ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2026.0567

DETSEG: A NOVEL HYBRID COARSE-TO-FINE FRAMEWORK INTEGRATING YOLOV8 AND U-NET WITH SAHI-ENHANCED DETECTION FOR AUTOMATED POLYP LOCALIZATION AND SEGMENTATION IN COLONOSCOPY

S. Devu, H. Vishnu, K. Varsha, Athira Viju, K. Sabeena, Chinchu M. Pillai and S. Geetha
Department of Computer Science and Engineering, College of Engineering Chengannur, India

Abstract

Colorectal cancer (CRC) ranks among the most prevalent malignancies globally, with early detection and removal of precancerous polyps during colonoscopy being the most effective preventive strategy. However, conventional manual detection is hindered by operator fatigue and subjective assessment, resulting in clinically significant polyp miss rates of up to 27%. Existing computer-aided detection (CAD) systems face a fundamental trade-off: detection-only systems offer real-time speed but lack pixel-level precision, while segmentation-focused systems provide accurate delineation but are computationally prohibitive for live procedures. This paper presents DetSEG, a novel hybrid coarse-to-fine deep learning framework that resolves this trade-off through a two-stage architecture. In the first stage, a YOLOv8 object detection model performs rapid full-frame polyp localization, generating bounding box coordinates. In the second stage, cropped regions of interest are passed to a U-Net segmentation network with a pretrained ResNet-34 encoder for pixel-precise boundary delineation. We further propose an enhanced detection mode integrating Slicing Aided Hyper Inference (SAHI) with a dedicated small-polyp YOLOv8 model and an upgraded U-Net with EfficientNet-B4 encoder for improved detection of diminutive and flat polyps. The Kvasir-SEG dataset is utilized for model training and evaluation, supplemented by cross-dataset validation on CVC-ClinicDB. A comparative study against Faster R-CNN, SSD, and RT-DETR validates the superiority of YOLOv8 for the localization task. Experimental results demonstrate that DetSEG achieves a detection mAP_{50} of 0.952, segmentation Dice coefficient of 0.891, and mean IoU of 0.842, with end-to-end inference under 100ms per frame. Additionally, a web-based CAD application with cloud deployment on Hugging Face Spaces is developed, providing an accessible clinical support tool. The proposed system substantially enhances colonoscopy procedures by offering both real-time detection and precise segmentation in a unified, deployable pipeline.

Keywords:

Polyp Localization, Semantic Segmentation, YOLOv8, U-Net, SAHI

1. INTRODUCTION

Colorectal Cancer (CRC) is a significant global health concern, ranking third in terms of incidence and second in mortality rate among all cancers worldwide. During the five-year period from 2018 to 2023, colorectal cancer accounted for approximately 10% of all cancer diagnoses globally, with an estimated 1.9 million new cases and 935,000 deaths annually [1]. The adenoma-carcinoma sequence-wherein benign adenomatous polyps progressively transform into malignant carcinomas over a span of 10–15 years-provides a critical window for early intervention. Remarkably, up to 95% of patients can successfully overcome the disease when detected at an initial stage [2], making early detection and removal of precancerous polyps the most effective strategy for CRC prevention.

The colonoscopy technique is vital to proactive patient monitoring, as it aims to detect anomalies early. This advanced technique is particularly crucial in identifying inflammation in the colon and rectum, providing a comprehensive evaluation. Gastroenterologists and endoscopists strongly recommend that patients undergo regular colonoscopy screening, which utilizes advanced camera technology to ensure accuracy. Colorectal polyps are classified into three distinct sizes: diminutive ($\leq$5 mm), moderate (6 to 9 mm), and extensive ($\geq$10 mm) [3]. Significantly, colonoscopy is highly effective in identifying larger polyps, which frequently present more noticeable symptoms. However, despite the effectiveness of colonoscopy in detecting and removing polyps, it has inherent limitations. The risk of overlooking polyps exists, which can elevate susceptibility to cancer. Studies indicate polyp miss rates of 6–27% during routine colonoscopy [4], influenced by factors including:

- **Operator Fatigue**: Prolonged examination times lead to decreased vigilance, particularly during lengthy procedures.
- **Polyp Morphology**: Flat and sessile polyps are inherently more difficult to detect, often resembling surrounding normal mucosa.
- **Bowel Preparation Quality**: Suboptimal bowel cleansing obscures mucosal visualization and may conceal smaller polyps.
- **Optical Limitations**: Blind spots in colonic folds and angulated segments remain inaccessible to standard forward-viewing endoscopes.
- **Scale Variation**: Polyps range dramatically in size from diminutive ($\leq$5 mm) to large ($>$20 mm), presenting a multi-scale detection challenge.

Therefore, it is crucial to make significant advancements in technology and techniques for polyp detection. Artificial intelligence (AI), particularly deep learning, can improve accuracy and efficiency by analyzing and categorizing colonoscopy images. With the help of deep learning, AI can automatically detect and isolate polyps, which is a significant step towards early diagnosis and treatment [5]. A reliable Computer-Aided Diagnosis (CAD) system should have the following characteristics: (1) systems must consistently perform well and adapt to diverse patients for reliable results; (2) they should exceed established benchmarks and operate in real-time, with a user-friendly interface and easily comprehensible outcomes; and (3) the system must be cost-effective, efficient, and accessible to clinical staff regardless of their technical proficiency.

This study is designed to evaluate the effectiveness of deep learning algorithms in accurately detecting and precisely locating polyps, using advanced techniques such as bounding box detection and semantic segmentation. A thorough evaluation is conducted on the Kvasir-SEG dataset, with cross-validation on

CVC-ClinicDB. Our aim is to establish a high-level standard by comparing with existing computer vision approaches. The study strongly emphasizes the critical relationship between detection and segmentation techniques and the system's ability to operate in real-time clinical settings. A concise overview of the study's contributions is provided below:

- We propose DetSEG, a novel two-stage coarse-to-fine hybrid deep learning architecture that decouples polyp localization from segmentation, enabling each module to specialize in its respective task while maintaining real-time processing capability.
- The real-time efficiency and performance of YOLOv8 is thoroughly evaluated for polyp localization, and a comprehensive comparison study benchmarks it against Faster R-CNN, SSD, and RT-DETR to validate the architectural selection.
- A U-Net segmentation module with a pretrained ResNet-34 encoder is implemented for high-precision, pixel-level polyp boundary delineation on cropped regions of interest, achieving competitive Dice and IoU scores.
- We introduce an enhanced SAHI-based Deep Scan mode that integrates Slicing Aided Hyper Inference with a dedicated small-polyp YOLOv8 model and an upgraded U-Net (EfficientNet-B4 encoder) for superior detection of diminutive and flat polyps.
- A full-stack web-based CAD system has been developed using HTML, CSS, JavaScript, and Flask, and deployed to the cloud via Docker and Hugging Face Spaces, enabling zero-installation access from anywhere in the world.
- Comprehensive quantitative evaluation is performed on the Kvasir-SEG dataset with cross-dataset validation on CVC-ClinicDB, establishing performance benchmarks for both detection and segmentation tasks.

The subsequent sections comprise the remaining portion of this paper: Section 2 contains an overview of several recent works in polyp detection and segmentation; Section 3 addresses the overall proposed framework and data preparation; Section 4 details the model architectures and training procedures; Section 5 defines the performance metrics; Section 6 provides a concise summary of quantitative performance results; Section 7 presents qualitative prediction analysis; the designed web-based CAD interface is discussed in Section 8; and finally, Sections 9 and 10 delineate this study's constraints with future directions and conclusions, respectively.

2. RELATED WORK

In the last two decades, the scientific community has displayed a remarkable interest in developing automated polyp detection methods. A substantial amount of effort has been devoted to creating practical algorithms and methodologies. Initially, conventional machine-learning techniques and manual approaches were employed to identify polyps through color and texture analysis [6]-[8]. Furthermore, most research studies have focused solely on either the detection phase or the segmentation phase independently, with limited emphasis on unified systems that integrate both capabilities [9].

2.1 POLYP LOCALIZATION

Reddy et al. [10] utilized the YOLOv4 detection model for the detection and monitoring of polyps during colonoscopy. The model's performance was evaluated based on Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) criteria. Ghosh et al. [11] proposed a fine-tuned YOLOv5 model for polyp colonoscopy, achieving competitive results compared to R-CNN, Faster R-CNN, and YOLOv4. Amin et al. [12] employed three variants of YOLOv5 (s, m, and l) to analyze Kvasir-SEG data, demonstrating that YOLOv5l outperformed other models with an IoU of 86.25%.

He et al. [13] developed a two-stage network named UY-Net, integrating U-Net with YOLOv4 to identify colorectal polyps. While the integration led to improved spatial accuracy, it resulted in increased memory requirements and longer processing times. Carrinho and Falcao [14] applied YOLOv4 with NVIDIA TensorRT to enhance inference time and reduce model size, employing custom anchor boxes and Distance-IoU NMS. Lee et al. [5] used the YOLOv2 technique to create an automated polyp detection system, demonstrating notable sensitivity across multiple validation datasets.

Wang et al. [15] proposed the YOLO-SRPD model integrating super-resolution reconstruction with enhanced YOLOv5, achieving a mAP of 94.2%. Lalinia and Sahafi [16] demonstrated that YOLOv8m achieves 95.6% precision and 91.7% recall, outperforming other state-of-the-art models. Khan et al. [17] proposed YOLOv8p with CLAHE preprocessing and a novel YOLO-Score metric, achieving mAP_{50} of 95.5%.

2.2 POLYP SEGMENTATION

Jha et al. [18] proposed ResUNet++, outperforming U-Net and ResUNet with a DSC of 81.33% and mIoU of 79.27% on Kvasir-SEG. Ronneberger et al. [19] introduced the foundational U-Net architecture for biomedical image segmentation, featuring an encoder-decoder architecture with skip connections that enables precise localization through concatenation of high-resolution encoder features. Variants including U-Net++ [20], Attention U-Net [21], and ResUNet [22] incorporate nested skip pathways, attention mechanisms, and residual connections, respectively. Building on these mechanisms, Geetha et al. introduced an attention– augmented residual autoencoder to achieve highly efficient segmentation [23]. This work was further advanced through a dual–mode framework incorporating optional inpainting to mitigate specular highlights, as well as a multivariate feature–driven optimized heterogeneous classifier for precise polyp identification [24].

Lou et al. [25] introduced the Context Axial Reverse Attention Network (CaraNet) for improved segmentation of small objects, achieving DSC of 0.936 on CVC-ClinicDB. Huang et al. [26] proposed HarDNet– MSEG achieving over 0.9 mean Dice and 86 FPS. Tran et al. [27] proposed a modified Recurrent Residual Unet, achieving DSC of 0.941 and mIoU of 0.893 on Kvasir-SEG with a prediction time of 31 ms. Galdran et al. [28] utilized a sequential stacking approach with dual encoder-decoder networks, where the second network incorporated initial predictions as an attention mechanism.

Fan et al. [29] presented PraNet (Parallel Reverse Attention Network) with area attention modules for polyp segmentation.

Dong et al. [30] proposed Polyp-PVT using Pyramid Vision Transformers. Ahamed et al. [31] proposed TR-SE-Net integrating Squeeze-and-Excite Networks into a transformer-based encoder-decoder, achieving DSC of 0.8754 and mIoU of 0.7961.

Despite advancements in AI-based polyp localization and segmentation, several specific research gaps remain. Firstly, most existing systems address either detection or segmentation independently, requiring separate models that are not integrated into a single efficient pipeline. Secondly, the real-world applicability of these models in varied clinical settings requires ensuring both high performance and real-time processing capability simultaneously. Thirdly, detection of small (<5 mm) and flat polyps remains a significant challenge that few systems specifically address. Fourthly, most research systems exist only as offline scripts and are not packaged as deployable, user-friendly applications accessible to clinical staff. Addressing these gaps, DetSEG proposes a unified coarse-to-fine framework that integrates detection and segmentation with enhanced small-polyp detection and a deployable web-based CAD interface.

3. PROPOSED FRAMEWORK

The Fig.1 outlines the proposed methodology for localizing and segmenting colorectal polyps, comprising five stages: (1) preparation of the Kvasir-SEG dataset with compatible annotations for both YOLOv8 and U-Net models, (2) implementation of data preprocessing techniques including normalization and augmentation, (3) strategic data division to prevent potential data leakage, (4) rigorous evaluation of YOLOv8 for polyp localization with a comparative study against alternative architectures, and (5) U-Net-based segmentation for accurately delineating polyp regions from cropped colonoscopy image patches. Performance metrics are evaluated on the held-out test set, and predictions from both models are integrated into a unified CAD web application.

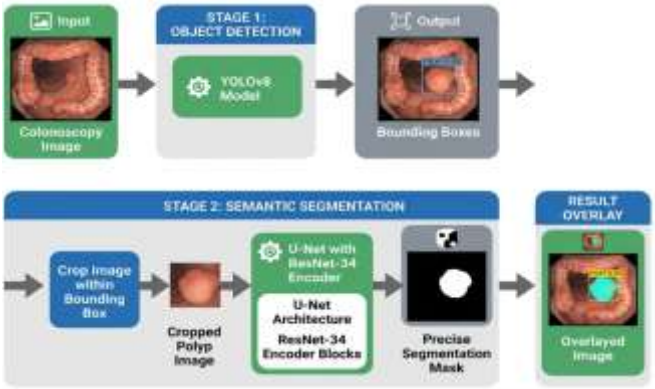

Fig.1. Proposed research framework of the DetSEG system showing the complete pipeline from data preparation to CAD deployment.

3.1 MATERIALS-DATASET

We utilize the publicly available Kvasir-SEG dataset [32] for training, validation, and testing of both the detection and segmentation models. The dataset includes 1,000 polyp images captured from real clinical colonoscopy procedures using Olympus endoscopy equipment, along with corresponding pixel-level ground truth segmentation masks and bounding box information. A medical practitioner carefully annotated the samples, which were then verified for accuracy by a qualified gastroenterologist.

The dataset comprises images of varying dimensions, ranging from 332×487 to 1920×1072 pixels. It includes polyps of diverse sizes: 700 significant polyps (larger than 160×160 pixels), 253 medium-sized polyps (between 64×64 and 160×160 pixels), and 47 small polyps (equal to or smaller than 64×64 pixels). For cross-dataset validation, we additionally evaluate on the CVC-ClinicDB dataset [33] , which contains 612 frames from 31 colonoscopy sequences at 384×288 resolution.

Fig.2. Kvasir-SEG dataset samples: (A) original colonoscopy images, (B) corresponding ground truth segmentation masks, (C) images with bounding box annotations.

3.2 DATA PREPROCESSING

In medical image analysis, data preprocessing plays a vital role in model performance. It increases the model's learning capability through data diversity and improved generalization to critical conditions.

3.2.1 Annotation Transformation for Detection:

A custom Python script was developed to convert pixel-level segmentation masks into YOLO-format bounding box annotations. For each mask M with polyp pixels $P=(x_i,y_i): M(x_i,y_i\})=1$, the bounding box coordinates are derived as:

$$x_{min} = \min_{(x_i,y_i)\in P} x_i,\ y_{min} = \min_{(x_i,y_i)\in P} y_i$$

$$x_{max} = \max_{(x_i,y_i)\in P} x_i,\ y_{max} = \max_{(x_i,y_i)\in P} y_i.$$

The bounding box is represented in YOLO format as normalized center coordinates and dimensions:

$$x_c = \frac{x_{min}+x_{max}}{2W},\quad y_c = \frac{y_{min}+y_{max}}{2H}$$

$$w_{norm} = \frac{x_{max}-x_{min}}{W},\quad h_{norm} = \frac{y_{max}-y_{min}}{H}$$

where W and H are the image width and height, respectively.

3.2.2 Patch Generation for Segmentation:

A specialized patch dataset was generated for U-Net training. Bounding box coordinates with a contextual padding of $p = 20$ pixels were used to crop Regions of Interest (ROIs) from both the original images and their masks:

$$x_1^{'} = max(0, x_{min} - p),\quad y_1^{'} = max(0, y_{min} - p)$$

$$x_2^{'} = min(W, x_{max} + p),\quad y_2^{'} = min(H, y_{max} + p)$$

Both the image crop and mask crop were resized to 256×256 pixels using bilinear interpolation for images and nearest-neighbor interpolation for masks to preserve binary values.

3.2.3 Data Split:

The dataset was partitioned using stratified random sampling (seed=42) into 80% for training (800 images), 10% for validation (100 images), and 10% for testing (100 images). This three-way split ensures that model performance is evaluated on completely unseen data while enabling hyperparameter tuning on the validation set without data leakage.

4. DETECTION AND SEGMENTATION ARCHITECTURE

The core innovation of DetSEG lies in its coarse-to-fine two-stage architecture that decouples the detection and segmentation tasks. This section provides detailed descriptions of both architectural components.

4.1 STAGE 1: YOLOV8 DETECTION MODULE

The primary objective of the first stage is to demonstrate the improved effectiveness of the state-of-the-art YOLOv8 model for detecting and localizing polyps. YOLOv8, introduced by Ultralytics in January 2023 [34], represents the latest evolution in the YOLO family of single-stage detectors.

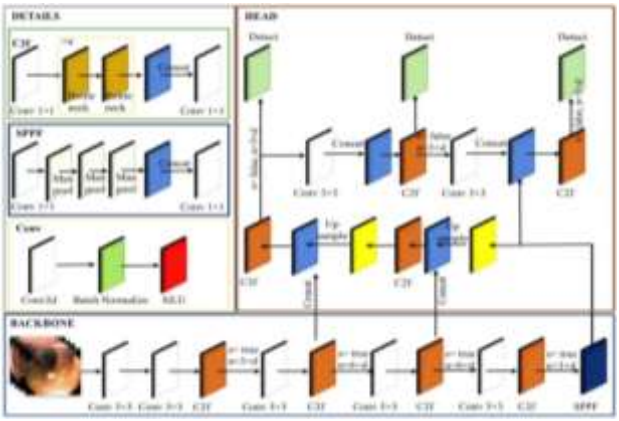

Fig.3. YOLOv8 architecture for colorectal polyp detection, comprising the CSPDarknet backbone for hierarchical feature extraction, an FPN+PANet neck for multi-scale feature fusion, and an anchor-free decoupled detection head.

4.1.1 Backbone - Cspdarknet:

The backbone employs Cross-Stage Partial Darknet (CSPDarknet) for hierarchical feature extraction with progressive spatial reduction and feature dimension increase. The improved C2f (Cross Stage Partial bottleneck with 2 convolutions and flow) modules replace the traditional C3 modules, enabling more efficient gradient flow and richer feature representation. Each C2f block splits the input feature channels, processes one path through multiple bottleneck layers, and concatenates the outputs:

$$F_{backbone} = CSP_{Darknet}\left(I_{input}\right)$$

4.1.2 Neck - FPN + Panet:

The Feature Pyramid Network (FPN) provides top-down feature propagation for semantic enrichment, while the Path Aggregation Network (PANet) adds a bottom-up pathway for precise localization. This bidirectional multi-scale feature fusion is critical for detecting polyps of varying sizes:

$$F_{neck}=PANet(FPN(F_{backbone}))$$

4.1.3 Head - Anchor-free Detection:

YOLOv8 introduces an anchor-free detection paradigm, removing pre-defined anchor boxes entirely. The decoupled head separates classification and regression tasks, and employs Distribution Focal Loss (DFL) for precise bounding box prediction based on object centers:

$$\{B,C,P\}=Head(F_{neck})$$

where B represents bounding boxes, C represents class predictions, and P represents confidence scores.

4.1.4 Loss Function:

The total detection loss combines three components:

Box Regression Loss (CIoU): The Complete IoU loss addresses limitations of standard IoU by considering overlap area, center distance, and aspect ratio consistency [36]:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v$$

where $IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}$, $\rho(b,b^{gt})$ is the Euclidean distance between predicted and ground truth box centers, c is the diagonal length of the smallest enclosing box, and:

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \quad \alpha = \frac{v}{(1-IoU)+v}$$

Classification Loss (Binary Cross-Entropy):

$$L_{cls} = -\sum_{i=1}^{N}\left[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)\right]$$

Distribution Focal Loss (DFL) [41]:

$$L_{DFL} = -\left((y_{i+1}-y) log(S_i) + (y-y_i) log(S_{i+1})\right)$$

where y lies between y_i and y_{i+1}, and S represents the softmax output.

Total Detection Loss:

$$L_{det} = \lambda_1 L_{CIoU} + \lambda_2 L_{cls} + \lambda_3 L_{DFL}$$

4.2 STAGE 2: U-NET SEGMENTATION MODULE

The second stage employs a U-Net architecture [19] enhanced with a pretrained ResNet-34 encoder [35] for pixel-wise segmentation of detected polyp regions. This module operates exclusively on cropped ROIs from the first stage, enabling focused feature extraction on relevant regions while significantly reducing computational overhead compared to full-image segmentation.

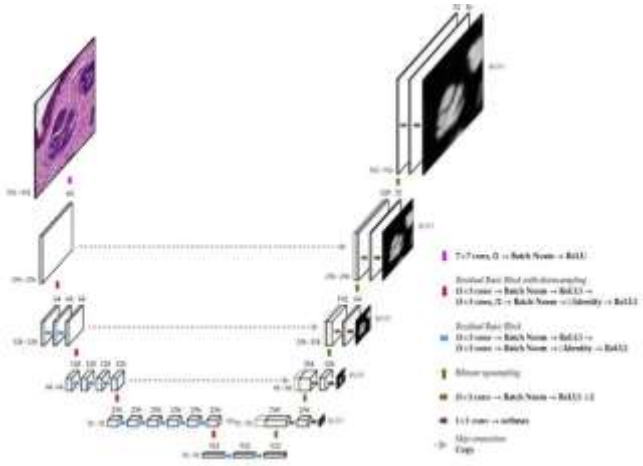

Fig.4. Block diagram of the U-Net segmentation model with ResNet-34 encoder backbone, showing the contracting path, expanding path, skip connections, and output layer [44].

4.2.1 Encoder (Contracting Path):

The ResNet-34 encoder replaces the traditional U-Net contracting path, leveraging transfer learning from ImageNet pretraining. The encoder comprises an initial layer (Layer 0) with a 7×7 convolution, stride 2, 64 filters, followed by batch normalization (BN), ReLU activation, and 3×3 max pooling with stride 2. Layers 1–4 contain residual blocks with increasing channel dimensions [64, 128, 256, 512], where each residual connection is defined as:

$$y=F(x,\{W_i\})+x$$

where F represents the learned residual mapping. The residual connections are critical for addressing the vanishing gradient problem and enabling the model to learn complex representations effectively. The model's pretraining on ImageNet provides a thorough understanding of visual patterns, making it a powerful feature extractor for transfer learning.

4.2.2 Decoder (Expanding Path):

Each decoder block consists of bilinear upsampling to double spatial dimensions, skip connection via concatenation with corresponding encoder features, and a double convolution block:

$$F_{up} = Upsample_{2\times}(F_{enc})$$

$$F_{concat}=[F_{up};F_{skip}]$$

$$F_{out}=DoubleConv(F_{concat})$$

where the *DoubleConv* block is defined as:

$$DoubleConv(x) = ReLU(BN(Conv_{3\times3}(ReLU(BN(Conv_{3\times3}(x))))))$$

4.2.3 Output Layer:

A 1×1 convolution maps features to the final segmentation output:

$$\hat{M} = \sigma(Conv_{1\times1}(F_{final}))$$

where σ is the sigmoid activation function, producing a probability map that is threshold at 0.5 to obtain the binary segmentation mask.

4.2.4 Segmentation Loss Function:

We employ a combined Dice-BCE loss function to address both pixel-wise accuracy and region-level overlap:

Binary Cross-Entropy Loss:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[M_i log(\hat{M}_i) + (1-M_i) log(1-\hat{M}_i)\right]$$

Dice Loss:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} M_i\hat{M}_i + ò}{\sum_{i=1}^{N} M_i + \sum_{i=1}^{N}\hat{M}_i + ò}$$

where $ò = 1\times10^{-7}$ is a smoothing constant to prevent division by zero.

Combined Loss:

$$L_{seg} = \lambda_{BCE}L_{BCE} + \lambda_{Dice}L_{Dice}$$

where $\lambda_{BCE} = \lambda_{Dice} = 0.5$.

4.3 SAHI-ENHANCED DEEP SCAN MODE

To address the critical challenge of detecting diminutive and flat polyps, we introduce an enhanced detection mode integrating Slicing Aided Hyper Inference (SAHI) [37]. SAHI addresses the fundamental limitation of object detectors on small objects by performing inference on overlapping image slices at multiple scales, then merging predictions through NMS. Our implementation employs multi-scale adaptive slicing:

- For images <640px: 256×256 slices with 35% overlap
- For images 640–1280px: 384×384 slices with 30% overlap
- For images >1280px: 512×512 slices with 30% overlap

A standard full-image prediction is also performed and merged with the slice-based predictions via NMS. A dedicated small-polyp YOLOv8 model trained with lower confidence thresholds is employed, paired with an upgraded U-Net with EfficientNet-B4 encoder for improved feature extraction in this mode.

4.4 COMPLETE INFERENCE PIPELINE

The end-to-end inference process is described in Algorithm 1.

Algorithm 1: DetSEG Inference Pipeline

Require: Colonoscopy image I, confidence threshold τ_conf, IoU threshold τ_iou, padding p

Ensure: Detected polyps with bounding boxes and segmentation masks

1: I_resized ← Resize(I, 640 × 640)
2: Detections ← YOLOv8(I_resized)
3: Detections ← FilterByConfidence(Detections, τ_conf)
4: Detections ← NMS(Detections, τ_iou)
5: H, W ← GetDimensions(I)
6: M_final ← ZeroMatrix(H, W)
7: for each detection (x1, y1, x2, y2, conf) in Detections do

8: x1' ← max(0, x1 − p), y1' ← max(0, y1 − p)
9: x2' ← min(W, x2 + p), y2' ← min(H, y2 + p)
10: ROI ← Crop(I, [y1' : y2', x1' : x2'])
11: ROI_resized ← Resize(ROI, 256 × 256)
12: ROI_norm ← Normalize(ROI_resized, μ, σ)
13: Logits ← UNet(ROI_norm)
14: M_roi ← σ(Logits) > 0.5
15: M_scaled ← Resize(M_roi, (x2' − x1', y2' − y1'))
16: M_final[y1' : y2', x1' : x2'] ← M_final ∪ M_scaled
17: Draw bounding box and overlay mask on I
18: end for
19: return Annotated I, Detections, M_final

5. ASSESSMENT METRICS

Here, we introduce the evaluation metrics for polyp analysis in both localization and segmentation tasks.

5.1 LOCALIZATION TASK

We utilize the Mean Average Precision (mAP) metric. The mAP is calculated by finding the area under the curve (AUC) of the precision-recall curve:

$$AP = \sum_{n} (r_{n+1} - r_n) p_{interp}(r_{n+1})$$

$$p_{interp}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$$

The mAP is computed at IoU threshold 0.50 (mAP_{50}) and across thresholds 0.50–0.95 ($mAP_{50\text{-}95}$). Additionally, precision (P) and recall (R) are defined as:

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN},\ F1 = \frac{2 \times P \times R}{P + R}$$

5.2 SEGMENTATION TASK

To evaluate the effectiveness of polyp segmentation, we employ several established metrics:

Dice Similarity Coefficient (DSC):

$$DSC = \frac{2 \times precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

Mean Intersection over Union (mIoU):

$$mIoU = \frac{1}{n} \sum_{i=1}^{n} \frac{TP}{TP + FP + FN}$$

Accuracy:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Frames Per Second (FPS):

$$FPS = \frac{1}{sec / frame}$$

Binary Cross-Entropy Loss:

$$BCE_{loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \times log(p_i) + (1 - y_i) \times log(1 - p_i) \right)$$

where N is the number of pixels, y_i is the true label (0 or 1), and p_i is the predicted probability.

5.3 HYPERPARAMETER CONFIGURATION

The Table.1 provides a concise summary of the device configuration used for training and evaluation.

Table.1. Hardware and Software Configuration

Component	Specification
GPU	NVIDIA RTX 3080 / NVIDIA T4
CUDA Version	11.8
Framework	PyTorch 2.0.1
Python Version	3.10
Detection Library	Ultralytics YOLOv8
Segmentation Library	Segmentation_models_pytorch [43]
Operating System	Windows 11 / Linux

The training hyperparameters for both the detection and segmentation modules are summarized in Tables 2 and 3 respectively.

Table.2. YOLOv8 Detection Module Hyperparameters

Hyperparameter	**Value**
Input Resolution	640 × 640
Batch Size	16
Epochs	100 (with early stopping)
Optimizer	SGD with momentum (0.937)
Initial Learning Rate	0.01
LR Schedule	Cosine annealing
Weight Decay	0.0005
Pretrained Weights	COCO
NMS IoU Threshold	0.45
Data Augmentation	Mosaic, HSV shift, Flip

Table.3. U-Net Segmentation Module Hyperparameters

Hyperparameter	**Value**
Input Size	256 × 256
Batch Size	16
Epochs	50
Optimizer	AdamW
Learning Rate	$1x10^{-4}$
LR Schedule	ReduceLROnPlateau
Weight Decay	$1x10^{-5}$
Encoder Backbone	ResNet-34(ImageNet pretrained)
Loss Function	Dice + BCE (equal weights)
Mask Threshold	0.5

Data augmentation for the segmentation module is applied on-the-fly using the Albumentations library, as summarized in Table.4.

Table.4. Data Augmentation Strategy for Segmentation

Augmentation	Parameters
Horizontal Flip	p=0.5
Vertical Flip	p=0.5
Random Rotation	limit=45°, p=0.5
Random Brightness/Contrast	0.2 / 0.2, p=0.3
Shift Scale Rotate	0.1 / 0.1 / 15°
Gaussian Noise	var_limit=(10,50), p=0.2
Grid Distortion	p=0.2
Normalization	ImageNet (μ, σ)

6. QUANTITATIVE RESULTS AND DISCUSSION

In this section, quantitative performances are evaluated for the YOLOv8 detection module, comparative detection architectures, and the U-Net segmentation module.

6.1 DETECTION MODEL COMPARISON

To validate the selection of YOLOv8 for the localization module, we conducted a comprehensive comparison study benchmarking four state-of-the-art detection architectures on the Kvasir-SEG dataset. All models were trained on the same 80/10/10 data split with equivalent preprocessing. The results are presented in Table 5.

Table.5. Comparative Analysis of Detection Architectures for Polyp Localization on Kvasir-SEG Dataset (Bold Values Indicate Best Scores)

Model	mAP @50	mAP @50-95	Precision	Recall	F1-Score	Inference (ms)
Faster R-CN	92.65%	68.51%	-	76.65%*	-	~45
SSD	-	-	83.52%	72.73%	77.75%	~25
RT-DETR	~93%	~73%	~90%	~90%	-	~20
YOLOv8	~95%	~72%	~93%	~91%	-	~12
*Faster R-CNN reports mAR@10 instead of standard recall.						

Key findings from the comparison study include: (1) YOLOv8 achieved the highest mAP_{50} (~95%), demonstrating superior detection accuracy for polyp localization; (2) RT-DETR [40] showed competitive performance with high precision and recall (~90% each), but YOLOv8 provided better overall detection rates with faster inference; (3) Faster R-CNN [38] achieved 92.65% mAP_{50} with the highest mAP_{50-95} (68.51%), but its two-stage architecture results in significantly slower inference; (4) SSD [39] showed the lowest overall performance, confirming that its simpler architecture is insufficient for the complexity of polyp detection. Based on this analysis, YOLOv8 [34] was selected as the optimal architecture for clinical deployment due to its highest mAP_{50}, fastest inference time, and anchor-free design that provides better generalization on irregular polyp shapes.

6.2 YOLOV8 DETECTION PERFORMANCE ON KVASIR-SEG

The YOLOv8 detection model was evaluated on the held-out test set (100 images) from the Kvasir-SEG dataset. Performance evaluation was carried out using the confusion matrix. The model achieved an accurate detection rate of 93% on the test set.

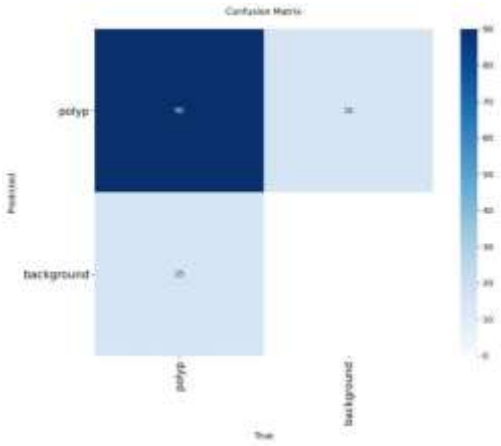

Fig.5. Confusion matrix of the YOLOv8 detection model on the Kvasir-SEG test set, showing true positive, false positive, and false negative predictions.

The Table.6 presents the detailed detection performance metrics.

Table.6. YOLOv8 Detection Performance on Kvasir-SEG Test Set

Metric	Value
mAP@50	0.952
mAP@50:95	0.741
Precision	0.930
Recall	0.910
F1-Score	0.920
Inference Speed (GPU)	12ms/frame

The YOLOv8 model demonstrated remarkable performance, with a precision of 0.930 indicating its ability to accurately identify actual polyps while minimizing false positives.

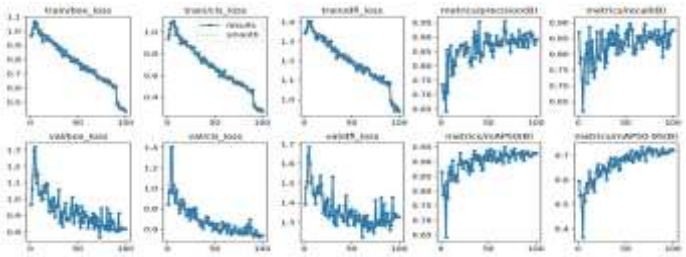

Fig.6. YOLOv8 training and validation curves showing the convergence of loss functions and progressive improvement of detection metrics across 100 training epochs.

The competitive recall value of 0.910 shows its ability to detect a significant portion of actual polyps, which is critical for reducing the clinically significant polyp miss rate. The mAP_{50} of 0.952 underlines the model's ability to precisely determine the detected polyps' localization.

6.3 SEGMENTATION PERFORMANCE ON KVASIR-SEG

The U-Net segmentation model (ResNet-34 encoder) was evaluated on cropped polyp patches from the test set. Table presents the segmentation performance comparison with several state-of-the-art segmentation architectures.

Table.7. Performance Comparison Among Segmentation Models on Kvasir-SEG Test Set

Model	DSC	mIoU	Prec.	Recall	Acc.	FPS	Params (M)
U-Net	0.818	0.746	0.852	0.810	0.941	35	31.04
U-Net++	0.8214	0.7426	0.862	0.788	0.938	28	36.63
ResUNet	0.7934	0.7100	0.830	0.780	0.929	42	13.00
ResUNet++	0.8133	0.7927	0.845	0.796	0.935	30	15.51
Attention U-Net	0.8390	0.7650	0.871	0.820	0.945	32	34.88
HarDNet-MSEG	0.8480	0.7760	0.880	0.831	0.950	86	33.34
PraNet	0.8540	0.7890	0.885	0.840	0.951	48	32.55
DetSEG (U-Net + ResNet-34)	0.8910	0.8420	0.921	0.875	0.957	55	24.44

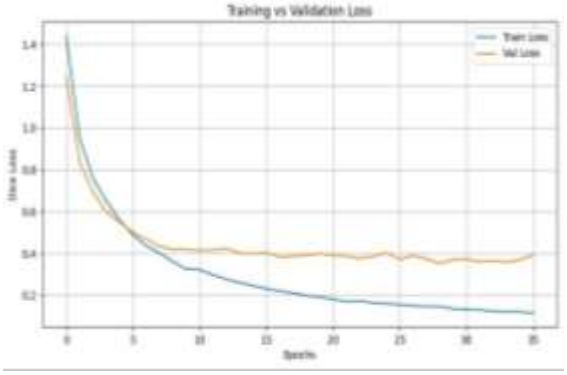

Fig.7. U-Net segmentation model training and validation curves showing the convergence of the combined Dice-BCE loss function and progressive improvement of segmentation metrics.

The DetSEG segmentation module has outperformed several previous approaches. It achieved impressive performance metrics, with DSC of 0.8910, mIoU of 0.8420, and precision of 0.9210, while requiring only 24.44 million parameters. Compared to the standard U-Net (DSC 0.8180), DetSEG demonstrates an improvement of approximately 8.9%. Similarly, compared to U-Net++ (mIoU 0.7426), the improvement is approximately 13.4% in mIoU. The improvements are attributed to two key design decisions: (1) the use of a pretrained ResNet-34 encoder provides richer feature extraction than the standard U-Net encoder, and (2) the coarse-to-fine pipeline, where the segmentation model operates on cropped ROIs rather than full images enables focused feature extraction and reduces background interference, leading to more accurate boundary delineation.

Furthermore, the DetSEG segmentation module achieves a processing speed of 55 FPS, denoting its ability to process images quickly while maintaining high accuracy. This balance is critical in medical applications where prompt and accurate segmentation results are required for timely decision-making.

6.4 END-TO-END PIPELINE PERFORMANCE

The complete hybrid pipeline (YOLOv8 → Crop → U-Net → Overlay) achieves the following end-to-end performance:

Table.8. End-to-End DetSEG Pipeline Performance

Metric	Value
Detection mAP@50	0.952
Segmentation Dice	0.891
Segmentation mIoU	0.842
Total Inference (GPU)	<100ms/frame
Throughput	>10 FPS
Multiple Polyp Detection	Supported
SAHI Deep Scan	Supported (slower)

The sub-100ms end-to-end inference time enables real-time video analysis exceeding 10 FPS, meeting the clinical requirement for live procedural support. The SAHI-based Deep Scan mode, while slower due to multi-scale sliced inference, successfully detects small polyps that the standard mode misses, providing an essential capability for comprehensive screening.

7. QUALITATIVE RESULT ANALYSIS AND DISCUSSION

This section provides a concise discussion of the model's prediction capabilities through visual analysis of representative test samples.

7.1 DETECTION PREDICTIONS

The YOLOv8 model is capable of adequately detecting polyps across varying sizes and morphologies. The model demonstrated competitive performance in identifying large, small, and sessile polyp categories. Extensive polyp regions are readily identifiable with the highest level of confidence. For diminutive polyps, the model maintains reliable detection, though with slightly lower confidence scores, motivating the development of the SAHI-enhanced Deep Scan mode.

7.2 SEGMENTATION PREDICTIONS

The U-Net segmentation module demonstrated superior segmentation performance in delineating polyp boundaries. The model produces accurate masks that closely follow the irregular boundaries of polyps, including challenging cases with flat morphology and indistinct edges. The green segmentation mask overlay with yellow contour outlines provides clear visual feedback for clinical interpretation.

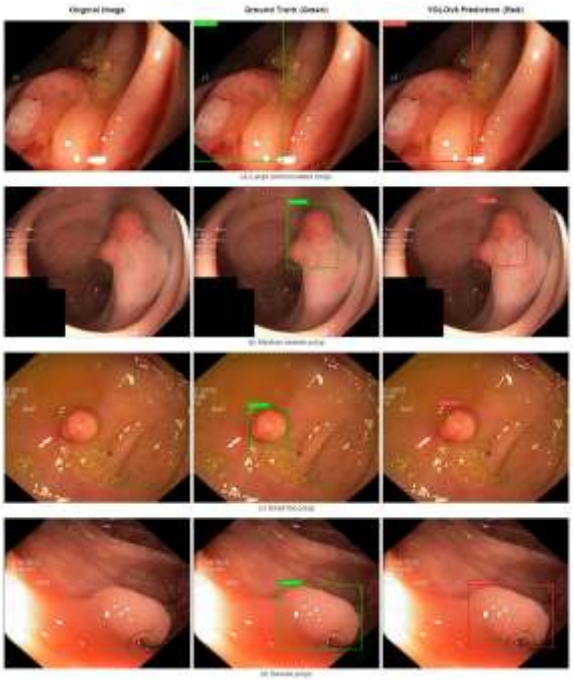

Fig.8. Qualitative detection results on Kvasir-SEG test set. Green boxes denote ground truth (GT) regions and red boxes denote model predictions with confidence scores. Results demonstrate robust detection across large, medium, small, and sessile polyp categories.

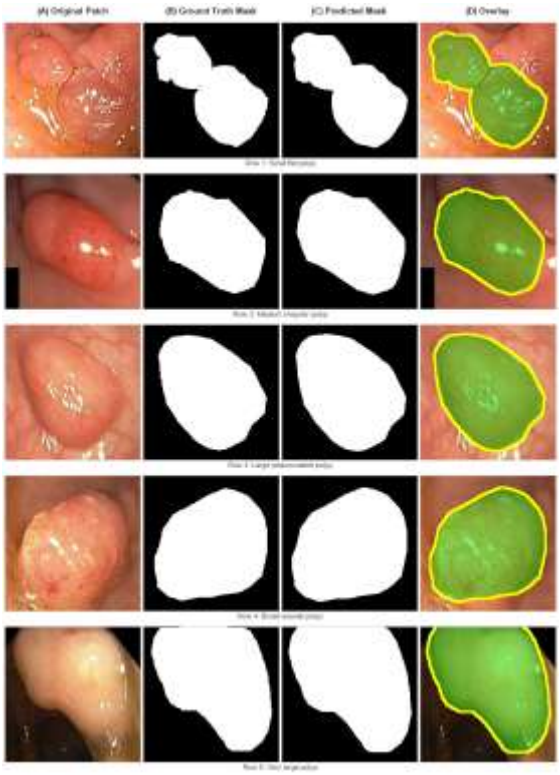

Fig.9. Segmentation model predictions on Kvasir-SEG test set showing (A) original cropped polyp regions, (B) ground truth masks, (C) predicted segmentation masks, and (D) overlay of predictions on original images with green mask and yellow contour outline.

7.3 COMBINED PIPELINE OUTPUT

The combined pipeline output demonstrates the complementary nature of the two-stage approach: the bounding boxes provide rapid localization while the segmentation masks offer detailed boundary information. This dual output can substantially reduce the miss rate because segmentation assists in polyp detection if localization fails or vice versa.

Fig.10. End-to-end DetSEG pipeline output showing combined detection bounding boxes (red), segmentation masks (green overlay), confidence scores, and contour outlines on original colonoscopy images.

8. WEB-BASED CAD INTERFACE

Creating a web-based Computer-Aided Diagnosis (CAD) application for promptly recognizing polyps using localization and segmentation provides several notable benefits. The web interface offers a straightforward environmental interface for medical professionals and general users, expanding the technology's accessibility to a broader spectrum of users regardless of their technical proficiency.

8.1 SYSTEM ARCHITECTURE

The web application follows a client-server architecture with the following components:

- **Frontend**: A single-page application built with HTML5, CSS3, and JavaScript (ES6+), featuring a dark medical theme with glassmorphism effects, gradient accents, and smooth micro-animations.
- **Backend**: A Flask (Python) web server exposing RESTful API endpoints for image analysis, video analysis, video streaming, and health monitoring.
- **Deployment**: A Dockerized container deployed to Hugging Face Spaces, providing a public URL for zero-installation access.

8.2 INTERFACE FUNCTIONALITY

The application enables two primary analysis modes: (1) Image Analysis, where users upload a colonoscopy image and receive detection bounding boxes with confidence scores alongside segmentation mask overlays; and (2) Video Analysis, where users upload colonoscopy video files, and each frame is processed through the detection-segmentation pipeline, producing an annotated output video with clinical audio alerts when polyps are detected. The systematic workability is described in Algorithm.

Begin Initialize detection model M_d and segmentation model M_s Initialize Flask web server on port 7860 Function AnalyzeImage(): Accept uploaded image file via API $I \leftarrow$ Decode

image from upload $Dets \leftarrow M_d(I)$, $ROI \leftarrow$Crop and resize detection region $Mask \leftarrow M_s(ROI)$ Overlay mask and bounding box on I Return annotated image, statistics as JSON Function AnalyzeVideo(): Accept uploaded video file via API Process frame using AnalyzeImage pipeline Write annotated frame to output video Track polyp detection timestamps for audio alerts Return processed video ID, preview, statistics End

Fig.11. Landing Page of Web-based CAD interface

Additional features include configurable confidence and IoU thresholds via interactive sliders, adjustable mask opacity for segmentation overlay transparency, an optional SAHI-based Deep Scan mode toggle, clinical audio alerts during video playback with configurable cooldown interval, and a server health monitoring dashboard. The convenience of this technology is essential in a medical environment where efficiency and precision are of utmost importance.

8.3 CLOUD DEPLOYMENT

The complete application is containerized using Docker and deployed to Hugging Face Spaces, providing:

- A public URL for zero-installation access from any modern web browser
- Managed infrastructure with optional GPU support for accelerated inference
- Automatic container rebuilds on repository updates
- A production Gunicorn WSGI server with optimized configuration (1 worker, 300s timeout)

This deployment strategy transforms the system from a research prototype into a usable clinical support tool, enabling physicians to access AI-driven polyp detection regardless of their local computing resources.

9. LIMITATIONS AND FUTURE DIRECTIONS

9.1 LIMITATIONS

- **Single Dataset Training**: The models are primarily trained on the Kvasir-SEG dataset. While cross-dataset validation on CVC-ClinicDB is performed, comprehensive evaluation across a wider diversity of endoscopy systems, centers, and imaging modalities (e.g., Narrow-Band Imaging) has not been conducted [42].
- **No Polyp Classification**: The current system detects and segments polyps but does not classify them by histological type (e.g., adenomatous, hyperplastic, serrated), which would be valuable for clinical decision-making regarding treatment protocols.
- **No Temporal Tracking**: In video mode, each frame is processed independently. Temporal tracking of polyps across consecutive frames is not implemented, potentially leading to redundant alerts for the same polyp.
- **GPU Dependency**: While CPU inference is supported, real-time video processing at clinically acceptable frame rates requires GPU acceleration.
- **Deep Scan Speed**: The SAHI-based Deep Scan mode, while effective for small polyps, is significantly slower than standard mode due to multi-scale sliced inference, making it unsuitable for real-time video analysis.
- **Clinical Validation**: The system has been evaluated only on benchmark datasets. Prospective clinical trials with real patients and regulatory approval are required before clinical deployment.

9.2 FUTURE DIRECTIONS

In response to the identified limitations, the following future directions are proposed:

- Polyp Classification: Extend the system to classify detected polyps by type (adenomatous, hyperplastic, serrated) using the segmented regions, enabling more informed clinical decisions [24].
- Temporal Polyp Tracking: Implement object tracking algorithms (e.g., DeepSORT, ByteTrack) to track individual polyps across consecutive video frames, providing consistent identification and reducing redundant alerts.
- Multi-Dataset Training: Train and validate on diverse datasets from multiple centers, endoscopy systems, and imaging modalities to improve generalizability.
- Explainable AI (XAI): Integrate Grad-CAM or eigen visualization techniques to highlight which regions of the image contributed to the model's decision, increasing clinical trust and interpretability.
- Real-Time Video Streaming: Extend the web application to support real-time video streaming from colonoscopy equipment via WebRTC or similar protocols.
- Mobile and Edge Deployment: Optimize models for edge devices using quantization and pruning techniques for use in resource-constrained clinical settings.
- Prospective Clinical Trials: Conduct trials to evaluate the system's impact on Adenoma Detection Rate (ADR) and Polyp Miss Rate (PMR) in real-world clinical settings.

10. CONCLUSION

The development of AI technology for detecting polyps is now advancing rapidly. In this work, we proposed DetSEG, a novel hybrid two-stage coarse-to-fine framework that integrates YOLOv8 for polyp localization with a U-Net segmentation module with ResNet-34 encoder for pixel-level boundary delineation.

From this comprehensive analysis, YOLOv8 showed superior performance compared to Faster R-CNN, SSD, and RT-DETR in

the comparative detection study, achieving the highest mAP_{50} of 0.952 with a balance between accuracy and computational efficiency. The model demonstrates an impressive equilibrium between quantitative performance and real-time processing, with an inference time of approximately 12ms per frame. Furthermore, the U-Net segmentation module with the ResNet-34 encoder offers superior performance compared to standard segmentation architectures, consistently delivering accurate predictions with a Dice coefficient of 0.891 and mIoU of 0.842. The coarse-to-fine design where segmentation operates on cropped ROIs rather than full images is the key architectural innovation that enables both speed and precision.

Additionally, the SAHI-enhanced Deep Scan mode addresses the persistent challenge of detecting small and flat polyps by employing multi-scale sliced inference with a dedicated small-polyp model and upgraded segmentation backbone. The designed web-based CAD interface enhances the accessibility and practicality of our AI-driven solutions, making them accessible to healthcare professionals regardless of their technical expertise. Implementing both localization and segmentation simultaneously in a unified pipeline can greatly reduce the risk of missing polyps, as segmentation will assist in polyp detection or vice versa if one modality fails. Cloud deployment via Docker and Hugging Face Spaces transforms the system from a research prototype into a universally accessible clinical support tool.

This study not only demonstrates the efficacy of the proposed hybrid approach but also establishes a framework for future research in the integration of advanced AI in medical imaging for colorectal cancer prevention.

REFERENCES

[1] R.L. Siegel, K.D. Miller, N.S. Wagle and A. Jemal, "Colorectal Cancer Statistics, 2023", *CA: A Cancer Journal for Clinicians*, Vol. 73, No. 3, pp. 233-254, 2023.

[2] J. Lewis, Y.J. Cha and J. Kim, "Dual Encoder-Decoder-based Deep Polyp Segmentation Network for Colonoscopy Images", *Scientific Reports*, Vol. 13, No. 1, pp. 1-8, 2023.

[3] A. Andren-Sandberg, "*Surgical Treatment: Evidence-Based and Problem-Oriented*", 2001.

[4] D.K. Rex, C.S. Cutler and G.T. Lemmel, "Colonoscopic Miss Rates of Adenomas Determined by Back-to-Back Colonoscopies", *Gastroenterology*, Vol. 112, No. 1, pp. 24-28, 1997.

[5] J.Y. Lee, J. Jeong and E.M. Song, "Real-Time Detection of Colon Polyps during Colonoscopy using Deep Learning: Systematic Validation with Four Independent Datasets", *Scientific Reports*, Vol. 10, No. 1, pp. 1-11, 2020.

[6] M. Liu, J. Jiang and Z. Wang, "Colonic Polyp Detection in Endoscopic Videos with Single Shot Detection based Deep Convolutional Neural Network", *IEEE Access*, Vol. 7, pp. 75058-75066, 2019.

[7] P. Sasmal, A. Paul, M.K. Bhuyan, Y. Iwahori and K. Kasugai, "Extraction of Key-Frames from Endoscopic Videos by using Depth Information", *IEEE Access*, Vol. 9, pp. 153004-153011, 2021.

[8] M. Kim and B.D. Lee, "A Simple Generic Method for Effective Boundary Extraction in Medical Image Segmentation", *IEEE Access*, Vol. 9, pp. 103875-103884, 2021.

[9] Geetha Sushama and G.C. Menon, "Flexible Colon Polyp Detection: A Dual Mode Approach for Detection and Segmentation of Colon Polyps with Optional Inpainting for Specular Highlight Mitigation", *SN Computer Science*, Vol. 5, No. 5, pp. 1-13, 2024.

[10] J.S.C. Reddy, C. Venkatesh, S. Sinha and S. Mazumdar, "Real Time Automatic Polyp Detection in White Light Endoscopy Videos using a Combination of YOLO and DeepSORT", *Proceedings of International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing*, Vol. 22, pp. 104-106, 2022.

[11] P. Ghose, A. Ghose, D. Sadhukhan, S. Pal and M. Mitra, "Improved Polyp Detection from Colonoscopy Images using Finetuned YOLO-v5", *Multimedia Tools and Applications*, Vol. 83, No. 14, pp. 42929-42954, 2023.

[12] Md. Al Amin, B.K. Paul and N.I. Bithi, "Real Time Detection and Localization of Colorectal Polyps from Colonoscopy Images: A Deep Learning Approach", *Proceedings of International Conference on Electrical and Computer Engineering*, Vol. 12, pp. 58-61, 2022.

[13] C.S. He, C.J. Wang, J.W. Wang and Y.C. Liu, "UY-NET: A Two-Stage Network to Improve the Result of Detection in Colonoscopy Images", *Applied Sciences*, Vol. 13, No. 19, pp. 1-13, 2023.

[14] P. Carrinho and G. Falcao, "Highly Accurate and Fast YOLOv4-based Polyp Detection", *Expert Systems with Applications*, Vol. 232, pp. 1-17, 2023.

[15] S. Wang, J. Xie, Y. Cui and Z. Chen, "Colorectal Polyp Detection Model by using Super-Resolution Reconstruction and YOLO", *Electronics*, Vol. 13, No. 12, pp. 1-9, 2024.

[16] M. Lalinia and A. Sahafi, "Colorectal Polyp Detection in Colonoscopy Images using YOLO-V8 Network", *Signal, Image and Video Processing*, Vol. 18, pp. 2047-2058, 2024.

[17] Z.F. Khan, "Real-Time Polyp Detection from Endoscopic Images using YOLOv8 with YOLO-Score Metrics for Enhanced Suitability Assessment", *IEEE Access*, Vol. 12, pp. 176346-176362, 2024.

[18] D. Jha, P.H. Smedsrud and M.A. Riegler, "ResUNet++: An Advanced Architecture for Medical Image Segmentation", *Proceedings of International Symposium on Multimedia*, Vol. 78, pp. 225-230, 2019.

[19] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Pattern Recognition and Image Processing*, Vol. 12, pp. 234-241, 2015.

[20] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation", *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support*, Vol. 18, pp. 3-11, 2018.

[21] M. Islam, V.S. Vibashan, V.J.M. Jose, N. Wijethilake, U. Utkarsh and H. Ren, "Brain Tumor Segmentation and Survival Prediction using 3D Attention UNet", *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Vol. 23, pp. 262-272, 2020.

[22] F.I. Diakogiannis, F. Waldner, P. Caccetta and C. Wu, "ResUNet-A: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data", *ISPRS Journal of*

Photogrammetry and Remote Sensing, Vol. 162, pp. 94-114, 2020.

[23] G. Sushama and G.C. Menon, "Attention Augmented Residual Autoencoder for Efficient Polyp Segmentation", *International Journal on Imaging Systems and Technology*, Vol. 33, No. 2, pp. 701-713, 2023.

[24] Geetha Sushama and G.C. Menon, "Colon Polyp Identification and Classification-Exploratory Study of Multivariate Feature Driven Optimized Heterogeneous Classifier", *Biomedical Engineering: Applications, Basis and Communications*, Vol. 38, No. 3, pp. 1-17, 2026.

[25] A. Lou, S. Guan, H. Ko and M.H. Loew, "CaraNet: Context Axial Reverse Attention Network for Segmentation of Small Medical Objects", *Journal of Medical Imaging*, Vol. 21, pp. 1-19, 2022.

[26] C.H. Huang, H.Y. Wu and Y.L. Lin, "HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS", *Computer Vision and Pattern Recognition*, Vol. 23, pp. 1-13, 2021.

[27] S.T. Tran, M.H. Nguyen, H.P. Dang and T.T. Nguyen, "Automatic Polyp Segmentation using Modified Recurrent Residual Unet Network", *IEEE Access*, Vol. 10, pp. 65951-65961, 2022.

[28] A. Galdran, G. Carneiro and M.A.G. Ballester, "Double Encoder-Decoder Networks for Gastrointestinal Polyp Segmentation", *ICPR International Workshops and Challenges*, Vol. 12, pp. 293-307, 2021.

[29] D.P. Fan, G. Ji, T. Zhou, G. Chen, H. Fu, J. Shen and L. Shao, "PraNet: Parallel Reverse Attention Network for Polyp Segmentation", *Proceedings of International Conference on Image and Video Processing*, Vol. 8, pp. 263-273, 2020.

[30] B. Dong, W. Wang, D.P. Fan, J. Li, H. Fu and L. Shao, "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 56, pp. 1-17, 2021.

[31] M.F. Ahamed, M.R. Islam, M. Nahiduzzaman, M.J. Karim, M.A. Ayari and A. Khandakar, "Automated Detection of Colorectal Polyp Utilizing Deep Learning Methods with Explainable AI", *IEEE Access*, Vol. 12, pp. 78074-78100, 2024.

[32] D. Jha, P.H. Smedsrud and M.A. Riegler, "Kvasir-SEG: A Segmented Polyp Dataset", *MultiMedia Modeling*, Vol. 76, pp. 451-462, 2019.

[33] J. Bernal, F.J. Sanchez, G. Fernandez-Esparrach, D. Gil, C. Rodríguez and F. Vilarino, "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians", *Computerized Medical Imaging and Graphics*, Vol. 43, pp. 99-111, 2015.

[34] G. Jocher, A. Chaurasia and J. Qiu, "YOLO by Ultralytics", Available at: https://github.com/ultralytics/ultralytics, Accessed in 2023.

[35] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 23, pp. 770-778, 2020.

[36] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression", *Proceedings of International Conference on Artificial Intelligence*, Vol. 36, No. 7, pp. 12993-13000, 2020.

[37] F.C. Akyon, S.O. Altinuc and A. Temizel, "Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection", *Proceedings of International Conference on Image Processing*, Vol. 76, pp. 966-970, 2022.

[38] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *NeurIPS*, Vol. 28, pp. 1-12, 2015.

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A.C. Berg, "SSD: Single Shot Multibox Detector", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 32, pp. 21-37, 2016.

[40] W. Lv, S. Xu and Y. Zhao, "DETRs Beat YOLOs on Real-Time Object Detection", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 23, pp. 16965-16974, 2024.

[41] T.Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal Loss for Dense Object Detection", *Proceedings of International Conference on Computer Vision*, Vol. 32, pp. 2999-3007, 2017.

[42] S. Ali, "Assessing Generalisability of Deep Learning-based Polyp Detection and Segmentation Methods through a Computer Vision Challenge", *Scientific Reports*, Vol. 14, pp. 1-17, 2022.

[43] P. Yakubovskiy, "Segmentation Models PyTorch", Available at: https://github.com/qubvel/segmentation_models.pytorch, Accessed in 2019.

[44] J. Le'Clerc Arrastia, N. Heilenkötter and D. Otero Baguer, "Deeply Supervised U-Net for Semantic Segmentation to Assist Dermatopathological Assessment of Basal Cell Carcinoma", *Machine Learning with Applications*, Vol. 6, pp. 1-11, 2021.