



AI Assisted Polyp Detection and Segmentation in Colonoscopy

Değer Ayata, PhD

Artificial Intelligence and Computer Vision Department
AIATUS AI
Istanbul, Türkiye
deger.ayata@aiatus.com

Abstract—This study presents a two-stage deep learning pipeline for real-time polyp localization and segmentation in endoscopic imagery. The system integrates a YOLO detector for rapid, frame-level candidate identification and a U-Net segmenter with a ResNet-50 encoder for pixel-accurate mask generation. Evaluation on a held-out test set demonstrated strong performance: the detection model achieved 92.3% precision, 86.0% recall, and a mean Average Precision (mAP@0.5) of 94.8%, while the segmenter reached a Dice score of 0.936 and IoU of 0.672. Model selection during training was guided by background-ignored IoU@0.5, contributing to a decoupling between Dice and IoU metrics due to their differing sensitivity and aggregation methods. Both components individually satisfied real-time throughput requirements on the target GPU.

Keywords—Polyp detection; image segmentation; colonoscopy; medical image processing; polyp detection; object detection

I. INTRODUCTION

Colorectal cancer remains a major global health concern, with nearly 1.9 million new cases reported each year [1]. Since the majority of CRC cases originate from adenomatous polyps, early detection and removal during colonoscopy are key to reducing both incidence and mortality. However, studies have shown that even experienced clinicians may miss 17–28% of polyps during routine procedures [2,3]. Missing polyps—especially small or flat ones—can be a serious issue, as they may later turn into cancer, reducing the effectiveness of screening efforts [4]. This makes it clear that there's an urgent need for supportive technologies that can help doctors detect polyps more consistently and accurately.

In recent years, AI has become a valuable tool for helping doctors detect polyps during colonoscopies [1]. Modern computer-aided detection systems can now analyze colonoscopy footage in real time, flagging polyps that might be missed by the human eye [2,4]. Deep learning model and computer vision techniques have shown impressive accuracy in medical imaging tasks, including both polyp detection and segmentation [5]. For instance, real-time object detectors such as those in the YOLO family have proven effective at spotting polyps with high precision during live procedures. AI's potential here is significant—it can help reduce miss rates and improve consistency across different clinicians, ultimately making colonoscopy screenings safer and more reliable. Still, building

accurate detection models is no easy task. Polyps vary greatly in appearance—some are small, flat, and hard to distinguish from healthy tissue [4]. Challenging visual conditions like folds, glare, or fluid in the colon further complicate detection [4]. To be practical in real-world settings, models also need to be computationally efficient, ideally running at 25 frames per second or more to keep up with live video.

To address this challenge, real-time AI-based computer vision systems are being developed to assist physicians during colonoscopies. These systems can process live colonoscopy camera feeds and provide consistent detection support without fatigue or variability.

II. LITERATURE REVIEW

Early polyp detection approaches applied conventional computer vision and older machine learning techniques (e.g. hand-crafted features, classic CNNs) [7]. In recent years, the field has shifted almost entirely to deep learning object detectors for this task. Two main families of object detection models have been explored: (1) two-stage detectors like Faster R-CNN (first generate region proposals, then classify) and (2) one-stage detectors like YOLO and SSD (which directly predict boxes in one pass) [1]. Given the real-time requirements of colonoscopy, one-stage models (particularly the YOLO series) have become the dominant choice due to their speed advantage, while still achieving high accuracy.

Recent advances in computer-aided detection have focused on fast, single-stage detectors that maintain the frame rates needed for clinical use while enhancing the ability to detect small lesions. Reviews and benchmarking studies highlight YOLO (You Only Look Once) style models as practical solutions well-suited for integration into endoscopy procedures [4,6,7]. For example, variant of YOLO with improvements for small target detection and attention mechanisms—has shown increased average precision on public datasets without sacrificing efficiency required for real-time deployment. Besides YOLO, Hyb-SSDNet offers an updated version of the SSD (Single Shot Detector) framework, incorporating lightweight backbone networks and multi-scale feature fusion to deliver competitive performance when speed is essential [5]. Traditional end-to-end approaches have also proven that high frame-per-second (FPS)



processing pipelines can successfully handle entire colonoscopy videos [3].

Another key task alongside polyp detection is polyp segmentation, which involves labeling every pixel in an image as either part of the polyp or the surrounding tissue, resulting in a clear binary mask highlighting the polyp area. Accurate segmentation is especially important because it helps determine the size of the polyp, defines precise margins for removal, and can assist in identifying the polyp type based on its shape and texture. However, segmenting polyps is a complex challenge due to their diverse appearances and often indistinct or fuzzy edges. These unclear boundaries arise from factors such as uneven lighting and the smooth transition between the polyp and the surrounding mucosa. On the segmentation front, UNet variants achieve high Dice scores for polyp mask accuracy [8, 9].

III. SCOPE AND CONTRIBUTION

In this study a pragmatic, two-module setup was targeted for single-workstation clinical use 1) computer-aided detection (CAD) and computer-aided segmentation (CAS). CAD module is a fast detector localizes candidates every frame, and a precise segmenter (CAS) refines boundaries on demand. Public, reproducible datasets and mainstream architectures (YOLO variants, U-Net variants) were emphasized due to adoptability and prior success in colonoscopy. The contributions consist of a compact real-time inference pipeline with measured latencies, a clear training recipe with reproducible open dataset splits, and a concise, project-relevant literature recap positioning the design.

IV. METHODS

A. Colonoscopy Data Source and Dataset

Colonoscopy images are collected as part of standard gastrointestinal diagnostic procedures. During the exam, a thin, flexible instrument known as a colonoscope—which includes a high-definition camera—is guided through the rectum to visually examine the colon's interior. As the device records live video, clinicians can pause the feed to capture individual frames that show suspicious regions, such as potential polyps. These selected images are then annotated by specialists, who outline abnormal areas for use in medical research. This expert-labeled image data plays a vital role in training and validating AI systems for tasks like polyp detection and segmentation.

For all experiments in this study, the Kvasir-SEG dataset was used exclusively. This dataset includes 1,000 colonoscopy images, each paired with a manually verified binary mask that outlines the polyp region [9]. These images were collected during routine diagnostic colonoscopy procedures, where a flexible endoscope equipped with a high-resolution video camera captures real-time footage of the colon's interior. From this video stream, still frames containing visible polyps are extracted and annotated by medical experts to serve as training

data for computer vision models. The images in Kvasir-SEG vary in resolution, ranging from approximately 332×487 up to 1920×1072 , and each image comes with exactly one corresponding mask, making it well-suited for supervised segmentation tasks. To structure the dataset for training and evaluation, a deterministic 80/10/10 split was applied, resulting in 800 images for training, 100 for validation, and 100 for testing.

B. Polyp Detection

Polyp detection refers to the task of identifying and localizing polyps within endoscopic images, typically by drawing bounding boxes around the regions of interest. This process plays a critical role in assisting clinicians by automatically highlighting potential polyps that may require closer examination. In recent years, object detection models—particularly those in the YOLO (You Only Look Once) family—have become widely used for this task due to their balance of speed and accuracy. YOLO models are designed for real-time detection, making them highly suitable for live colonoscopy procedures. They work by dividing an image into grids and predicting both the bounding boxes and class probabilities for each region in a single forward pass. In this study, various YOLO variants were employed for polyp detection, leveraging their proven efficiency in medical imaging applications.

The effectiveness of YOLO models in polyp detection is typically assessed using several well-established evaluation metrics. Precision reflects the proportion of correctly identified polyps among all predicted instances, indicating how often the model is right when it makes a detection. Recall, on the other hand, measures the model's ability to find all actual polyps in the dataset, highlighting its sensitivity. A commonly used metric, mAP@0.5 (mean Average Precision at an Intersection over Union threshold of 0.5), provides an overall accuracy score based on how well the predicted bounding boxes align with the ground truth. For a more comprehensive assessment, mAP@0.5:0.95 averages the model's performance across a range of stricter IoU thresholds—from 0.5 to 0.95—offering deeper insight into its consistency and robustness.

To ensure consistent supervision across both detection and segmentation tasks, bounding boxes were automatically generated from the existing segmentation masks. This approach allowed both branches of the model to learn from a unified annotation source, reducing discrepancies between tasks. All input images were resized to a fixed resolution of 640×640 using letterboxing, a technique that maintains the original aspect ratio by adding padding where necessary. The detection model was trained for 100 epochs using a cosine learning rate schedule, starting from an initial learning rate of 0.01. This schedule gradually reduced the learning rate over time in a smooth, wave-like manner to promote stable convergence. The model checkpoint for final evaluation was selected based on its performance on the validation set, specifically using IoU@0.5 as the selection criterion.

During inference, the model operated using a single-scale input, matching the resolution used during training to ensure consistent performance. Figure 2 provides a visual example of the detector's output, showing both the predicted bounding box and confidence score on a sample frame. This visualization underscores how detection complements segmentation by identifying object locations, while segmentation further refines the boundaries.

C. Polyp Segmentation

In medical image analysis, segmentation refers to the task of assigning each pixel in an image to a specific anatomical or pathological class—such as identifying tumors, lesions, or in this case, polyps. Accurate segmentation is essential for diagnosis, treatment planning, and monitoring, especially in endoscopic procedures where polyps must be precisely localized.

This study employed U-Net architecture variants for segmentation, using a supervised learning approach with binary masks where each pixel was labeled as either “polyp” or “background.” This binary formulation simplifies the task while directing the model’s capacity toward clinically relevant regions.

The input images were resized to a fixed resolution of 512×608 pixels to standardize dimensions across the dataset. Prior to feeding into the model, images were normalized using ImageNet mean and standard deviation, which aligns data distributions with pretrained backbones. To enhance the model’s robustness against positional variance, horizontal flip augmentation was applied during training.

Training was performed over 100 epochs using the U-Net architecture, optimized with the Adam optimizer (learning rate: $1e-4$, batch size: 2). To address the significant class imbalance—where polyp regions are often small and sparsely distributed compared to the background—the Dice loss function was employed. Dice loss is particularly effective in medical segmentation tasks, as it directly optimizes for spatial overlap between predicted and ground truth masks, emphasizing accuracy in the regions of interest.

Fig. 1 illustrates the overall data pipeline, from preprocessing to model selection, highlighting the parallel flow of detection and segmentation. The U-Net is trained with binary masks and evaluated using Dice and IoU metrics, while detection differs only in its target outputs—bounding boxes instead of pixel-wise masks.

Model selection was based on $\text{IoU}@0.5$, calculated while excluding background pixels, ensuring that the evaluation focused solely on the model’s ability to segment polyp regions accurately. To provide a more comprehensive view of performance, two complementary evaluation strategies were

employed: Dice score was computed directly from the raw (non-thresholded) probability maps and averaged across the entire dataset. This approach captures overall segmentation quality in a smooth, probabilistic manner. IoU, on the other hand, was calculated after applying a 0.5 threshold to binarize the model outputs. These values were then averaged on a per-image basis, offering insight into prediction consistency at the individual image level.

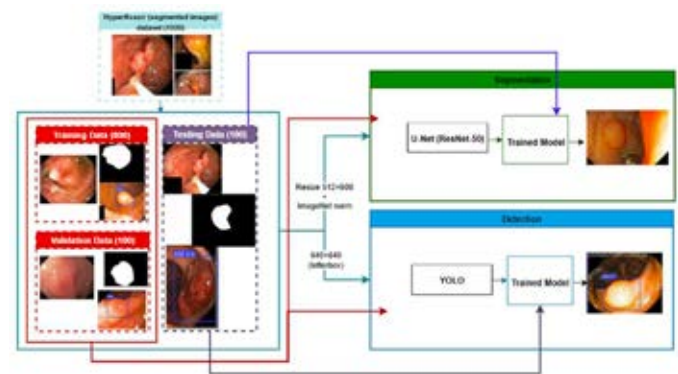


Fig. 1. Training and data workflow for the models.

Fig. 2 presents a sample inference result, where segmentation masks and their contours are overlaid on an original frame, showcasing boundary refinement around localized polyps.

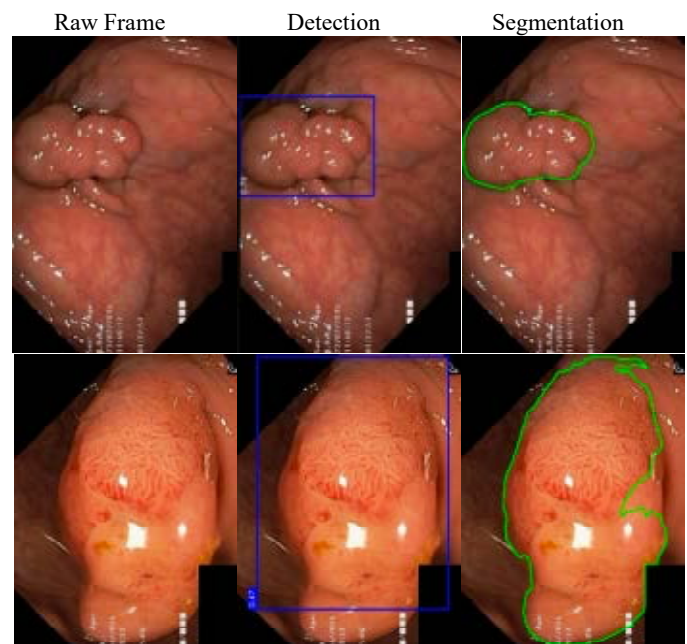


Fig. 2. Qualitative example on a representative colonoscopy frame.

Fig. 3 shows the real-time deployment pipeline, beginning with frame capture and shared preprocessing, followed by independent detection and segmentation stages, and concluding with overlay composition.

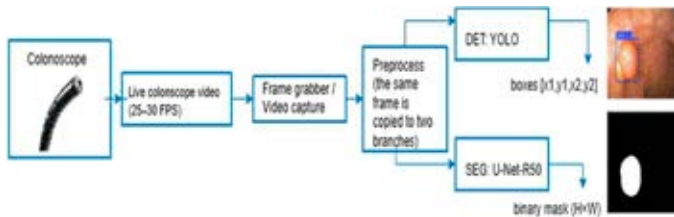


Fig. 3. Real-time inference pipeline

V. RESULTS AND DISCUSSION

The YOLOv8-m object detection model demonstrated strong performance in localizing polyp regions within the held-out test set. As shown in Table 1, the model achieved high scores across key evaluation metrics—indicating both accurate and consistent detection behavior.

Specifically, the model reached a precision of 0.923, meaning the majority of its predicted bounding boxes corresponded to true polyp locations, with minimal false positives. It also attained a recall of 0.860, reflecting its ability to detect most of the true polyp instances in the test images. These values combine to produce an F1 score of 0.890, which balances precision and recall into a single metric. Finally, the model achieved a mean Average Precision at IoU 0.5 (mAP@0.5) of 0.948, indicating high overall detection quality across varying confidence thresholds.

Table 1: YOLO variants detection performance

Method	Precision	Recall	F1	mAP@0.5
YOLOv8-m	0.923	0.860	0.890	0.948

Following detection, the U-Net segmentation model—based on a ResNet-50 backbone—was evaluated on the same test split. It demonstrated a high degree of pixel-level alignment between predicted masks and ground truth, confirming its effectiveness in detailed polyp boundary delineation. Performance metrics for segmentation are summarized in Table 2.

Table 2: U-Net segmentation performance

Method	Dice	IoU
U-Net (ResNet-50)	0.936	0.672

VI. CONCLUSION

The proposed system combined a fast object detector and a precise segmentation model to achieve both reliable localization and accurate boundary refinement. YOLOv8-m effectively identified polyp candidates on a frame-by-frame basis, while the U-Net segmenter refined these regions into high-resolution, pixel-level masks. Both models, when run independently, met real-time performance requirements on the target GPU.

Despite promising results, the current pipeline has limitations. Training was conducted on a single dataset, which restricts generalizability to other clinical environments.

Additionally, the system has not been stress-tested in continuous video streams, where temporal noise and performance drift may emerge. Future improvements should prioritize cross-dataset validation, integration of lightweight temporal smoothing to reduce frame-level inconsistencies, and latency-aware inference strategies that maintain a balance between responsiveness and accuracy in real-world use.

REFERENCES

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, pp. 209–249, 2021, doi: 10.3322/caac.21660.
- [2] S. Zhao, S. Wang, P. Pan, T. Xia, X. Chang, X. Yang, L. Guo, Q. Meng, F. Yang, W. Qian, Z. Xu, Y. Wang, Z. Wang, L. Gu, R. Wang, F. Jia, J. Yao, Z. Li, and Y. Bai, "Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-Analysis," *Gastroenterology*, vol. 156, no. 6, pp. 1661–1674.e11, 2019, doi: <https://doi.org/10.1053/j.gastro.2019.01.260>.
- [3] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning," *IEEE Access*, vol. 9, pp. 40496–40510, Mar. 2021, doi: <https://doi.org/10.1109/ACCESS.2021.3063716>.
- [4] Y. Yu, K. Lin, J. Hong, R.-G. Tsai, and Y. Huang, "PD-YOLO: Colon Polyp Detection Model Based on Enhanced Small-Target Feature Extraction," *Computers, Materials & Continua*, vol. 82, no. 1, pp. 913–928, 2025, doi: <https://doi.org/10.32604/cmc.2024.058467>.
- [5] Souaidi M, El Ansari M. Multi-Scale Hybrid Network for Polyp Detection in Wireless Capsule Endoscopy and Colonoscopy Images (Hyb-SSDNet). *Diagnostics*. 2022;12:2030. <https://www.mdpi.com/2075-4418/12/8/2030> MDPI
- [6] J. Wan, W. Zhu, B. Chen, L. Wang, K. Chang, and X. Meng, "CRH-YOLO for precise and efficient detection of gastrointestinal polyps," *Scientific Reports*, vol. 14, no. 1, p. 30033, Dec. 2024, doi: <https://doi.org/10.1038/s41598-024-81842-9>.
- [7] S. Gupta and J. Singla, "YOLO-LAN: Precise Polyp Detection via Optimized Loss, Augmentations and Negatives," *arXiv preprint arXiv:2509.19166 [cs.CV]*, 2025. doi: 10.48550/arXiv.2509.19166
- [8] S. Safarov and T. K. Whangbo, "A-DenseUNet: Adaptive Densely Connected UNet for Polyp Segmentation in Colonoscopy Images with Atrous Convolution," *Sensors*, vol. 21, no. 4, p. 1441, 2021, doi: <https://doi.org/10.3390/s21041441>.
- [9] Jha, D., Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, (2020). Kvasir-SEG: A Segmented Polyp Dataset. In: Ro, Y., et al. MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science(), vol 11962. Springer, Cham. https://doi.org/10.1007/978-3-030-37734-2_37