AI Assisted Colonscopy Navigation

Yimeng Duan
Johns Hopkins University

Abstract—**This paper presents a novel AI-assisted approach for colonoscopy navigation. Our method leverages a binary image classifier to categorize colonoscopy frames into distinct groups based on extracted features. For lumen segmentation, we employ an adapted Med-SAM model (PC-SAM) to generate precise lumen masks, which are then used to determine the centroid of the optimal insertion path. The proposed framework is evaluated using Intersection-over-Union (IoU) and Dice Similarity Coefficient (DSC) metrics, demonstrating superior performance compared to established segmentation models, including U-Net, U-Net++, and TransUNet. Future work will focus on enhancing model robustness by incorporating larger and more diverse training datasets.**

I. Introduction

Colorectal cancer (CRC) is one of the most prevalent and deadly cancers worldwide, with early detection critical for improving patient outcomes. Colonoscopy enables both early tumor detection and precise examination of structural abnormalities. By identifying lesions, polyps, and malignancies at early stages, colonoscopy significantly reduces mortality rates.However, the procedure's success relies on the endoscopist's ability to navigate the colon's complex anatomy, which is often obscured by collapsed lumen segments, fluid retention, and tortuous pathways. Poor visibility and difficulty in determining the optimal path prolong examinations, increase patient discomfort, and elevate procedural risks—particularly for less experienced clinicians. These challenges highlight the need for intelligent navigation aids to improve safety and consistency.A key requirement for effective colonoscopy navigation is real-time identification of the camera's insertion direction, which aligns with the lumen's center. Existing lumen detection methods rely on traditional image-processing techniques, such as dark region detection or haustral fold analysis, but struggle with real-world challenges like poor lighting, specular reflections, and fluid artifacts. While deep learning models like MedSAM show promise in medical segmentation, they lack specialized training for colonoscopy and suffer from high computational costs.

To address these limitations, we propose an AI-assisted navigation system that combines adaptive computer vision with efficient deep learning. Our approach first classifies input frames into two groups: (1) images with a complete O-ring lumen and (2) images where the lumen is obscured, redirecting focus to dark regions. This categorization optimizes feature extraction, which is then processed by PC-SAM—an enhanced, prompt-free adaptation of MedSAM for lumen segmentation. Further refinement via Low-Rank Adaptation (LoRA) reduces computational overhead while improving domain-specific performance. By integrating traditional and deep learning techniques, our system delivers robust navigation guidance even in challenging conditions.

This paper is presented as the followng: Section 2 reviews related work in colonoscopy navigation and medical image segmentation. Section 3 details our methodology, including the binary classifier, PC-SAM architecture, and LoRA refinement. Section 4 presents experimental results and comparisons with baseline models. Finally, Section 5 discusses limitations, clinical implications, and future directions.

II. Related Works

Lumen segmentation techniques for colonoscopy navigation have been extensively studied around the globe. Below we summarize the key research directions.

A. Optical Based Dark Region Detection

Dark region detection is a classical colonoscopy navigation method that identifies the darkest image areas as potential lumen centers. This approach is based on the optical principle that deeper anatomical structures appear darker under endoscopic illumination. The colon lumen naturally forms a dark, hollow space, while surrounding tissues reflect more light, making low-intensity pixel regions reliable indicators of either the central lumen or depth variations in the colon's 3D structure.

The method basically include 3 essential stages: (1) Preprocessing to normalize illumination variations, (2) Adaptive thresholding (using Otsu's method or local contrast enhancement) to segment dark regions, and (3) Morphological processing (erosion/dilation) to refine shape boundaries. These computationally efficient techniques have been extensively studied, with advanced implementations incorporating shape-from-shading or structured light for enhanced 3D depth estimation.

The fundamental limitation lies in assuming the darkest region always represents the optimal navigation path. While valid in straight segments like the descending colon, this fails when lighting artifacts, focal distance changes, or tissue geometry distort intensity-depth relationships. Crucially, the method ignores endoscope dimensions and anatomical constraints—forcing alignment with the darkest point risks tangential approaches to curved walls. True navigation requires parallel lumen wall alignment, achievable only in straight segments through dark-region methods, highlighting the need for hybrid solutions in complex anatomies.

B. Haustral Folds Detection

Traditional haustral fold-based navigation methods utilize the colon's crescent-shaped ridges as anatomical landmarks, leveraging their characteristic radial convergence toward the lumen center. These approaches typically follow a three-stage pipeline.

Initial preprocessing commonly applies contrast-limited adaptive histogram equalization (CLAHE) to enhance tissue contrast while suppressing imaging artifacts. For edge detection, several classical computer vision techniques have been employed:

The Canny edge detector provides gradient-based boundary identification through its multi-stage algorithm involving Gaussian smoothing, non-maximum suppression, and hysteresis thresholding. Alternatively, the Hough transform approach detects parametric shapes by converting image features into parameter space representations. Separately, adaptive thresholding methods, including mean-based local thresholding, offer solutions for fold detection in varying illumination conditions by computing spatially-variant intensity thresholds.

Morphological processing typically follows edge detection, employing dilation to bridge discontinuities and erosion to eliminate noise. Size-based filtering then removes non-physiological features according to anatomical constraints.

These conventional image processing techniques demonstrate particular effectiveness in well-defined anatomical regions where haustral folds maintain regular patterns and visibility. Their performance characteristics vary according to implementation choices and specific colonoscopic conditions.

C. Deep Learning in Medical Segmentation

The evolution of medical image segmentation has been profoundly shaped by advances in deep learning architectures. Early approaches predominantly utilized convolutional neural networks (CNNs), with models like U-Net establishing the effectiveness of encoder-decoder structures with skip connections for preserving spatial information. These CNN-based methods demonstrated strong performance in segmenting well-defined anatomical structures but faced limitations in handling complex, variable morphology due to their local receptive fields and dependence on large annotated datasets.

The field advanced significantly with the introduction of vision transformers (ViTs), which incorporated self-attention mechanisms to model long-range dependencies in medical images. Hybrid architectures such as TransUNet successfully combined the spatial precision of CNNs with the global context modeling of transformers, achieving state-of-the-art results across various segmentation tasks. However, these models still required extensive task-specific fine-tuning and substantial computational resources.

A transformative development came with the Segment Anything Model (SAM), a foundation model trained on an unprecedented scale of diverse images. SAM introduced a promptable architecture capable of zero-shot generalization to new segmentation tasks without domain-specific retraining. Its success led to MedSAM, a specialized adaptation for medical imaging that maintained SAM's flexible prompting while optimizing performance on biomedical data through targeted fine-tuning. These models offer several advantages including flexible interaction through various prompt types, strong generalization across imaging modalities, and efficient adaptation through their decoupled architecture of heavy image encoders and lightweight mask decoders.

In clinical practice, SAM and MedSAM have shown promising results in tasks ranging from tumor delineation to endoscopic polyp detection. Their ability to handle diverse anatomical structures makes them particularly valuable in medical imaging. However, when applied to colonoscopy navigation, these models encounter significant challenges. The complex and dynamic nature of colonoscopic environments presents unique difficulties not adequately addressed by general medical segmentation models. The lumen's variable morphology, frequent visual obstructions from fluids or folds, and rapid scene changes require specialized adaptation beyond typical medical imaging scenarios.

Furthermore, these models depend heavily on high-quality labeled datasets that are especially challenging to obtain for colonoscopy procedures. Precise lumen direction annotation demands substantial endoscopic expertise and labor-intensive frame-by-frame validation. The computational intensity of models like MedSAM also poses practical barriers, as their prompt engineering requirements may not align with the need for fully automated, real-time navigation assistance during procedures. These limitations highlight the need for domain-specific optimizations to make such advanced models clinically viable for colonoscopy applications.

III. Methodology

A. Method Pipeline

Our proposed dual-path framework introduces PC-SAM (Prompt-Controlled Segment Anything Model), an enhanced adaptation of MedSAM that incorporates two key innovations to improve colonoscopy navigation. The model first employs an automatic binary classifier within its prompt encoder to categorize input images into two anatomical groups: O-ring dominant scenes with dark central regions and haustral fold-dominant scenes featuring prominent crescent-shaped ridges. This classification enables subsequent processing through specialized pathways, each optimized through LoRA (Low-Rank Adaptation) fine-tuning that maintains the base model's weights while efficiently training low-rank matrices for anatomy-specific segmentation.

For O-ring dominant images, PC-SAM combines adaptive thresholding with prompt-guided segmentation to accurately localize the lumen center. The haustral fold-dominant pathway utilizes an edge-enhanced variant with morphological constraints to detect fold convergence points. By training separate weight sets through parallel processing branches while sharing the foundational MedSAM architecture, our framework achieves precise lumen detection while maintaining computational efficiency across diverse colonoscopic scenes. The system's performance is further enhanced by the binary classifier's dynamic routing, which ensures optimal processing for each anatomical presentation without manual intervention.

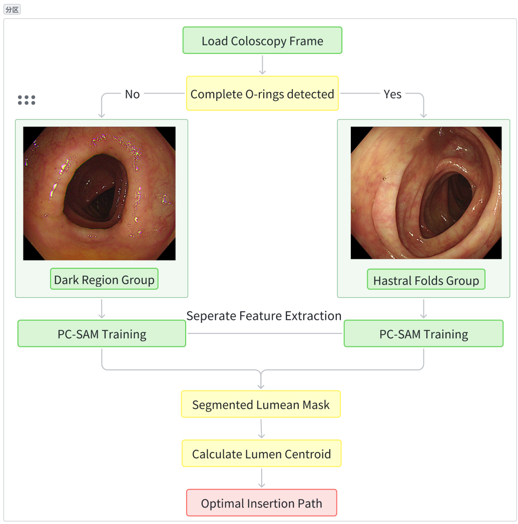

Fig. 1. Overall Pipeline of Methods

B. *Dataset Description and Processing*

The study employs a dataset of 3,000 expert-annotated colonoscopy images collected during clinical examinations. The dataset maintains an equal distribution between dark region characteristics and haustral fold patterns. Each image is accompanied by comprehensive metadata, including precise coordinate lists defining ground-truth lumen mask boundaries and classification labels indicating the dominant anatomical feature group.

For preprocessing, we first (1) convert the coordinate-based annotations into binary mask images by systematically transforming the lumen boundary coordinates, then (2) apply rigorous quality control to exclude images with incomplete anatomical annotations or insufficient resolution, and finally (3) implement a five-fold cross-validation scheme, partitioning the data into balanced subsets that preserve consistent representation across anatomical variations. Each fold contains designated training, validation, and testing sets to ensure a reproducible experimental setup.

1) Data Augmentation: To address the small dataset constraint, the augmentation strategy uses tiered transformations. For training, it first standardizes image dimensions then introduces variability via random resized cropping (70-100 percent of original size, 1.3-1.8 aspect ratio) and subtle brightness/contrast shifts (±10 percent), expanding effective sample diversity. Validation data undergoes only fixed resizing to preserve integrity for accurate performance assessment.

C. *Evaluation Protocol*

To rigorously assess our method's performance, we established a comprehensive evaluation protocol employing two complementary metrics that measure different aspects of segmentation quality. The Intersection over Union (IoU) metric quantifies the regional overlap between predicted segmentations and ground truth annotations by calculating the ratio of intersection area to union area, providing a global measure of detection accuracy. For finer assessment of boundary precision, we utilize the Dice Coefficient which computes twice the overlap area divided by the sum of individual mask areas, making it particularly sensitive to edge alignment quality.

IV. Experiments and Results

A. *Current model performance on Haustral Folds Group*

Model Name	Val IOU	VAl DSC
UNet	0.632	0.742
UNet++	0.688	0.794
PC-SAM	0.767	0.871

TABLE I
Model Performance

More experimetns results will come soon