HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2024 December 01.

Colonoscopy 3D video dataset with paired depth from 2D-3D registration

Taylor L. Bobrow[a], **Mayank Golhar**[a], **Rohan Vijayan**[a], **Venkata S. Akshintala**[b], **Juan R. Garcia**[c], **Nicholas J. Durr**[a,*]

[a]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

[b]Division of Gastroenterology and Hepatology, Johns Hopkins Medicine, Baltimore, MD 21287, USA

[c]Department of Art as Applied to Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Abstract

Screening colonoscopy is an important clinical application for several 3D computer vision techniques, including depth estimation, surface reconstruction, and missing region detection. However, the development, evaluation, and comparison of these techniques in real colonoscopy videos remain largely qualitative due to the difficulty of acquiring ground truth data. In this work, we present a Colonoscopy 3D Video Dataset (C3VD) acquired with a high definition clinical colonoscope and high-fidelity colon models for benchmarking computer vision methods in colonoscopy. We introduce a novel multimodal 2D-3D registration technique to register optical video sequences with ground truth rendered views of a known 3D model. The different modalities are registered by transforming optical images to depth maps with a Generative Adversarial Network and aligning edge features with an evolutionary optimizer. This registration method achieves an average translation error of 0.321 millimeters and an average rotation error of 0.159 degrees in simulation experiments where error-free ground truth is available. The method also leverages video information, improving registration accuracy by 55.6% for translation and 60.4% for rotation compared to single frame registration. 22 short video sequences were registered to generate 10,015 total frames with paired ground truth depth, surface normals, optical flow, occlusion, six degree-of-freedom pose, coverage maps, and 3D models. The dataset also includes screening videos acquired by a gastroenterologist with paired ground truth pose and 3D surface models. The dataset and registration source code are available at durr.jhu.edu/C3VD.

Keywords

*Corresponding author: ndurr@jhu.edu (Nicholas J. Durr).

1. Introduction

Colorectal cancer (CRC) is the second most lethal form of cancer in the United States (Siegel et al., 2021). At least 80% of CRCs are believed to develop from premalignant adenomas (Cunningham et al., 2010). Screening colonoscopy remains the gold-standard for detecting and removing precancerous lesions, effectively reducing the risk of developing CRC (Rex et al., 2015). Still, an estimated 22% of precancerous lesions go un-detected during screening procedures (van Rijn et al., 2006). These missed lesions are thought to be a primary contributor to interval CRC — the development of CRC within 5 years of a negative screening colonoscopy — which represents 6% of CRC cases (Samadder et al., 2014).

Colonoscopy remains an active application for computer vision researchers working to reduce lesion miss rates and improve clinical outcomes (Ali et al., 2020; Fu et al., 2021; Chadebecq et al., 2023). Recent works have employed Convolutional Neural Networks (CNNs) to detect and alert clinicians of visible, but sometimes subtle, polyps in colonoscopy video frames (Hassan et al., 2020; Livovsky et al., 2021; Luo et al., 2021). While these algorithms have demonstrated impressive detection rates, they require that polyps appear in the colonoscope field of view (FoV) during a procedure to be detected. Missed regions — areas of the colon never imaged during a screening procedure — were found to make up an estimated 10% of the colon surface in a retrospective analysis of endoscopic video (McGill et al., 2018), putting patients at risk of experiencing interval CRC.

To reduce the extent of missed regions, research has explored measuring observational coverage of the colon during screening procedures (Hong et al., 2007, 2011; Armin et al., 2016). Freedman et al. (2020) describe a data-driven method for directly regressing a numerical visibility score for a small cluster of frames using deep learning. Some methods utilize deep learning to regress pixel-level depth (Mahmood and Durr, 2018; Rau et al., 2019; Cheng et al., 2021), and this information can be incorporated into simultaneous localization and mapping (SLAM) techniques to reconstruct the colon surface (Chen et al., 2019b). Holes in these reconstructions could indicate tissue areas that have gone unobserved, and these regions have been flagged in real time to alert the colonoscopist (Ma et al., 2019). Other relevant applications of 3D computer vision techniques in colonoscopy include polyp size prediction (Abdelrahim et al., 2022), surface topography reconstruction (Parot et al., 2013), visual odometry estimation (Yao et al., 2021), and enhanced lesion classification with 3D augmentation (Mahmood et al., 2018b).

1.1. Related work

1.1.1. Endoscopy reconstruction datasets—Datasets for evaluating endoscopic reconstruction methods differ by intended application, model type, recording setup, and ground truth data availability. Acquiring datasets with accurate surface information in the surgical environment is generally impractical. As an alternative, both commercial (Stoyanov et al., 2010) and computed tomography (CT)-derived (Penza et al., 2018) silicone models have been explored. Surface information can be directly measured with either CT or optical scanning (OS), and the stereo sensor configuration in some laparoscopes can be used to generate ground truth depth (Recasens et al., 2021). Beyond synthetic models, which suffer from relatively homogeneous optical properties, both *ex-vivo* (Mahmood and Durr, 2018;

Edwards et al., 2022; Allan et al., 2021; Maier-Hein et al., 2014; Ozyoruk et al., 2021) and *in-vivo* (Ye et al., 2017) animal tissues have been used to generate data with a more realistic bidirectional scattering distribution functions (BSDF).

More recently, game engines, such as Unity (Unity Technologies) have been used to render synthetic images from 3D anatomical models, such as CT colonography volumes (Mahmood et al., 2018a; Rau et al., 2019; Ozyoruk et al., 2021; Zhang et al., 2021; Rau et al., 2022). Rendered data is advantageous because error-free, pixel-level ground truth labels such as depth and surface normals are available from rendering primitives. Additionally, large quantities of data may be quickly produced. One significant drawback is the limited ability of rendering engines to simulate real-world camera optics, non-global illumination, proprietary post-acquisition processing, sensor noise, and light-tissue interaction.

The majority of endoscopic reconstruction datasets with real images are designed for laparoscopic imaging of the abdomen and thorax. Laparoscopic datasets are unsuitable for benchmarking colonoscopic imaging because of large differences in the angular FoV (and corresponding distortion), sensor arrangement, and organ geometry. Furthermore, it is challenging to mimic the realistic motions of a colonoscope with a rigid laparoscope. Ozyoruk et al. provide a video dataset for imaging the stomach, small bowel, and colon with several different camera types with one sequence recorded using a clinical colonoscope (Ozyoruk et al., 2021). Each video sequence in this dataset is paired with a ground truth camera trajectory and 3D surface model, but no pixel-level ground truth data was generated, limiting the utility of the data. Generating pixel-level ground truth information is particularly challenging for colonoscopy, due to the space-constrained imaging environment, high resolution requirements, and large range of working distances that are relevant for clinical applications.

1.1.2. Registering real and virtual endoscopy images—Registering endoscopic frames with ground truth surface models enables derivative ground truth data and metrics to be extracted. For example, ground truth depth frames may be assigned to real endoscopy images provided the endoscope pose relative to the surface model is known. However, registration using conventional, feature-based methods is challenging due to a lack of robust corner points and variable specular reflections common in endoscopic images. To circumvent this challenge, segmented fiducials have been used to register optical images to ground truth CT volumes (Rau et al., 2019; Stoyanov et al., 2010). While this method is robust, one drawback is the presence of unrealistic fiducials throughout the image FoV. Edwards et al. (2022) opt to remove the fiducials and instead rely on the manual alignment of a virtual camera with a ground truth CT volume. The manual nature of this method works well for producing small quantities of data, but it is a barrier to registering large quantities of data, and its accuracy is limited by inter-operator variability. Penza et al. (2018) use a calibration target to calibrate the coordinate systems of a fixed camera and laser scanner, allowing for simultaneous recording of endoscopy frames and 3D surface information in laparoscopic imaging environments that are not size-constrained.

A summary of existing and the proposed endoscopic 3D datasets and acquisition methods is reported in Table 1.

1.1.3.   2D-3D registration—2D-3D registration enables the registration of a 2D image with a 3D spatial volume, and it is frequently used for aligning 3D preoperative CT volumes with 2D intraoperative X-ray images. Most methods rely on optimizing the similarity between a target 2D image and a simulated 2D radiograph of the 3D volume acquired at the estimated pose (Markelj et al., 2012). Common similarity measures are gradient- (Livyatan et al., 2003), intensity- (Birkfellner et al., 2003), and feature-based (Groher et al., 2007) metrics. More recently, 2D-3D registration methods have evolved to include learning-based algorithms to address challenges in cross-modality registration (Oulbacha and Kadoury, 2020) and feature extraction (Grupp et al., 2020).

1.2.   Contributions

While the body of computer vision research in colonoscopy is extensive, evaluating and benchmarking methods remains a challenge due to a lack of ground truth annotated data. In this work, we present a High Definition (HD) Colonoscopy 3D Video Dataset (C3VD) for quantitatively evaluating computer vision methods. To the best of our knowledge, this is the first video dataset with 3D ground truth that is recorded entirely with a clinical colonoscope. In this work, we contribute:

1.   a 2D-3D video registration algorithm for aligning real 2D optical colonoscopy video sequences with ground truth 3D models. GAN-estimated depth frames are compared with rendered predicted views along a measured camera trajectory for minimizing an edge-based loss.

2.   a technique for generating high-fidelity silicone phantom models with varying textures and colors to facilitate domain randomization.

3.   a ground truth dataset with pixel-level registration to a known 3D model for quantitatively evaluating computer vision techniques in colonoscopy. This dataset contains 10,015 HD video frames of realistic colon phantom models obtained with a clinical colonoscope. Each frame is paired with ground truth depth, surface normals, occlusion, optical flow, and a six degree-of-freedom camera pose. Each video is paired with a ground truth surface model and coverage map (Figure 1).

The dataset, 3D colon model, phantom molds, and 2D-3D registration algorithm are all made publicly available at durr.jhu.edu/C3VD.

2.   Methods

We propose a method for generating video sequences through a clinical colonoscope with paired, pixel-level ground truth. We first describe a protocol for producing high-fidelity phantom models (Section 2.1) and recording video sequences with ground truth trajectory (Section 2.2). We then introduce a novel technique for registering the acquired video and trajectory sequences with a ground truth 3D surface model (Section 2.3).

2.1. Phantom model production

A complete 3D colon model - from sigmoid colon to cecum - was digitally sculpted in Zbrush (Pixologic) by a boardcertified anaplastologist (JRG) using reference anatomical imagery from colonoscopic procedures. This method introduces higher frequency detail in the digital models as compared to models derived from resolution-limited CT colonography. We split the sculpted model into five segments: the sigmoid colon, descending colon, transcending colon, ascending colon, and cecum. Three-part molds were generated for each segment. Two parts comprised two halves of the outer shell and one insert part formed the colon lumen and mucosal surface. All molds were 3D printed with an Objet 260 Connex 3 with 16 micrometer resolution. Casts of each mold were created with silicone (Dragon Skin™, Smooth-On, Inc.). Silicone pigments (Silc Pig™, Smooth-On, Inc.) were used to vary the color and texture. Silicone was manually applied in 5–12 layers with varying degrees of opacity to emulate patient-specific tissue features and vasculature patterns at varying optical depths. A silicone lubricant (015594011516, BioFilm, Inc.) was applied to the surface of the models at recording time to simulate the highly specular appearance of the mucosa.

2.2. Data acquisition

The setup used for data acquisition is shown in Figure 2. An Olympus CF-HQ190L video colonoscope, CV-190 video processor, and CLV-190 light source were used to record video sequences of the phantom models. The models were placed inside the molds with inserts removed to keep them static and free of deformation during video recording. The tip of the colonoscope was rigidly mounted to a UR-3 (Universal Robotics) robotic arm. For each video segment, 4–8 sequential poses were manually programmed to mimic a typical colonoscopy trajectory. The robotic arm traversed the colon with interpolation between these poses with 10 micrometer repeatability. A pose log was recorded from the arm at a sampling rate of 63 Hz. Colonoscopy videos were recorded in an uncompressed HD format using an Orion HD (Matrox) frame grabber connected to the SDI output of the Olympus video processor.

2.3. Registration pipeline overview

Pixel-level ground truth for each video frame was generated by moving a virtual camera along the recorded trajectory and rendering ground truth frames of the 3D model. While the colonoscope trajectory for the virtual camera was known, the location of the phantom model relative to this trajectory was unknown. Assuming the phantom model was stationary for the duration of a video sequence, the phantom pose can be expressed as a single rigid body transform ($\mathbf{T}_{\text{final}}$), consisting of a rotational component (θ) and translational component ($\mathbf{t}$). We parameterize this transformation using three Euler angles and a translation vector ($[\theta_\alpha\ \theta_\beta\ \theta_\gamma\ \mathbf{t}_x\ \mathbf{t}_y\ \mathbf{t}_z]$). To estimate this unknown transform, we utilize a 2D-3D registration approach to align geometric features shared between the 2D video frames and the virtual 3D model of the phantom. The registration pipeline samples the parameter space for a model transform prediction ($\mathbf{T}_i$), evaluates the feature alignment between target depth frames and renderings at the current model transform, and updates the model transform prediction using an evolutionary optimizer. This registration method is outlined in Figure 3 and detailed in the following subsections.

2.3.1. Data preprocessing—Before beginning the registration optimization, each video sequence, pose log, and ground truth 3D model were preprocessed. For a video sequence consisting of N frames, keyframes were sampled at an interval of δ, resulting in K total keyframes. A depth frame was then estimated for each keyframe using a Generative Adversarial Network trained to estimate depth (Section 2.3.2).

For each keyframe, a pose $\mathbf{A}_i$ describing the position of the robotic arm end-effector relative to the base was sampled from the pose log. Synchronization between the video sequence and pose log was achieved by solving for the relative temporal offset that resulted in the maximum correlation between the optical flow magnitude and the pose displacement of the camera. A handeye calibration was performed to characterize the transformation, $\mathbf{X}$, between robotic arm pose, $\mathbf{A}_i$, and colonoscope camera pose ($\mathbf{B}_i$). This calibration allows the relationship

$$\mathbf{A}_a^{-1}\mathbf{A}_b\mathbf{X} = \mathbf{X}\mathbf{B}_a^{-1}\mathbf{B}_b,\qquad(1)$$

where $\mathbf{A}_a/\mathbf{B}_a$ and $\mathbf{A}_b/\mathbf{B}_b$ are pairs of robot and camera poses captured for calibration. Once calibrated, robotic arm poses were transformed to camera poses by solving a rearranged version of the handeye relationship

$$\mathbf{B}_i = \mathbf{B}_{cal}\mathbf{X}^{-1}\mathbf{A}_{cal}^{-1}\mathbf{A}_i\mathbf{X},\qquad(2)$$

where $\mathbf{A}_{cal}$ and $\mathbf{B}_{cal}$ are a pair of poses retained from the calibration.

Finally, the ground truth triangulated mesh was converted to a Split Bounding Volume Hierarchy (SBVH), and a rendering context was created for rendering depth frames from the 3D ground truth model (Section 2.3.3). The initial model transform ($\mathbf{T}_{\text{initial}}$) for each video sequence was manually aligned. We used a custom graphical user interface that overlayed the first keyframe with frames rendered at camera pose $\mathbf{B}_1$ as the model transform was manually perturbed.

2.3.2. Target depth estimation—Predicting pixel-level depth for each keyframe was formulated as an image-to-image translation task. A conditional generative adversarial network (cGAN) was trained with synthetically rendered input-output image pairs, and inference was performed on real images. Previous studies demonstrating strong domain generalizability when trained on synthetic data and applied to real data motivated using a cGAN network architecture (Chen et al., 2018; Rau et al., 2019). 1,000 pairs of synthetic colonoscopy images with paired depth were rendered using the virtual 3D models and 3 unique BSDFs for domain randomization. The descending colon model and a fourth BSDF were omitted from the training data and saved for validation experiments. To enable training at HD-resolution, multi-scale discriminator models and a multi-layer feature matching loss were employed in addition to the traditional GAN loss (Wang et al., 2018). Keyframes were fed to the trained generator to produce target depth frames for alignment.

2.3.3. Rendering depth frames—Depth frames at each camera pose were rendered using a virtual camera in the 3D colon model, and these frames were compared to the target

depth frames to evaluate the accuracy of the current model transform prediction. A wide FoV ($\approx$170°) is a key characteristic of clinical colonoscopes that maximizes the visibility of the peripheral colon surface. However, this fisheye effect has not been simulated in synthetic endoscopy datasets (Ozyoruk et al., 2021; Rau et al., 2022). To model the entire FoV of a commercial colonoscope, we adopted a spherical camera intrinsic model (Scaramuzza et al., 2006) as shown in Figure 4.

For each pixel [u v], there exists a ray direction $\vec{V} \in \mathbb{N}^3$ emanating from the camera origin $O \in \mathbb{R}^3$ into world space. Pixel coordinates are first shifted with respect to the optical center of the camera $c \in \mathbb{R}^{2 \times 1}$, and then skewed with an affine transformation $A \in \mathbb{R}^{2 \times 2}$ to account for lens misalignment, producing distorted pixel coordinates $[u'v'] \in \mathbb{R}^2$:

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = A^{-1} \begin{bmatrix} u - c_x \\ v - c_y \end{bmatrix}, \ A = \begin{bmatrix} e & f \\ g & 1 \end{bmatrix}. \tag{3}$$

Distorted pixel coordinates and corresponding ray directions are related by a parametric equation

$$\begin{bmatrix} \vec{V}_x \\ \vec{V}_y \\ \vec{V}_z \end{bmatrix} = \begin{bmatrix} u' \\ v' \\ f(\rho) \end{bmatrix} \tag{4}$$

where $\rho = \sqrt{u'^2 + v'^2}$, $f(\rho) = \alpha_0 + \alpha_2\rho^2 + \alpha_3\rho^3 + \alpha_4\rho^4$, and $\alpha_0 \ldots \alpha_4$ are coefficient parameters solved for during calibration. $\vec{V}$ is transformed by the rotational matrix for the current camera pose and the resulting ray direction is cast from the ray origin.

For each optimization iteration, the model transform is first updated to the current sample ($T_{i-1} \rightarrow T_i$), and depth frames ($\mathscr{D}_T$) are rendered at each virtual camera pose ($B_1 \cdots B_K$). For efficient computation, we implement a custom raycasting engine using the Nvidia®OptiX™ raytracing API. This enables direct access to the GPU RT cores (Parker et al., 2010). The model SBVH is used to initialize an acceleration structure, and the structure is updated with each new model transform.

2.3.4. Edge loss E optimization—Target and rendered depth frames are compared to evaluate the accuracy of the current prediction. To minimize the effect of scale inconsistency that is common in depth predicted by deep learning, we compare geometric contours from depth discontinuities in the target and model depth frames. Contours are extracted using Canny edge extraction and binarized. To provide a continuous and smooth loss function when comparing target and model depth frames, we blurred these edges with a normalized Gaussian kernel, denoting this operation as $\hat{\mathscr{E}} = E(\mathscr{D})$. The registration optimization then aims to maximize the overlap of these blurred edge frames by computing a similarity metric between frame pairs,

$$sim\left(\hat{\mathscr{E}}, \mathscr{E}\right) = \frac{1}{KHW} \sum_{k=1}^{K} \sum_{v=1}^{H} \sum_{u=1}^{W} \hat{\mathscr{E}}_k\left(u, v\right) \cdot \mathscr{E}_k\left(u, v\right), \tag{5}$$

where $\hat{\mathscr{E}}$ is the set of rendered edge frames, $\mathscr{E}$ is the set of target edge frames, and K is the number of frame pairs. The output of this similarity function is a value between 0.0 and 1.0 that is maximal when the target and rendered edges are equivalent and perfectly aligned. To reframe the function as a minimization problem, the similarity value is subtracted from 1.0, yielding the full objective function

$$\mathbf{T}_{\text{final}} = \underset{\mathbf{T}}{\text{argmin}}(1.0 - sim(E(\mathscr{D}_{\mathbf{T}}), \mathscr{E})), \tag{6}$$

Evaluation samples for the model transform are iteratively refined by an evolutionary optimizer called Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen et al., 2003).

2.4. Simulated screening colonoscopies

In addition to the video segments with pixel-level ground truth that can be computed with robot-arm fine pose information, we recorded four simulated screening colonoscopy video sequences with coarse six degree-of-freedom pose and 3D surface models. The phantom model segments were adhered together with silicone adhesive (Sil-Poxy™, Smooth-On, Inc.) and mounted in a laser cut foam scaffold. An electromagnetic field generator (Aurora, Northern Digital Inc.) was positioned above the model, and a six degree-of-freedom electromagnetic sensors (EM, 610016, Northern Digital Inc.) was rigidly affixed to the distal tip of the scope for recording pose information at 40Hz. Beginning at the cecum, the scope was withdrawn by a trained gastroenterologist (VSA) while video and pose information was recorded. The EM poses were synchronized and transformed to camera poses using the same process described in Section 2.3.1. If the sensor failed to track for a portion of the trajectory, temporally neighboring poses were linearly interpolated.

3. Experiments and results

3.1. Implementation

A cross-section of the complete sculpted colon model is shown in Figure 5. Model segment lengths as well as lesion types and major axis diameters are reported in Table 2. The 3D printed molds for this model were used to cast four complete sets of phantoms. Each phantom set was cast with unique material properties.

Camera intrinsic parameters for the CF-HQ190L video colonoscope were measured using 30 images of a 10×15 checkerboard with 10×10 millimeter squares and the Matlab 2022a Camera Calibration Toolbox (Mathworks). The calibration resulted in a 0.47 pixel mean reprojection error. The final results of this calibration are reported in Table 3. The camera extrinsic poses estimated from the calibration were used with the corresponding robot poses

to compute the handeye transform **X**. The solution for **X** was found using the optimization method proposed by Park and Martin (1994).

The depth estimation network architecture was implemented in Pytorch. The initial learning rate of 0.0002 was held constant for the first 50 epochs, then linearly decayed for the final 50 epochs. Input frames were zero-padded to a size of 1088×1376 and fed to the network with a batch size of 1 The training computation was split across 4 Nvidia® RTX A5000 Graphics Processing Units, taking 31 hours to complete. Real colonoscopy video frames were then input into the trained generator to produce target depth frames. Because the target depth frames remain constant during registration optimization, the target edge features are extracted once before beginning iterative optimization.

The complete registration pipeline was implemented in C++ and CUDA and computed using an Nvidia® RTX 2080 TI Graphics Processing Unit. Sample evaluations were rendered, edges extracted, and a loss computed at an average rate of 4.0 milliseconds per keyframe.

CMA-ES iterations were computed with a population size of 100 samples. The parameter space was bounded to $\theta = \pm 0.1$ radians and $\mathbf{t} = \pm 7.5$ millimeters with respect to ($\mathbf{T}_{\text{initial}}$). The bounds were linearly scaled to a uniform space during parameter sampling ($\sigma = 0.1$) to account for the difference in scale between rotation and translation parameters.

3.2. Evaluation with synthetic data

To quantitatively validate the registration error of the proposed algorithm, 10 synthetic video sequences, each composed of 200 frames, were rendered with a 3D model and a BSDF not included in the dataset used to train the depth estimation network. In each experiment, the model initial transform was first manually aligned, then optimized using the registration pipeline. To quantify the registration accuracy, we compute the error $\mathbf{T}_{\text{error}}$ as

$$\mathbf{T}_{\text{error}} = \left(\mathbf{T}_{\text{gt}}\right)^{-1}\mathbf{T}_{\text{final}}, \tag{7}$$

where $\mathbf{T}_{\text{gt}}$ is the ground truth model transform and $\mathbf{T}_{\text{final}}$ is the final registered transform, both in homogeneous form. The homogeneous transform $\mathbf{T}_{\text{error}}$ is broken into rotation and translation components, converting the rotation component to Euler form (θ), then independently evaluating $\| \theta \|$ and $\| \mathbf{t} \|$.

Using these synthetic sequences and validation metrics, we first investigated the effect of number of keyframes K on registration accuracy. Optimizations were run for all 10 video sequences with K ranging from 1 to 10. Results are presented in Figure 6. We find that the rate of improvement in registration accuracy significantly decreases at K equal to 5, achieving a 0.321 millimeter and 0.159 degree accuracy. This resulted in an improvement of the translational accuracy by 55.6% and rotational accuracy by 60.4% compared to using only a single frame. With this in mind, K is set equal to 5 for the remaining experiments to balance registration accuracy with time-efficiency of computation.

To evaluate the effect of using blurred edge features in the loss function, we experimented with optimizing the alignment of depth frames prior to edge extraction. Depth-based

registrations were optimized using L_1, L_2, Normalized Cross Correlation (NCC), and Gradient Correlation (GC) loss functions (De Silva et al., 2016). To validate our loss function selection for edges, we also experimented with L_1, L_2, DICE (using binarized edges), and NCC losses with edge frames. The results for each input type and loss function are reported in Table 4.

To assess the benefit of registering in the depth domain, we registered synthetic video frames without depth transformation to predicted views rendered with a Lambertian BSDF. Edges were extracted from the color synthetic frames and rendered predicted views and a loss was computed using the proposed function. This method resulted in an average registration accuracy that was an order of magnitude larger than registration in the depth domain (Figure 6).

Lastly, we examined how trajectory complexity affects registration accuracy. Synthetic colonoscopy sequences were rendered using three trajectory types with increasing complexity: simple (linear translation only), moderate (helical translation only), and complex (helical translation with pitch and yaw rotation). 10 synthetic sequences per trajectory type were rendered and registered to the ground truth 3D model. This experiment resulted in 0.117-degree, 0.063-degree, and 0.070-degree rotational error for simple, moderate, and complex trajectories, respectively. The translational errors were 0.150-millimeter, 0.070-millimeter, and 0.089-millimeter for simple, moderate, and complex trajectories, respectively.

3.3. Evaluation with real video sequences

Once the registration pipeline was validated using synthetic data with available ground truth, the algorithm was then applied to real recorded videos and poses. While the ground truth model transform for real data is unknown, the quality of each registration may be evaluated by qualitatively comparing the edge alignment quality across each video. Sample results for a recorded sequence are reported in Figure 7. The results demonstrate that the registration pipeline aligns sample frames with the target frames throughout the entire video.

4. Dataset and distribution format

22 video sequences consisting of 10,015 frames were registered for inclusion in the dataset. The illumination conditions in each sequence were varied by adjusting the illumination mode (automatic versus manual) and illumination power settings on the clinical light source. The sequences also include a combination of "down-the-barrel", "en-face", and partially occluded views. A listing of the dataset video sequences is included in Table 5. Once registered, the virtual camera was moved to the camera pose corresponding with each video frame, and ground truth depth, surface normals, occlusion, and optical flow were rendered for each frame. Sample dataset frames are displayed in Figure 8. For each video sequence, a coverage map was generated by accumulating the surface faces observed in each frame throughout the video, then marking those faces which went unobserved. A sample coverage map for a video sequence from the dataset is shown in Figure 9. For every frame in the video dataset, we provide a corresponding:

- Depth frame: Depth along the camera frame's *z*-axis, clamped from 0–100 millimeters. Values are linearly scaled and encoded as a 16-bit grayscale image.

- Surface normal frame: Reported with respect to the camera coordinate system. X/Y/Z components are stored in separate R/G/B color channels. Components are linearly scaled from ±1 to 0–65535. Values are encoded as a 16-bit color image.

- Optical flow frame: Computed flowing from the current frame to the previous frame, meaning the first frame in the sequence has no value: $I_{i-1} = I_i(u + \Delta u, v + \Delta v)$. Values are saved in a color image, where the R-channel contains X-direction motion (left $\rightarrow$ right, −20 to 20 pixels), and the G-channel contains Y-direction motion (up $\rightarrow$ down, −20 to 20 pixels). Values are linearly scaled from 0 to 65535 and encoded as a 16-bit color image.

- Occlusion frame: Encoded as an 8-bit binary image. Pixels occluding other mesh faces within 100 millimeters of the camera origin are assigned a value of 255, and all other pixels are assigned a value of 0.

- Camera pose: Saved in homogeneous form.

For each video sequence, we also provide:

- 3D model: Ground truth triangulated mesh, stored as a Wavefront OBJ file.

- Coverage map: Binary texture indicating which mesh faces were observed by the camera during the video sequence (1=observed, 2=unobserved).

Finally, we include four video sequences of simulated screening procedure, consisting of 20,058 frames, and ground truth camera poses in homogeneous form.

5. Discussion and conclusion

This work presents a 2D-3D registration method for acquiring and registering phantom colonoscopy video data with ground truth surface models. Unlike traditional 2D-3D registration techniques that register single views, this method operates on video sequences and measured pose. Our results demonstrate that leveraging this temporal information improves the registration accuracy compared to registering a single frame (Figure 6). To circumvent registration errors caused by specular reflections and surface textures, we transformed keyframes to a depth domain for similarity evaluation. The transformed depth frames show scale inconsistency (Figure 7), as is common in GAN-predicted depth. We found the impact of this inconsistency to be reduced by aligning edge features extracted from the depth frames, further improving the registration accuracy (Table 4). Additionally, our loss function outperforms other metrics such as NCC and DICE, while also being computationally cheaper.

We also found that the registration accuracy for moderate and complex trajectories outperformed simple trajectories. One possible explanation is an increase in the number of unique edge features used for alignment and their increased translation in the imaging plane in moderate and complex trajectories. Prior to registration optimization, the input sequence is sampled to generate a set of keyframes, and this number of keyframes is constant (K=5)

for all trajectories. While simple trajectories included only linear translation primarily in the colon major axis direction, moderate and complex trajectories also included lateral translation as well as pitch and yaw variation. The camera poses in complex trajectories are more diverse than simple trajectories, which results in a larger diversity of edge features for alignment. Additionally, simple linear translation of the camera parallel to the major colon axis results in much smaller displacements in image features compared to camera translation perpendicular to the colon axis.

This registration method was used to generate C3VD, the first colonoscopy reconstruction dataset that includes real colonoscopy videos labeled with registered ground truth. Unlike the EndoSLAM dataset (Ozyoruk et al., 2021), C3VD is recorded entirely with an HD clinical colonoscope and includes depth, surface normal, occlusion, and optical flow frame labels. While EndoSLAM uses real porcine tissue mounted on nontubular scaffolds, we opt for tubular phantom models to simulate a geometrically realistic lumen. By recording with a real colonoscope, C3VD overcomes the limited ability of renderers to simulate non-global illumination, light scattering, and nonlinear post-processing (Rau et al., 2022). Compared to other datasets which omit large portions of the angular FoV, C3VD is the first dataset to model the entire colonoscope FoV through an omnidirectional camera model.

C3VD can be used to validate the sub-components of SLAM reconstruction algorithms, including estimating pixel-level depth, surface normals, optical flow, and occluded regions. The ground truth surface models and coverage maps can be used to evaluate entire SLAM reconstruction methods and techniques for identifying missed regions. Additionally, the 3D model assets are open-sourced and offer a higher-resolution alternative to CT colonography volumes for rendering synthetic training data. Mold files and fabrication protocols are also provided so that other researchers may produce and modify phantom colon models. The simulated screening colonoscopy videos may be used to test algorithms on full sequences with realistic scope motion.

The proposed methods have several important limitations. Precise pose measurements are required to generate the registered video sequences, and these trajectories are recorded by mounting the colonoscope to a robotic arm. While this method enables several additional forms of ground truth information, the scope range and types of motion are limited. Additionally, the phantom models must remain static during video acquisition, whereas colonic tissue is flexible, dynamic, and in frequent contact with the scope during imaging. Lastly, small errors in the handeye calibration accumulate, causing a gradual drift in the camera trajectory. Future work could improve upon these limitations with adjustments to the data acquisition protocol and registration algorithm. A more-accurate radiofrequency positional sensor could circumvent trajectory limitations imposed by the robotic arm. Handeye calibration errors could possibly be reduced by co-optimizing both the model transformation and the handeye transformation. This method could also be paired with dynamic CT, similar to Stoyanov et al. (2010), to produce video data with deformable colon models.

The C3VD dataset presented here also has important limitations in its diversity and scope. Future work could expand this dataset to encompass a larger variety of endoscopic imaging

and tissue conditions. For quality metric applications, such as fractional coverage, additional trajectories and colon shapes with a variety of missed regions should be developed. Among the challenges to applying 3D reconstruction algorithms to colonoscopy are loose stool and debris that often occlude the field of view. To simulate these artifacts, artificial stool could be applied to the silicone models to span varying levels of bowel preparation quality. The auxiliary water jet and suction systems of the colonoscope could be used to simulate bowel cleaning and underwater imaging in the videos. Similarly, the air and water nozzles may be used to remove debris from the objective lens. Future datasets may also benefit from including alternative endoscope systems and modalities, such as narrow band imaging (NBI) and chromoendoscopy. The use of distal attachments such as EndoCuff (Rex et al., 2018) and the effect on observational coverage could also be explored. Finally, we also envision generating ground truth data for evaluating structured illumination Parot et al. (2013) and hyperspectral methods (Yoon et al., 2021). Validation data for these methods could include phantoms with tuned optical scattering and absorption properties to realistically simulate light-tissue interactions (Ayers et al., 2008; Chen et al., 2019a; Sweer et al., 2019).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Abdelrahim M, Saiga H, Maeda N, Hossain E, Ikeda H, Bhandari P, 2022. Automated sizing of colorectal polyps using computer vision. Gut 71, 7–9. URL: https://gut.bmj.com/content/71/1/7, doi:10.1136/gutjnl-2021-324510,arXiv:https://gut.bmj.com/content/71/1/7.full.pdf. [PubMed: 34266967]

Ali H, Sharif M, Yasmin M, Rehmani MH, Riaz F, 2020. A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract. Artificial Intelligence Review 53, 2635–2707.

Allan M, Mcleod J, Wang C, Rosenthal JC, Hu Z, Gard N, Eisert P, Fu KX, Zeffiro T, Xia W, et al. , 2021. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133

Armin MA, Chetty G, De Visser H, Dumas C, Grimpen F, Salvado O, 2016. Automated visibility map of the internal colon surface from colonoscopy video. International journal of computer assisted radiology and surgery 11,1599–1610. [PubMed: 27492067]

Ayers F, Grant A, Kuo D, Cuccia DJ, Durkin AJ, 2008. Fabrication and characterization of silicone-based tissue phantoms with tunable optical properties in the visible and near infrared domain, in: Design and Performance Validation of Phantoms Used in Conjunction with Optical Measurements of Tissue, SPIE. pp. 56–64.

Birkfellner W, Wirth J, Burgstaller W, Baumann B, Staedele H, Hammer B, Gellrich NC, Jacob AL, Regazzoni P, Messmer P, 2003. A faster method for 3d/2d medical image registration-a simulation study. Physics in Medicine and Biology 48, 2665–2679. URL: 10.1088/0031-9155/48/16/307, doi:10.1088/0031–9155/48/16/307. [PubMed: 12974581]

Chadebecq F, Lovat LB, Stoyanov D, 2023. Artificial intelligence and automation in endoscopy and surgery. Nature Reviews Gastroenterology & Hepatology 20, 171–182. [PubMed: 36352158]

Chen MT, Mahmood F, Sweer JA, Durr NJ, 2019a. Ganpop: generative adversarial network prediction of optical properties from single snapshot wide-field images. IEEE transactions on medical imaging 39, 1988–1999. [PubMed: 31899416]

Chen R, Mahmood F, Yuille A, Durr NJ, 2018. Rethinking monocular depth estimation with adversarial training. arXiv preprint arXiv:1808.07528.

Chen RJ, Bobrow TL, Athey T, Mahmood F, Durr NJ, 2019 b. Slam endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv: 1907.00283.

Cheng K, Ma Y, Sun B, Li Y, Chen X, 2021. Depth estimation for colonoscopy images with self-supervised learning from videos, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 119–128.

Cunningham D, Atkin W, Lenz HJ, Lynch HT, Minsky B, Nordlinger B, Starling N, 2010. Colorectal cancer. The Lancet 375, 1030–1047. URL: https://www.sciencedirect.com/science/article/pii/S0140673610603534, doi:https://doi.org/10.1016/S0140-6736(10)60353-4.

De Silva T, Uneri A, Ketcha M, Reaungamornrat S, Kleinszig G, Vogt S, Aygun N, Lo S, Wolinsky J, Siewerdsen J, 2016. 3d-2d image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. Physics in Medicine & Biology 61, 3009. [PubMed: 26992245]

Edwards PE, Psychogyios D, Speidel S, Maier-Hein L, Stoyanov D, 2022. Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. Medical Image Analysis 76, 102302. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003479, doi:https://doi.org/10.1016/j.media.2021.102302 [PubMed: 34906918]

Freedman D, Blau Y, Katzir L, Aides A, Shimshoni I, Veikherman D, Golany T, Gordon A, Corrado G, Matias Y, et al. , 2020. Detecting deficient coverage in colonoscopies. IEEE Transactions on Medical Imaging 39,3451–3462. [PubMed: 32746092]

Fu Z, Jin Z, Zhang C, He Z, Zha Z, Hu C, Gan T, Yan Q, Wang P, Ye X, 2021. The future of endoscopic navigation: A review of advanced endoscopic vision technology. IEEE Access 9, 41144–41167.

Fulton MJ, Prendergast JM, DiTommaso ER, Rentschler ME, 2020. Comparing visual odometry systems in actively deforming simulated colon environments, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 4988–4995.

Groher M, Jakobs TF, Padoy N, Navab N, 2007. Planning and intraoperative visualization of liver catheterizations: New cta protocol and 2d-3d registration method. Academic Radiology 14, 13251340. URL: https://www.sciencedirect.com/science/article/pii/S1076633207003996, doi:https://doi.org/10.1016/j.acra.2007.07.009.

Grupp RB, Unberath M, Gao C, Hegeman RA, Murphy RJ, Alexander CP, Otake Y, McArthur BA, Armand M, Taylor RH, 2020. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2d/3d registration. International journal of computer assisted radiology and surgery 15, 759–769. [PubMed: 32333361]

Hansen N, Müller SD, Koumoutsakos P, 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). Evolutionary Computation 11, 1–18. doi:10.1162/106365603321828970. [PubMed: 12804094]

Hassan C, Wallace MB, Sharma P, Maselli R, Craviotto V, Spadaccini M, Repici A, 2020. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut 69, 799–800. URL: https://gut.bmj.com/content/69/5/799, doi:10.1136/gutjnl-2019-319914,arXiv:https://gut.bmj.com/content/69/5/799.full.pdf [PubMed: 31615835]

Hong D, Tavanapong W, Wong J, Oh J, De Groen PC, 2011. Colon fold contour estimation for 3d visualization of colon structure from 2d colonoscopy images, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE. pp. 121–124.

Hong W, Wang J, Qiu F, Kaufman A, Anderson J, 2007. Colonoscopy simulation, in: Medical Imaging 2007: Physiology, Function, and Structure from Medical Images, SPIE. pp. 212–219.

Livovsky DM, Veikherman D, Golany T, Aides A, Dashinsky V, Rabani N, Shimol DB, Blau Y, Katzir L, Shimshoni I, et al. , 2021. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). Gastrointestinal Endoscopy 94, 1099–1109. [PubMed: 34216598]

Livyatan H, Yaniv Z, Joskowicz L, 2003. Gradient-based 2-d/3-d rigid registration of fluoroscopic x-ray to ct. IEEE transactions on medical imaging 22, 1395–1406. [PubMed: 14606673]

Luo Y, Zhang Y, Liu M, Lai Y, Liu P, Wang Z, Xing T, Huang Y, Li Y, Li A, et al. , 2021. Artificial intelligence-assisted colonoscopy for detection of colon polyps: a prospective, randomized cohort study. Journal of Gastrointestinal Surgery 25, 2011–2018. [PubMed: 32968933]

Ma R, Wang R, Pizer S, Rosenman J, McGill SK, Frahm JM, 2019 Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 573–582.

Mahmood F, Chen R, Sudarsky S, Yu D, Durr NJ, 2018a. Deep learning with cinematic rendering: fine-tuning deep neural networks using photorealistic medical images. Physics in Medicine & Biology 63, 185012. [PubMed: 30113015]

Mahmood F, Durr NJ, 2018. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical Image Analysis 48, 230243. URL: https://www.sciencedirect.com/science/article/pii/S1361841518303761, doi:

Mahmood F, Yang Z, Ashley T, Durr NJ, 2018b. Multimodal densenet. arXiv preprint arXiv:1811.07407.

Maier-Hein L, Groch A, Bartoli A, Bodenstedt S, Boissonnat G, Chang PL, Clancy N, Elson DS, Haase S, Heim E, et al. , 2014. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. IEEE transactions on medical imaging 33, 1913–1930. [PubMed: 24876109]

Markelj P, Tomaževi D, Likar B, Pernuš F, 2012. A review of 3d/2d registration methods for image-guided interventions. Medical Image Analysis 16, 642–661. URL: https://www.sciencedirect.com/science/article/pii/S1361841510000368, doi: [PubMed: 20452269]

McGill SK, Rosenman J, Zhao Q, Wang R, Ma R, Fan M, Niethammer M, Alterovitz R, Frahm JM, Tepper J, Pizer S, 2018. Missed colonic surface area at colonoscopy can be calculated with computerized 3d reconstruction. Gastrointestinal Endoscopy 87, AB254. URL: https://www.sciencedirect.com/science/article/pii/S0016510718307272, doi:

Oulbacha R, Kadoury S, 2020. Mri to c-arm spine registration through pseudo-3d cyclegans with differentiable histograms. Medical Physics 47, 6319–6333. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14534, doi: [PubMed: 33063353]

Ozyoruk KB, Gokceler GI, Bobrow TL, Coskun G, Incetan K, Almalioglu Y, Mahmood F, Curto E, Perdigoto L, Oliveira M, Sahin H, Araujo H, Alexandrino H, Durr NJ, Gilbert HB, Turan M, 2021. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis 71, 102058. URL: https://www.sciencedirect.com/science/article/pii/S1361841521001043, doi:10.1016/j.media. 2021.102058. [PubMed: 33930829]

Park FC, Martin BJ, 1994. Robot sensor calibration: solving ax=xb on the euclidean group. IEEE Transactions on Robotics and Automation 10, 717–721.

Parker SG, Bigler J, Dietrich A, Friedrich H, Hoberock J, Luebke D, McAllister D, McGuire M, Morley K, Robison A, et al. , 2010. Optix: a general purpose ray tracing engine. Acm transactions on graphics (tog) 29, 1–13.

Parot V, Lim D, González G, Traverso G, Nishioka NS, Vakoc BJ, Durr NJ, 2013. Photometric stereo endoscopy. Journal of biomedical optics 18, 076017. [PubMed: 23864015]

Penza V, Ciullo AS, Moccia S, Mattos LS, De Momi E, 2018. Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. The International Journal of Medical Robotics and Computer Assisted Surgery 14, e1926. [PubMed: 29968295]

Pratt P, Stoyanov D, Visentini-Scarzanella M, Yang GZ, 2010. Dynamic guidance for robotic surgery using image-constrained biomechanical models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 77–85.

Rau A, Bhattarai B, Agapito L, Stoyanov D, 2022. Bimodal camera pose prediction for endoscopy. arXiv preprint arXiv:2204.04968

Rau A, Edwards PE, Ahmad OF, Riordan P, Janatka M, Lovat LB, Stoyanov D, 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International journal of computer assisted radiology and surgery 14, 1167–1176. [PubMed: 30989505]

Recasens D, Lamarca J, Fácil JM, Montiel J, Civera J, 2021. Endodepth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters 6, 7225–7232.

Rex DK, Repici A, Gross SA, Hassan C, Ponugoti PL, Garcia JR, Broadley HM, Thygesen JC, Sullivan AW, Tippins WW, et al. , 2018. High-definition colonoscopy versus endocuff versus endorings versus fullspectrum endoscopy for adenoma detection at colonoscopy: a multicenter randomized trial. Gastrointestinal endoscopy 88, 335–344. [PubMed: 29530353]

Rex DK, Schoenfeld PS, Cohen J, Pike IM, Adler DG, Fennerty MB, Lieb JG, Park WG, Rizk MK, Sawhney MS, et al. , 2015. Quality indicators for colonoscopy. Gastrointestinal endoscopy 81, 31–53. [PubMed: 25480100]

van Rijn JC, Reitsma JB, Stoker J, Bossuyt PMM, Deventer SJH, Dekker E, 2006. Polyp miss rate determined by tandem colonoscopy: A systematic review. The American Journal of Gastroenterology 101, 343–350. [PubMed: 16454841]

Samadder NJ, Curtin K, Tuohy TM, Pappas L, Boucher K, Provenzale D, Rowe KG, Mineau GP, Smith K, Pimentel R, Kirchhoff AC, Burt RW, 2014. Characteristics of missed or interval colorectal cancer and patient survival: A population-based study. Gastroenterology 146, 950–960. URL: https://www.sciencedirect.com/science/article/pii/S0016508514000262, doi: [PubMed: 24417818]

Scaramuzza D, Martinelli A, Siegwart R, 2006. A toolbox for easily calibrating omnidirectional cameras, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 5695–5701.

Siegel RL, Miller KD, Fuchs HE, Jemal A, 2021. Cancer statistics, 2021. CA: A Cancer Journal for Clinicians 71, 7–33. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21654, doi: [PubMed: 33433946]

Stoyanov D, Scarzanella MV, Pratt P, Yang GZ, 2010. Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 275–282.

Sweer JA, Chen MT, Salimian KJ, Battafarano RJ, Durr NJ, 2019. Wide-field optical property mapping and structured light imaging of the esophagus with spatial frequency domain imaging. Journal of biophotonics 12, e201900005. [PubMed: 31056845]

Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B, 2018. Highresolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

Yao H, Stidham RW, Gao Z, Gryak J, Najarian K, 2021. Motion-based camera localization system in colonoscopy videos. Medical Image Analysis 73, 102180. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002267, doi: [PubMed: 34303888]

Ye M, Johns E, Handa A, Zhang L, Pratt P, Yang GZ, 2017. Selfsupervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv: 1705.08260.

Yoon J, Joseph J, Waterhouse DJ, Borzy C, Siemens K, Diamond S Tsikitis VL, Bohndiek SE, 2021. First experience in clinical application of hyperspectral endoscopy for evaluation of colonic polyps. Journal of Biophotonics 14, e202100078. [PubMed: 34047490]

Zhang S, Zhao L, Huang S, Ye M, Hao Q, 2021. A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images. IEEE Transactions on Medical Robotics and Bionics 3, 85–95. doi:10.1109/TMRB. 2020.3044108.

Highlights

- 2D-3D registration that leverages temporal information in optical colonoscopy videos

- conditional Generative Adversarial Network used to transform frames to depth domain

- High-fidelity silicone phantom models with 3D ground truth

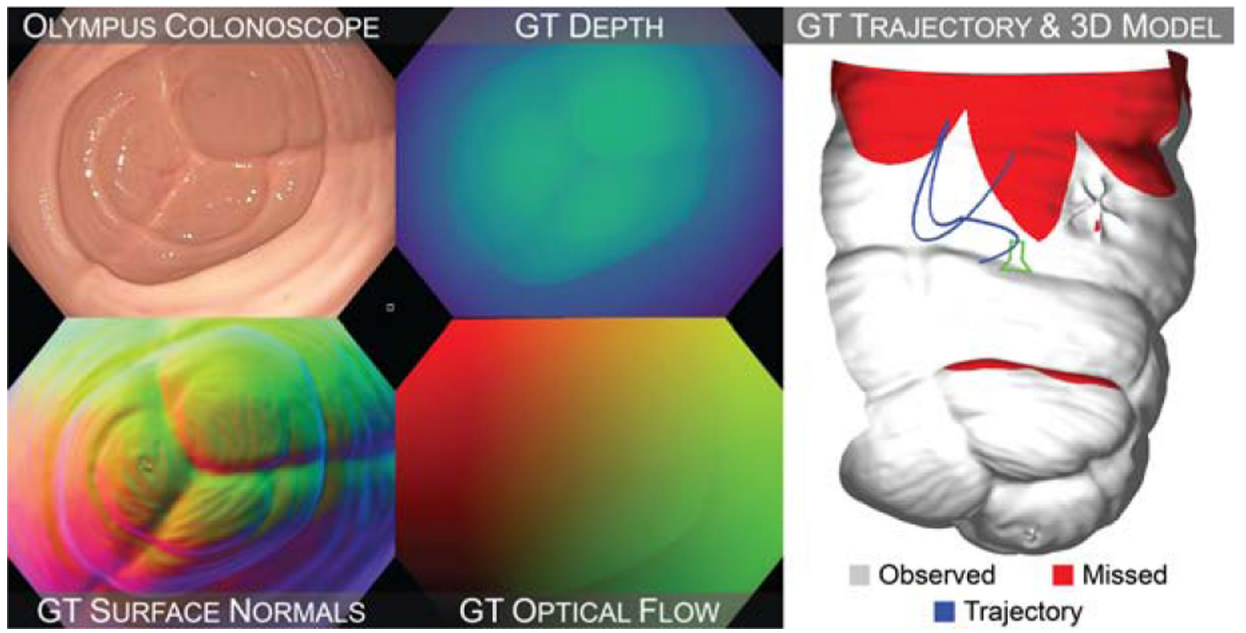- Public dataset for quantitative benchmarking of computer vision tasks in colonoscopy

Fig. 1: Sample frame from the proposed dataset.
Real colonoscope frames are paired with registered ground truth (GT) depth, surface normals, and optical flow frames (left). Each video is paired with a ground truth camera trajectory, 3D surface model, and coverage map (right).
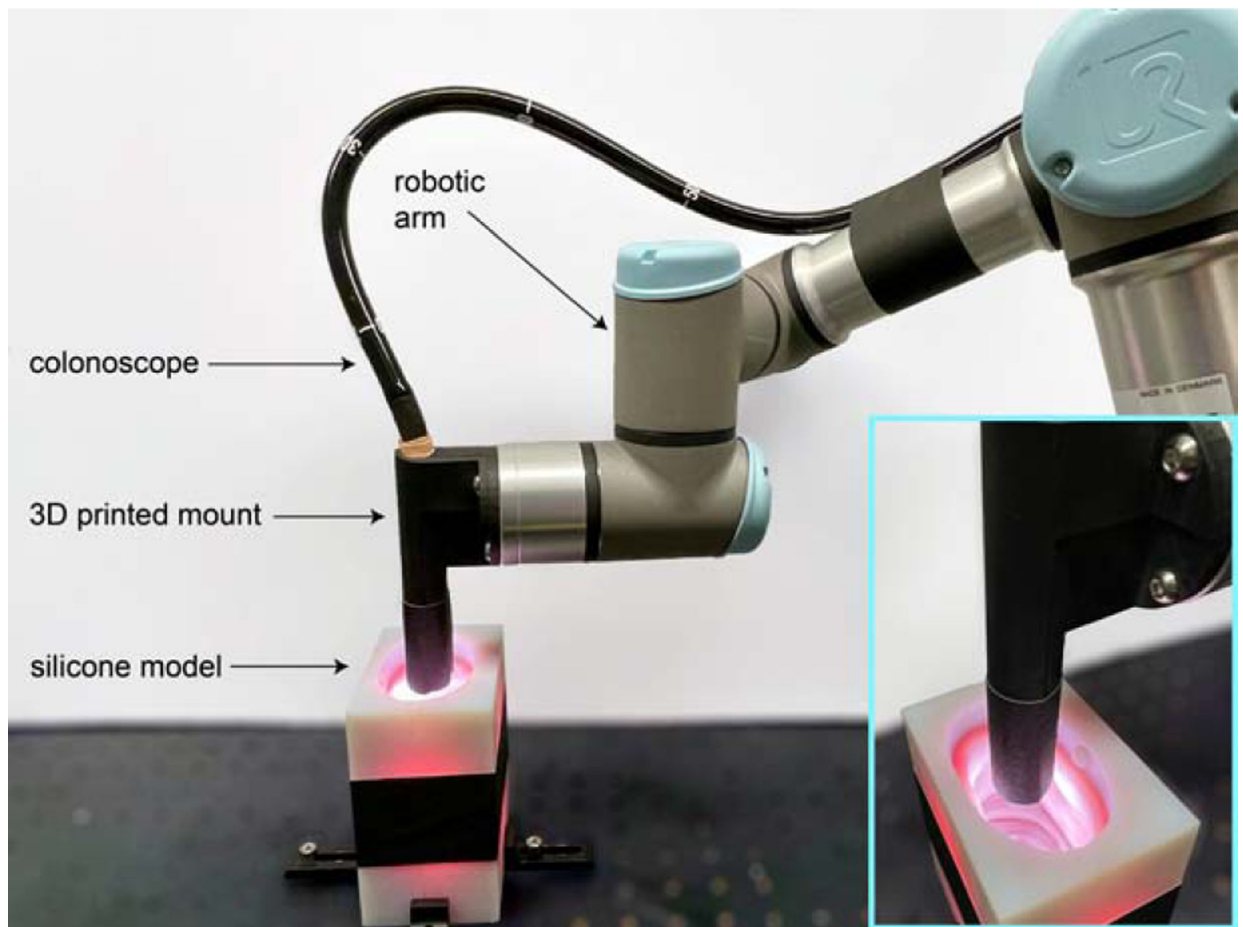
Fig. 2: Data acquisition setup.
A commercial colonoscope is rigidly affixed to a robotic arm using a 3D printed mount. The colonoscope is navigated through the silicone colon with known 3D shape while video and pose are simultaneously recorded.
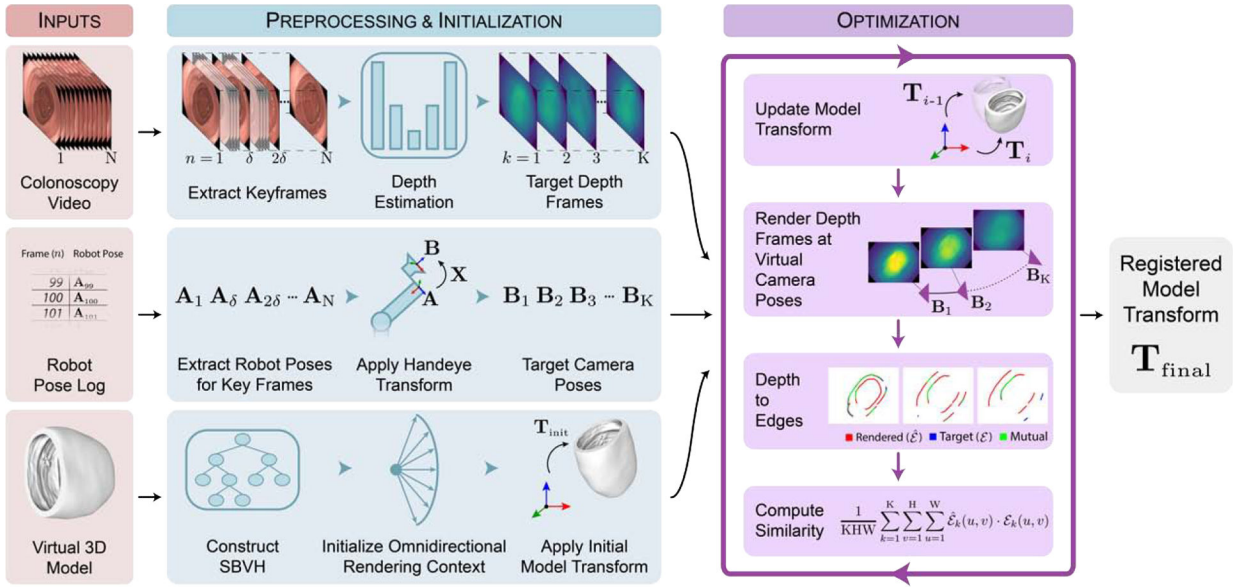
Fig. 3: 2D-3D video registration method.
Temporally synchronized video and pose measurements are sampled to produce a set of keyframes for registration with the ground truth 3D model. Keyframes are individually transformed to a depth domain by a generative model to produce target depth images. A virtual omnidirectional camera is moved to each keyframe pose to render depth images of the 3D model. Rendered and target depth frames are compared while the model transform is updated with an evolutionary optimizer until convergence.
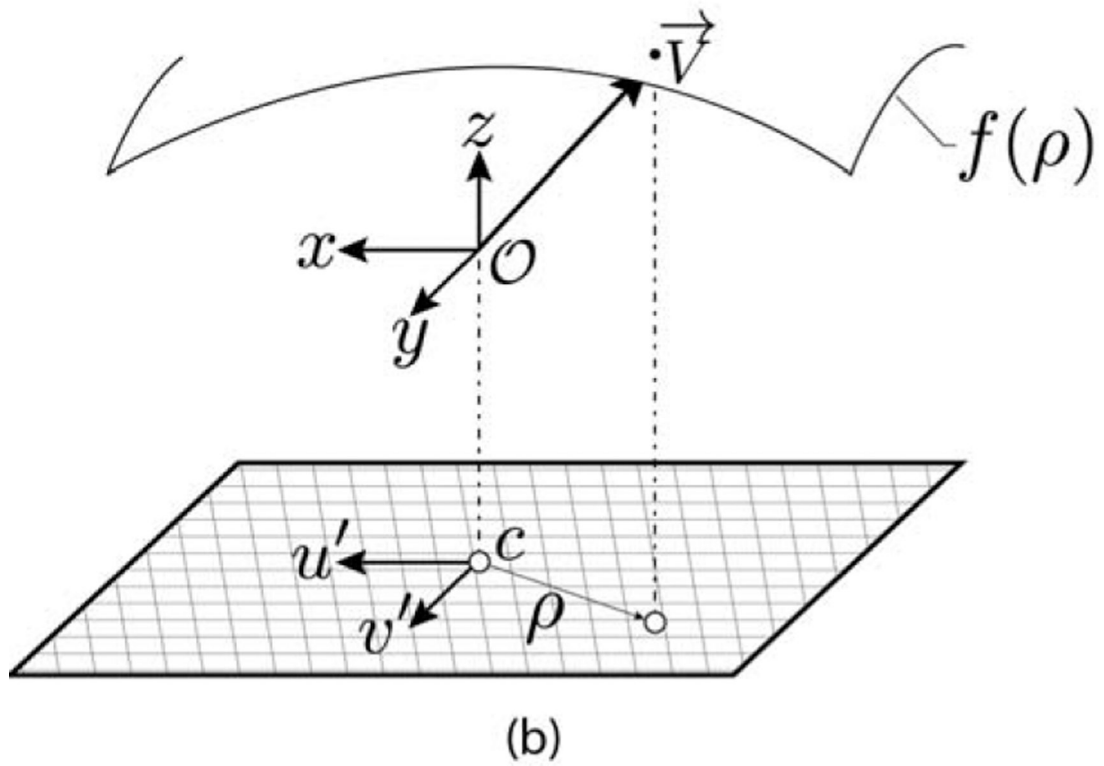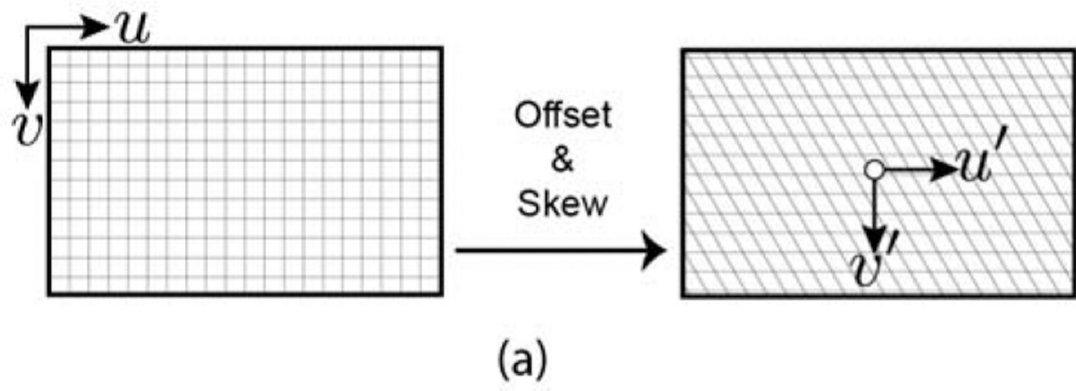
Fig. 4: Omnidirectional camera model.

(a) Pixel coordinates are skewed by an affine transform to model lens misalignment and offset to the optical center. (b) Each pixel is assigned a ray direction defined by a parametric surface that is a function of the pixel's radial distance from the optical center.
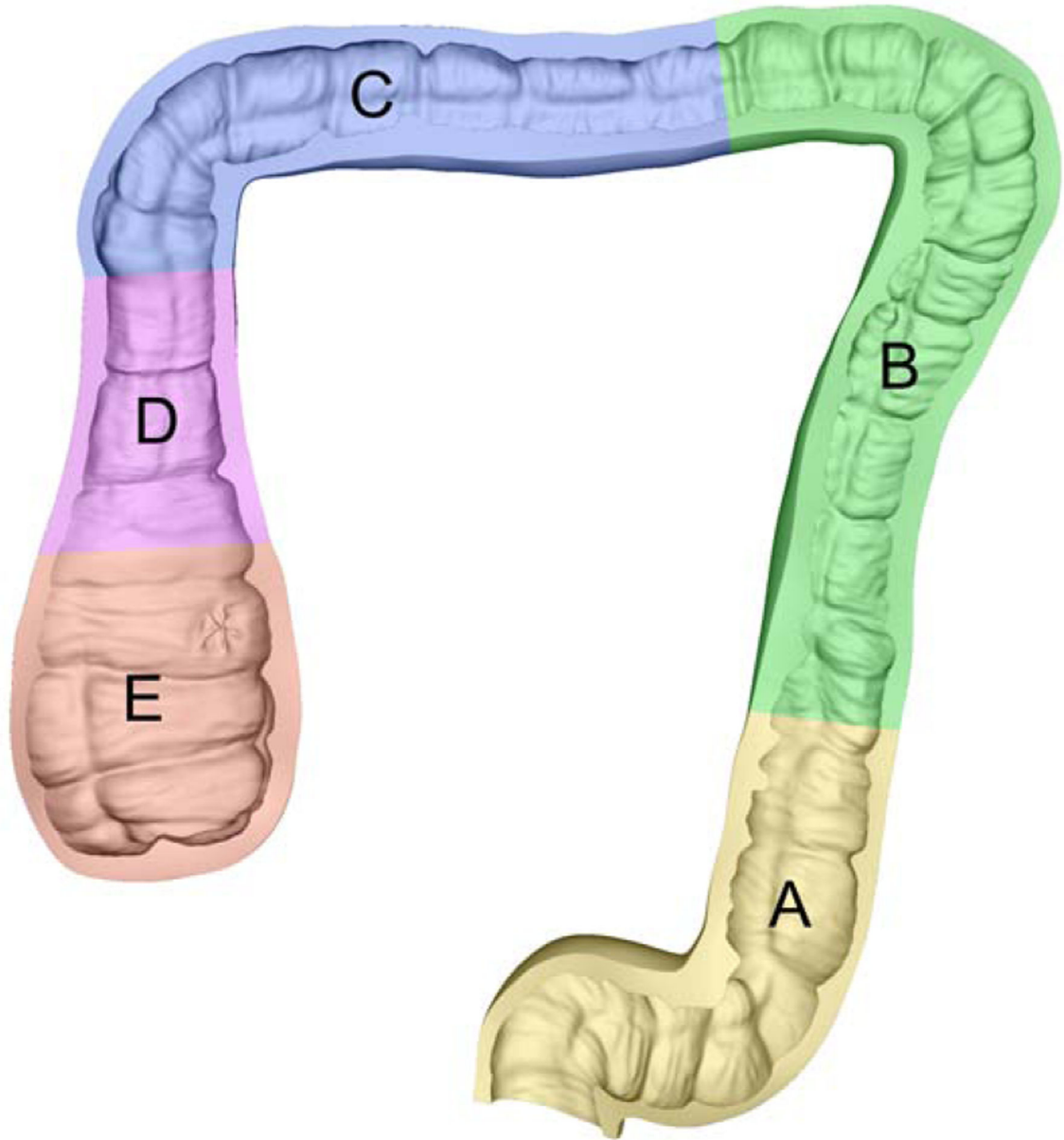
Fig. 5: Cross section of the ground truth 3D colon model.
Sculpting the 3D model using reference anatomical images introduces higher-resolution detail than CT colonography volumes.

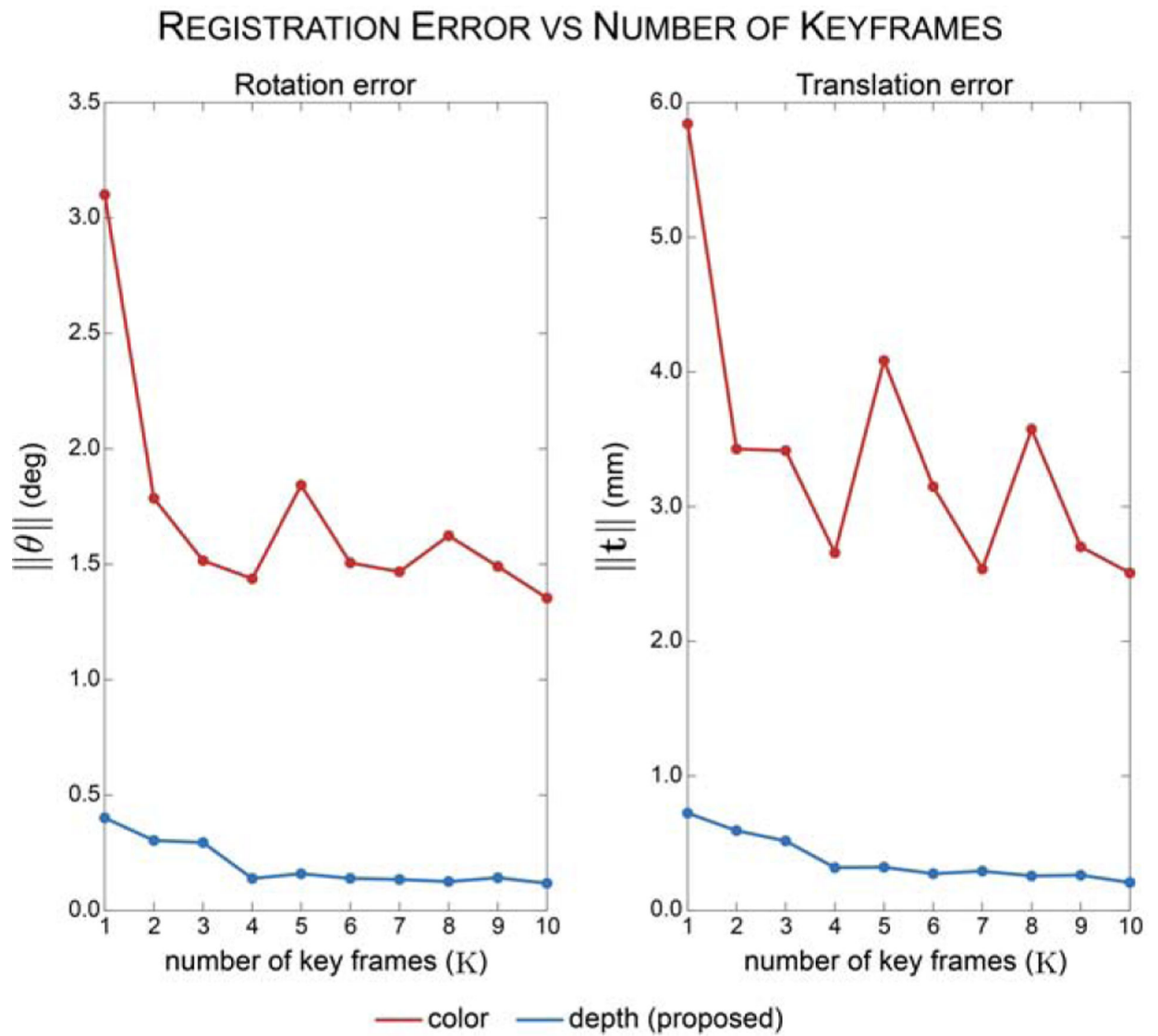# REGISTRATION ERROR VS NUMBER OF KEYFRAMES

Fig. 6: Average registration error for 10 synthetic video sequences with increasing number of keyframes K.

The registration error decreases with increasing number of keyframes. Registering frames in the depth domain (blue line) results in improved registration accuracy compared to registering frames in the color domain.
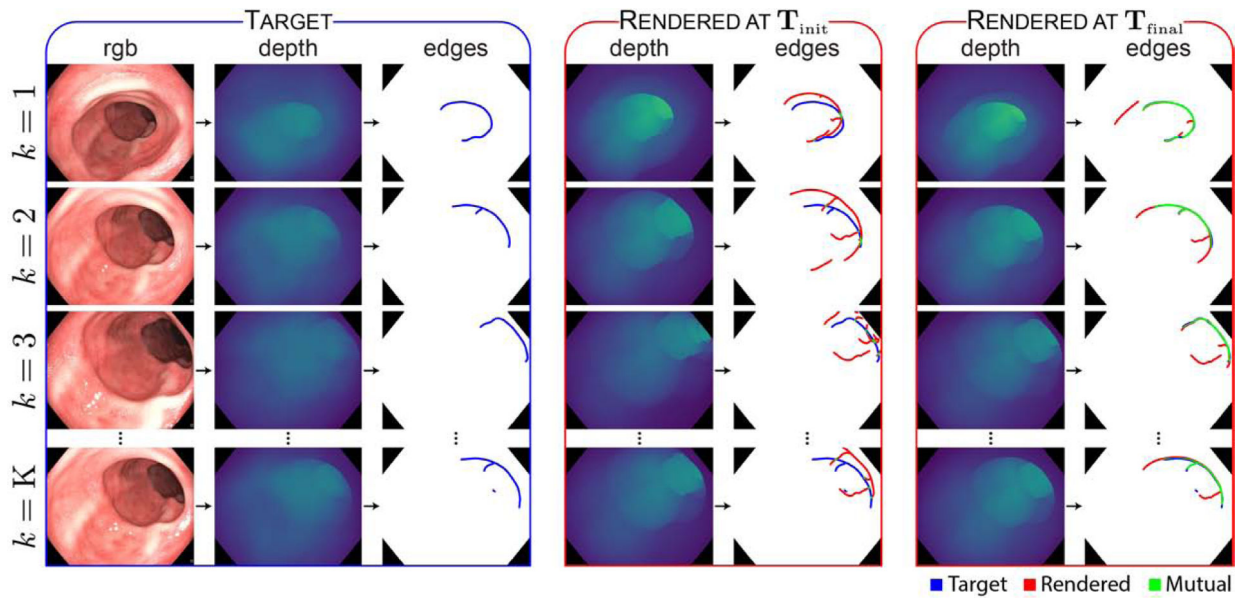
Fig. 7: Qualitative registration results.

RGB video frames are transformed to the depth domain by a generative model, and target edge features are extracted. Edge features from rendered depth frames are compared with the target edge features to compute an alignment loss, and the predicted model transform is iteratively updated until alignment is achieved. Blurred edge features are binarized in this figure for visibility. See Supplementary Video I for a visual depiction of the improvement in registration as the optimization progresses.
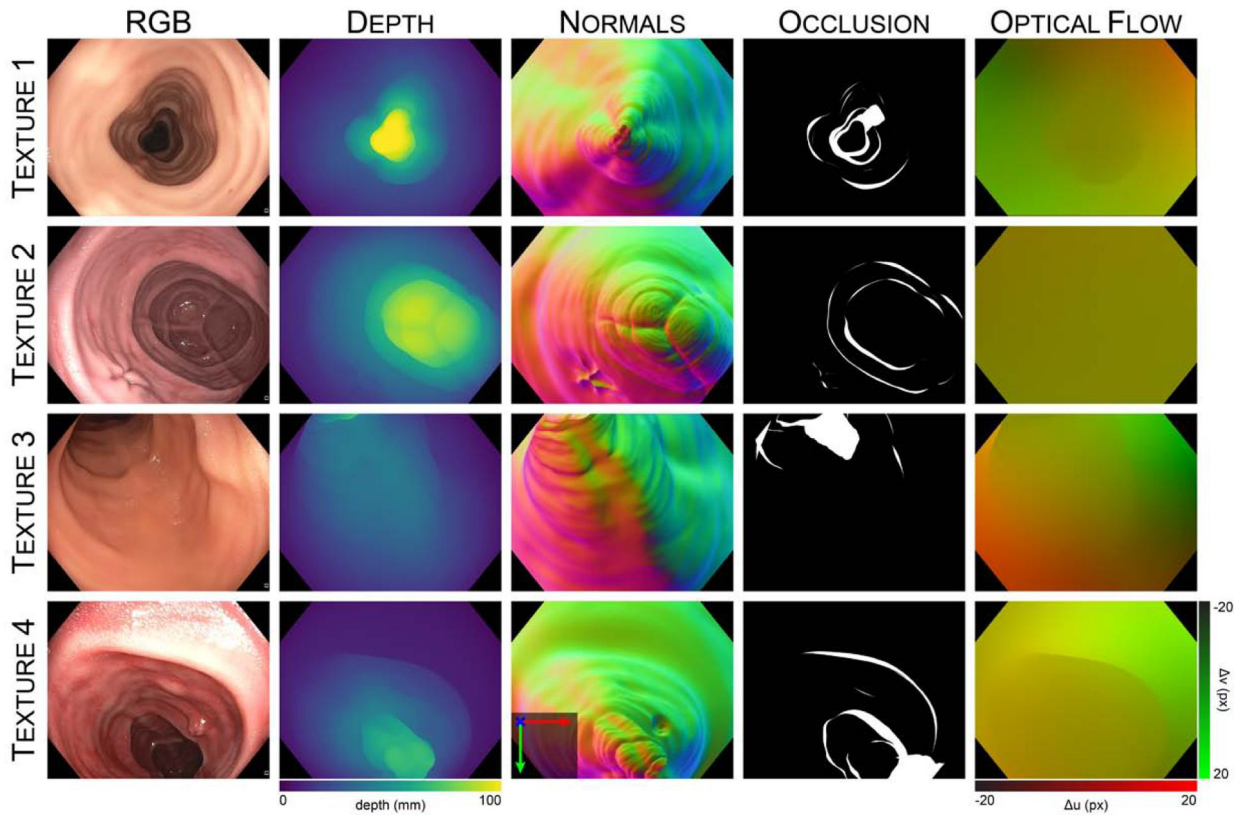
Fig. 8: Sample frames from the C3VD dataset.

Frames are recorded using a clinical colonoscope and silicone phantom models constructed with four unique textures. Each real video frame is paired with ground truth depth, surface normals, occlusion, and optical flow. A sample video sequence is shown in Supplementary Video II.
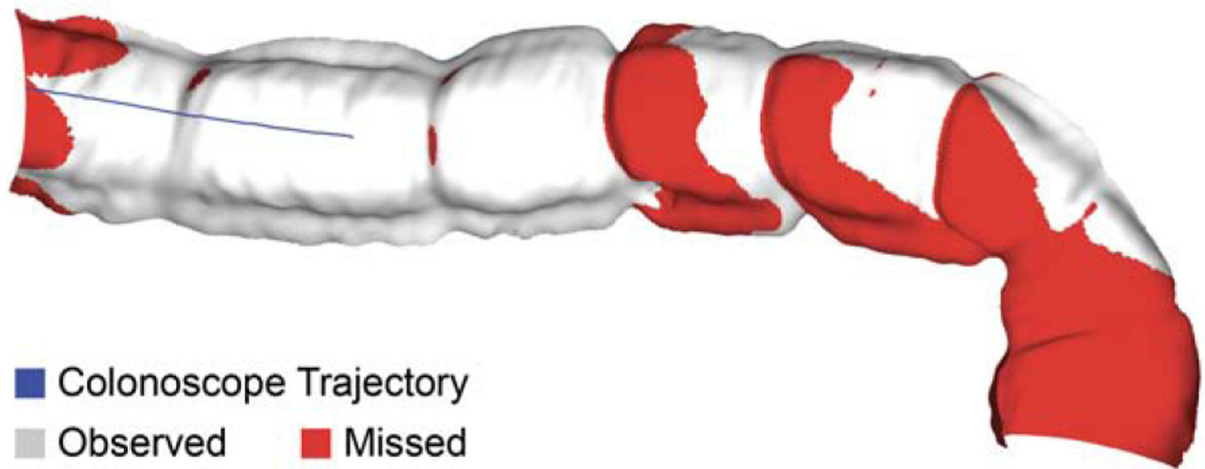
Fig. 9: Sample coverage map for a recorded video sequence.
Unobserved surface regions due to occlusion are marked as red.

Table 1:

Comparison of endoscopy reconstruction datasets

Camera			Tissue Type	3D GT	Registration	Frames	Ground Truth					Dataset
Type	FoV	Res.					Pose	Depth	Normals	Flow	3D Model	
Stereo da Vinci	Narrow	SD	Ex-vivo Porcine	CT	Manual	16	✓				✓	Edwards et al. (2022)
Stereo CMOS	Narrow	SD	Phantom	CT	Calib Plate	120		✓			✓	Penza et al. (2018)
Stereo da Vinci	Narrow	SD	Phantom	CT	Fiducials	5,816					✓	Stoyanov et al. (2010); Pratt et al. (2010)
Stereo da Vinci	Narrow	SD	In-vivo Animal	None	None	20,000						Ye et al. (2017)
Stereo da Vinci	Narrow	SD	In-vivo Animal	None	None	92,672		✓			✓	Recasens et al. (2021)
Stereo da Vinci	Narrow	SD	Ex-vivo Porcine	SI	None	40		✓				Allan et al. (2021) *
Stereo da Vinci	Narrow	SD	Ex-vivo Porcine	CT	Fiducials	131					✓	Maier-Hein et al. (2014)
Stereo Rendered	Narrow	SD	Virtual	CT	None	6,320	✓					Zhang et al. (2021)
Mono USB	Narrow	SD	Phantom	None	None	23,935	✓					Fulton et al. (2020)
Mono Rendered	Narrow	SD	Virtual	CT	None	16,016		✓				Rau et al. (2019)
Mono Rendered	Narrow	SD	Virtual	CT	None	18,000	✓	✓				Rau et al. (2022) *
Mono USB	Wide	SD/HD	Ex-vivo Porcine	OS	None	39,406	✓				✓	Ozyoruk et al. (2021)
Mono Pill Cam	Wide	SD	Ex-vivo Porcine	None	None	3,294	✓					Ozyoruk et al. (2021)
Mono Colonoscope	Wide	HD	Phantom	CT	None	12,250	✓				✓	Ozyoruk et al. (2021)
Mono Rendered	Narrow	SD	Virtual	CT	None	21,887	✓	✓				Ozyoruk et al. (2021)
Mono Colonoscope	*Wide*	*HD*	*Phantom*	*Sculpted*	*Optimized*	*10,015*	✓	✓	✓	✓	✓	*Proposed*

Ground Truth (GT), Computed Tomography (CT), Structured Illumination (SI), Optical Scan (OS)

*
Only available to challenge participants

Table 2:

Ground truth 3D colon model attributes

Segment	Length (mm)	Lesion Type	Major axis (mm)
A	192	Hyperplastic	4.2
		Traditional adenomatous	6.6
B	339	Sessile serrated	6.2
C	241	Sessile serrated	10.0
D	93	-	-
E	97	-	-

Table 3:

Omnidirectional camera intrinsics for Olympus CF-HQ190L video Colonoscope

Size	H×W	1080 × 1350
Optical Center	c_x	679.54
	C_y	543.98
Polynomial	a_0	769.24
	a_2	-8.13×10^{-4}
	a_3	-6.26×10^{-7}
	a_4	-1.20×10^{-9}
Stretch	e	0.9999
	f	2.88×10^{-3}
	g	-2.96×10^{-3}

Table 4:

Effect of loss functions on registration accuracy

Input	Loss	$\|\theta_{\mathbf{terror}}\|$ (deg)	$\|\mathbf{t_{terror}}\|$ (mm)
Depth	L_1	8.008	14.941
	L_2	8.729	16.109
	NCC	2.904	6.014
	GC	0.558	0.843
Edge	L_1	0.466	0.654
	L_2	0.417	0.750
	DICE	0.374	0.756
	NCC	0.450	0.681
	Proposed	**0.159**	**0.321**

Table 5:

C3VD dataset video sequences

Segment	Texture	Video	Frames
A	1	a	700
	2	a	514
	3	a	613
	3	b	536
B	4	a	148
C	1	a	61
	1	b	700
	2	a	194
	2	b	103
	2	c	235
	3	a	250
	3	b	214
	4	a	382
	4	b	597
E	1	a	276
	1	b	765
	2	a	370
	2	b	1,142
	2	c	595
	3	a	730
	3	b	465
	4	a	425
A-E[*]	1	a	5,458
	2	a	5,100
	3	a	4,726
	4	a	4,774

[*] Simulated screening colonoscopies