (2025).
43. L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng and H. Zhao, Depth anything: Unleashing the power of large-scale unlabeled data, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
44. L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng and H. Zhao, Depth anything v2, *Advances in Neural Information Processing Systems* **37**, 21875 (2024).
45. 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*, 2024.
46. Q. Tian, Z. Chen, H. Liao, X. Huang, L. Li, S. Ourselin and H. Liu, Endoomni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels, *arXiv preprint arXiv:2409.05442* (2024).
47. T.-X. Xu, X. Gao, W. Hu, X. Li, S.-H. Zhang and Y. Shan, Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors, *arXiv preprint arXiv:2504.01016* (2025).
48. Y. Xiao, Q. Wang, S. Zhang, N. Xue, S. Peng, Y. Shen and X. Zhou, Spatialtracker: Tracking any 2d pixels in 3d space, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
49. X. Xiong, A. D. Beltran, J. M. Choi, M. Niethammer and R. Sengupta, Pps-ctrl: Controllable sim-to-real translation for colonoscopy depth estimation, *arXiv preprint arXiv:2504.17067* (2025).
50. Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong and C. Xu, Inversion-based style transfer with diffusion models, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.
51. Z. Wang, L. Zhao and W. Xing, Stylediffusion: Controllable disentangled style transfer via diffusion models, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.