

ORIGINAL RESEARCH ARTICLE

Real-time You Only Look Once-based colorectal polyp detection for cancer screening across independent colonoscopy datasets

Ishak Pacal^{1,2,3*} ¹Department of Computer Engineering, Faculty of Engineering, Igdir University, Igdir, Türkiye²Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Fenerbahçe University, Istanbul, Türkiye³Department of Electronics and Information Technologies, Faculty of Architecture and Engineering, Nakhchivan State University, Nakhchivan, Azerbaijan

Abstract

Introduction: Colorectal cancer remains a major cause of cancer-related mortality, although many cases are preventable when precursor polyps are detected and removed during colonoscopy. Real-time computer-aided detection may assist endoscopists, but reliable localization is challenging because polyps vary in size, shape, color, contrast, boundary clarity, and visibility across heterogeneous endoscopic conditions.

Objective: This study benchmarked recent You Only Look Once (YOLO)-based detectors for single-class colorectal polyp localization by comparing architectural families and model scales ranging from nano to x-large.

Methods: YOLO12, YOLOv13, and YOLO26 models were evaluated using PolypGen2021, Polyp Imaging Classification for Colonoscopy using Optical Learning Outcomes (PICCOLO), and ETIS-Larib. PolypGen2021 was split at the patient level for model development; the predefined patient-level PICCOLO split was preserved, and ETIS-Larib was reserved exclusively for external testing. Performance was assessed using precision, recall, mean average precision (mAP) 50, mAP50–95, parameter count, giga floating-point operations per second (GFLOPs), and inference time.

Results: On ETIS-Larib, YOLO26x achieved the highest mAP50–95 (0.6475) and the highest recall (0.7736). On the PICCOLO test set, YOLOv13x achieved the highest mAP50–95 of 0.6630, whereas YOLO26x reached the highest precision of 0.9118. Across both test sets, YOLO26x achieved the highest average mAP50–95 (0.6462). YOLO26n provided the most efficient profile, with 2.3750 million parameters, 5.1894 GFLOPs, and inference times below 3 ms per image.

Conclusion: YOLO26x was the strongest accuracy-oriented detector, while YOLO26n offered the most practical lightweight alternative for real-time colonoscopy support.

Keywords: Colorectal cancer; Colonoscopy; Polyp detection; You Only Look Once; Deep learning; Computer-aided detection

***Corresponding author:**Ishak Pacal
(ishakpacal@igdir.edu.tr)

Citation: Pacal I. Real-time You Only Look Once-based colorectal polyp detection for cancer screening across independent colonoscopy datasets. *Eurasian J Med Oncol*. doi: 10.36922/EJMO026220238

Received: May 26, 2026**Revised:** June 27, 2026**Accepted:** July 3, 2026**Published online:** July 16, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Colorectal cancer remains a major contributor to cancer incidence and cancer-related mortality worldwide.^{1,2} Most sporadic colorectal cancers develop gradually from

precursor lesions through well-characterized neoplastic pathways, which makes early detection and removal of polyps a clinically meaningful opportunity for prevention.³ Colonoscopic polypectomy has been associated with a sustained reduction in colorectal cancer incidence and mortality, supporting the central role of colonoscopy in screening and surveillance programs.⁴

The effectiveness of colonoscopy depends not only on access to screening but also on the quality of mucosal inspection. Adenoma detection rate is one of the most important quality indicators, and even modest differences in detection performance are associated with clinically relevant differences in interval cancer risk.^{5,6} Missed lesions remain a persistent problem, particularly when polyps are small, flat, pale, serrated, partially hidden behind folds, or visually similar to the surrounding mucosa.^{7,8} Endoscopic visibility is further affected by bowel preparation, withdrawal technique, camera motion, illumination, bubbles, residual fluid, mucus, specular reflections, and endoscopist fatigue.⁹ These sources of variability pose a difficult visual recognition problem and have driven sustained interest in computer-aided detection systems to support real-time colonoscopy.

Artificial intelligence-assisted colonoscopy has moved from experimental image analysis toward clinical evaluation. Prospective studies have reported improved adenoma or polyp detection with computer-aided detection, although the measured benefit may depend on screening indication, withdrawal time, endoscopist behavior, adjunctive imaging techniques, and the baseline quality of colonoscopy.^{10,11} From a technical perspective, this clinical setting requires algorithms that are not only accurate but also fast, stable, and capable of operating across heterogeneous endoscopic appearances. Earlier computer-aided approaches relied heavily on handcrafted features, including color, texture, shape, edge, and geometric descriptors.^{12,13} Such methods provided useful proof of concept, but their robustness was limited by variation in illumination, mucosal background, lesion morphology, and imaging devices. Deep learning reduced this dependence on manually designed features by learning hierarchical representations directly from image data and quickly became the dominant approach in colonoscopy image analysis.^{14,15}

Public datasets have been essential for this progress because they allow methods to be evaluated under more transparent conditions. ETIS-Larib, computer vision center-derived benchmarks, Kvasir-SEG, Polyp Imaging Classification for Colonoscopy using Optical Learning Outcomes (PICCOLO), and PolypGen have contributed to the development and validation of automated polyp analysis systems.^{12,16-19} These datasets differ in acquisition

centers, imaging modes, annotation protocols, lesion appearances, frame selection, and test-set composition. Such diversity is valuable, but it also complicates direct comparison across studies. A detector that performs strongly on one benchmark may not preserve the same behavior when evaluated on an independent distribution.

Object detection models are especially relevant for real-time colonoscopy because they provide immediate lesion localization without requiring dense pixel-level annotation. The You Only Look Once (YOLO) family has been widely adopted in this setting because it performs detection in a single forward pass and offers a favorable balance between speed and accuracy.²⁰⁻²² Several YOLO-based studies have improved colorectal polyp detection through transfer learning, negative-sample training, activation-function optimization, hyperparameter search, feature-pyramid refinement, and attention-based feature enhancement.²³⁻²⁸ More recent work has explored lightweight attention mechanisms, wavelet-enhanced reverse attention, custom detection necks, real-time transformer-based detectors, and synthetic data augmentation to improve small-lesion sensitivity and robustness in complex backgrounds.²⁹⁻³³ These studies show the rapid evolution of the field, but they also make the literature harder to interpret because reported gains may reflect differences in architecture, model scale, dataset choice, training protocol, or evaluation setting.

Segmentation studies provide a complementary view of the same problem. Boundary-aware and U-Net-derived models, lightweight generative adversarial network-based approaches, adaptive feature-fusion networks, and efficient segmentation architectures have improved pixel-level delineation of polyps under challenging conditions.³⁴⁻³⁶ These methods are important when precise contour estimation is required. Detection, however, remains a distinct and clinically practical task: the immediate goal is to alert the endoscopist to the location of a suspicious lesion with sufficiently low latency. For this reason, bounding-box detection remains a relevant benchmark, particularly when localization accuracy is evaluated alongside computational cost.

Several methodological issues remain insufficiently resolved in current polyp detection benchmarks. First, many studies evaluate a single modified detector against selected baselines, making it difficult to separate the detector's contribution from model capacity or training configuration. Second, within-family scaling is often treated as secondary, although nano, small, medium, large, and x-large variants differ substantially in memory demand, latency, and deployment feasibility. Third, external testing is still not applied consistently, even

though cross-dataset behavior is one of the most important indicators of practical robustness. Finally, mean average precision (mAP) 50 alone may be too permissive for medical object detection, as a predicted box can satisfy the 0.50 Intersection over Union (IoU) threshold even if it only partially covers the lesion. mAP50–95 provides a stricter localization estimate by averaging performance across multiple IoU thresholds and is therefore more informative when spatial precision matters.

The present benchmark was designed to examine how recent YOLO detector families behave when trained and evaluated under the same protocol for single-class colorectal polyp localization. YOLO12, YOLOv13, and YOLO26 were compared across available model scales to assess both family-level differences and capacity-dependent behavior.^{37–39} PolypGen2021 and PICCOLO were used for model development, with patient-level splitting applied to PolypGen2021 and the predefined patient-level PICCOLO split preserved. ETIS-Larib was reserved exclusively for external testing. The evaluation combined precision, recall, mAP50, and mAP50–95 with parameter count, giga floating-point operations per second (GLOPs), and inference time. This design was intended to provide a standardized benchmark rather than to introduce a new detector architecture or to infer a macroscopic or histological polyp subtype. The main contributions of this work are as follows:

- (i) A unified benchmark of YOLO12, YOLOv13, and YOLO26 is provided for colorectal polyp localization under a shared experimental protocol, reducing the influence of inconsistent training and evaluation settings.
- (ii) Model scales from nano to x-large are evaluated within the same framework, allowing accuracy gains to be interpreted together with computational cost, latency, and deployment feasibility.
- (iii) The experimental design uses public datasets with clearly separated roles, including patient-level development data, a held-out patient-level test subset, and an external test set.
- (iv) mAP50–95 is used as the primary ranking metric to emphasize strict localization quality across multiple IoU thresholds, with mAP50, precision, and recall reported as complementary detection indicators.
- (v) The benchmark distinguishes between accuracy-oriented and lightweight deployment-oriented choices, which is essential for real-time colonoscopy systems where the largest model may not always be the most practical option.

2. Materials and methods

2.1. Datasets

Three publicly available colorectal polyp datasets were used in this benchmark: PolypGen2021, PICCOLO, and ETIS-Larib.^{12,18,19} The experiments were formulated as single-class object detection, and the target class was defined as polyp. The detection subsets used in the reported experiments consisted of images with visible polyp bounding-box annotations. Negative-only images without polyp bounding-box annotations were not included in the evaluated subsets, because the objective was to compare localization performance for the annotated polyp class under a consistent protocol.

PolypGen2021 was used as one of the model development datasets. Since patient-level information was available, the dataset was divided into training and validation subsets at the patient level. This strategy was preferred over image-level random splitting because multiple colonoscopy frames from the same patient may contain visually correlated lesions, repeated camera views, similar mucosal texture, or near-duplicate appearances. The final PolypGen2021 split contained 1,342 training images and 194 validation images with polyp annotations.

The PICCOLO dataset was included because it provides a predefined patient-level split. Its original training, validation, and test partitions were preserved without reshuffling. The PICCOLO training subset contained 2,203 polyp-annotated images, the validation subset contained 897 polyp-annotated images, and the test subset contained 333 polyp-annotated images. The predefined test subset was not used during training, validation, early stopping, threshold inspection, or model selection, and was treated as an independent held-out test set.

ETIS-Larib was used only for final external testing. It contained 196 polyp-annotated images and was not included in any stage of model development. This design enabled an additional assessment of cross-dataset generalization under visual conditions that differ from the training and validation data. The datasets were used as still-image detection datasets; no video-level frame extraction or temporal sampling was performed by the authors. Representative examples from PolypGen2021, PICCOLO, and ETIS-Larib are presented in [Figure 1](#).

2.2. Data split strategy

The data split strategy was designed to reduce leakage and provide a realistic estimate of model generalization. In

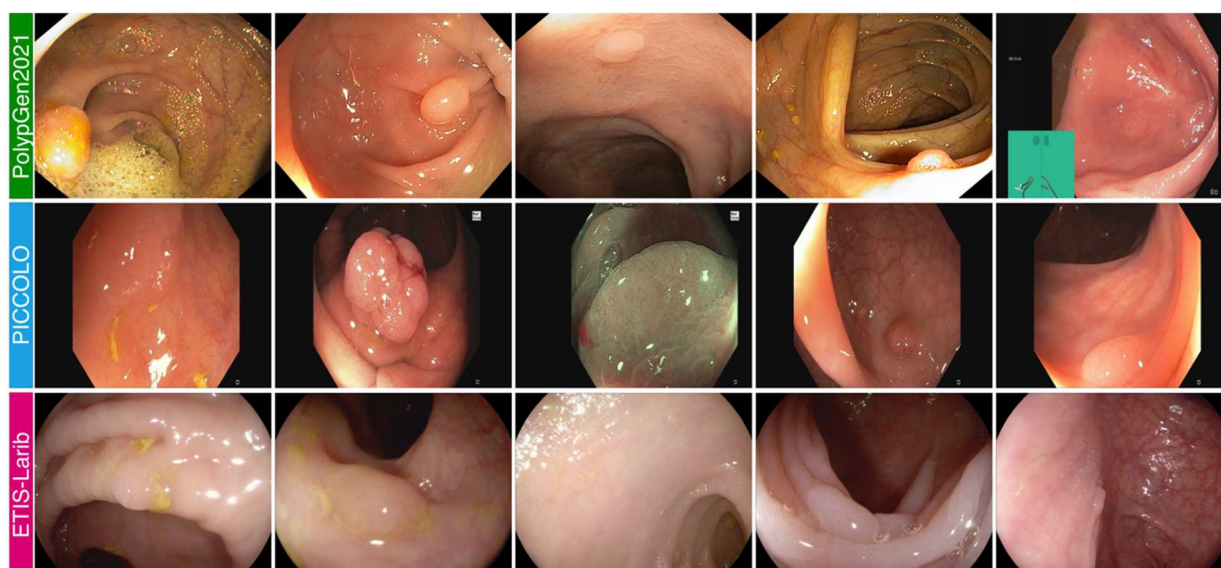


Figure 1. Representative examples from the colorectal polyp datasets used in this study. Representative colonoscopy images from PolypGen2021, PICCOLO, and ETIS-Larib are shown. The examples illustrate differences in polyp size, morphology, illumination, boundary visibility, mucosal texture, and background complexity.

colonoscopy datasets, image-level random partitioning can inflate performance because frames from the same patient or video may share highly similar lesion appearance, camera trajectory, background texture, and illumination. To reduce this risk, PolypGen2021 was split at the patient level, and the predefined patient-level split of PICCOLO was preserved.

The final model development data consisted of the training and validation subsets from PolypGen2021 and PICCOLO. Two test settings were then used for final evaluation. The first was the predefined PICCOLO test subset, which provided a held-out evaluation within the official dataset structure. The second was ETIS-Larib, which was kept completely independent from model development and served as an external test set. No image from the PICCOLO test or ETIS-Larib was used during training, validation, early stopping, hyperparameter selection, confidence threshold inspection, or model selection. The dataset distribution and experimental role of each dataset are summarized in Table 1.

2.3. You Only Look Once model families and architectural rationale

The evaluated detector families were selected because they represent three recent and distinct directions in YOLO development: attention-centric real-time detection in YOLO12, hypergraph-enhanced adaptive visual

perception in YOLOv13, and deployment-oriented end-to-end detection in YOLO26.³⁷⁻³⁹

2.3.1. General You Only Look Once detection formulation

The YOLO family formulates object detection as a single-stage prediction problem. Instead of generating region proposals in a separate stage, YOLO detectors estimate bounding boxes, objectness scores, and class probabilities directly from the input image. This design is well suited to colonoscopy-based computer-aided detection because the model must localize suspicious polyp regions with low latency while preserving sensitivity to small, low-contrast, or partially visible lesions.

For an input colonoscopy image I , a YOLO detector with trainable parameters θ produces a set of candidate detections (**Equation 1**):

$$\mathcal{D} = f_{\theta}(I) = \{(\hat{b}_i, \hat{s}_i, \hat{c}_i)\}_{i=1}^N \quad (1)$$

where $\hat{b}_i$ denotes the predicted bounding box, $\hat{s}_i$ is the confidence score, $\hat{c}_i$ is the predicted class label, and N is the number of candidate detections. Since the present benchmark was formulated as single-class object detection, all positive annotations were assigned to the polyp class. The detector output therefore represents the localization of

Table 1. Dataset distribution and experimental role

Dataset	Experimental role	Training images with polyp annotations	Validation images with polyp annotations	Test images with polyp annotations	Target class	Split strategy
PolypGen2021	Model development	1,342	194	–	Polyp	Patient-level split
PICCOLO	Model development and held-out testing	2,203	897	333	Polyp	Predefined patient-level split
ETIS-Larib	External testing	–	–	196	Polyp	Test-only dataset

Note: The benchmark was formulated as single-class polyp detection. The table reports images with polyp bounding-box annotations used in each split. Negative-only images without polyp annotations were not included in the evaluated detection subsets.

a visible polyp candidate rather than macroscopic subtype classification or histological prediction.

Spatial agreement between a predicted bounding box and a ground-truth bounding box was measured using IoU (Equation 2):

$$IoU(\hat{b}, b) = \frac{|\hat{b} \cap b|}{|\hat{b} \cup b|} \quad (2)$$

where $\hat{b}$ is the predicted bounding box and b is the corresponding ground-truth bounding box. A higher IoU indicates stronger spatial overlap between the predicted and annotated lesion regions. This measure is central to the calculation of mAP50 and mAP50–95, because a detection is considered correct only when its overlap with the ground truth satisfies the relevant IoU threshold.

During training, YOLO models optimize a multi-component detection objective that combines localization, classification, and regression refinement terms. The general form of the loss can be written as Equation 3:

$$\mathcal{L}_{total} = \lambda_{box} \mathcal{L}_{box} + \lambda_{cls} \mathcal{L}_{cls} + \lambda_{reg} \mathcal{L}_{reg} \quad (3)$$

where $\mathcal{L}_{box}$ penalizes bounding-box localization error, $\mathcal{L}_{cls}$ penalizes class prediction error, and $\mathcal{L}_{reg}$ represents the architecture-specific regression refinement term used by the implementation. The coefficients λ_{box} , λ_{cls} , and λ_{reg} control the relative contribution of these loss components.

These coefficients were not optimized separately in this study. They were kept fixed according to the training configuration used by the YOLO implementation to maintain a consistent comparison across detector families and model scales. Specifically, the loss weights were set

to 7.5 for bounding-box loss, 0.5 for classification loss, and 1.5 for the regression refinement term. No separate grid search, Bayesian optimization, or dataset-specific tuning was performed for these coefficients. This choice was aligned with the benchmark design, where the aim was to compare recent YOLO families under a shared protocol rather than to obtain an individually optimized configuration for each architecture.

The benchmark included both cross-family and within-family comparisons. YOLO12, YOLOv13, and YOLO26 were evaluated as recent YOLO detector families, and each family was examined across available capacity scales. YOLO12 included nano, small, medium, large, and x-large variants. YOLOv13 included nano, small, large, and x-large variants because the medium checkpoint was not included in the evaluated model set. YOLO26 included nano, small, medium, large, and x-large variants. This design allowed the analysis to capture differences between YOLO generations as well as the effect of increasing model capacity within the same architectural family.

2.3.2. You Only Look Once 26 family

The YOLO26 family was included as the most recent detector family in the benchmark and was selected for detailed visualization because it provided the strongest cross-dataset balance in the final analysis [39]. The family is designed for deployment-oriented object detection, with emphasis on efficient inference, simplified prediction flow, and practical use in real-time systems. These properties are relevant for colonoscopy, where computer-aided detection must operate continuously and with minimal delay.

Five YOLO26 variants were evaluated: YOLO26n, YOLO26s, YOLO26m, YOLO26l, and YOLO26x. These variants represent a capacity progression from nano to x-large. The nano variant was expected to provide the most

favorable computational profile, whereas the x-large variant provided the highest representational capacity. Evaluating the full-scale range was important because the best model for clinical deployment may not be the largest model. A smaller detector can be preferable if it retains adequate sensitivity while reducing latency and computational demand. A general overview of the YOLO26-based detector is provided in Figure 2. Figure 2 is intended to describe the backbone, neck, and detection head used by the detector at a high level, rather than to present a newly proposed architecture.

2.3.3. You Only Look Once v13 family

The YOLOv13 family was included because it represents a correlation-enhanced detection direction within recent YOLO development.³⁸ Its design emphasizes richer feature interaction and adaptive correlation modeling, which may be useful in medical object detection where lesion boundaries are subtle, and background structures can resemble target objects. In colonoscopy frames, local texture alone may not be sufficient to separate true polyps from folds, reflections, bubbles, or mucosal irregularities.

Four YOLOv13 variants were evaluated: YOLOv13n, YOLOv13s, YOLOv13l, and YOLOv13x. These variants

allowed scale-wise analysis within the family. The x-large version provided the highest capacity and achieved strong performance on the PICCOLO test subset, whereas the smaller variants allowed assessment of whether the architectural design remained competitive under more efficient configurations.

2.3.4. You Only Look Once 12 family

The YOLO12 family was included as an attention-oriented real-time detection family.³⁷ Attention-based feature refinement is relevant in colonoscopy because polyps may occupy small regions and may be distinguished from surrounding mucosa only by subtle texture, contour, or color changes. At the same time, attention mechanisms must remain computationally efficient for real-time medical imaging.

The evaluated YOLO12 variants were YOLO12n, YOLO12s, YOLO12m, YOLO12l, and YOLO12x. This enabled an internal capacity comparison from lightweight to high-capacity models. The inclusion of YOLO12 also allowed attention-oriented detection behavior to be compared with the correlation-enhanced YOLOv13 family and the deployment-oriented YOLO26 family under the same dataset and evaluation protocol.

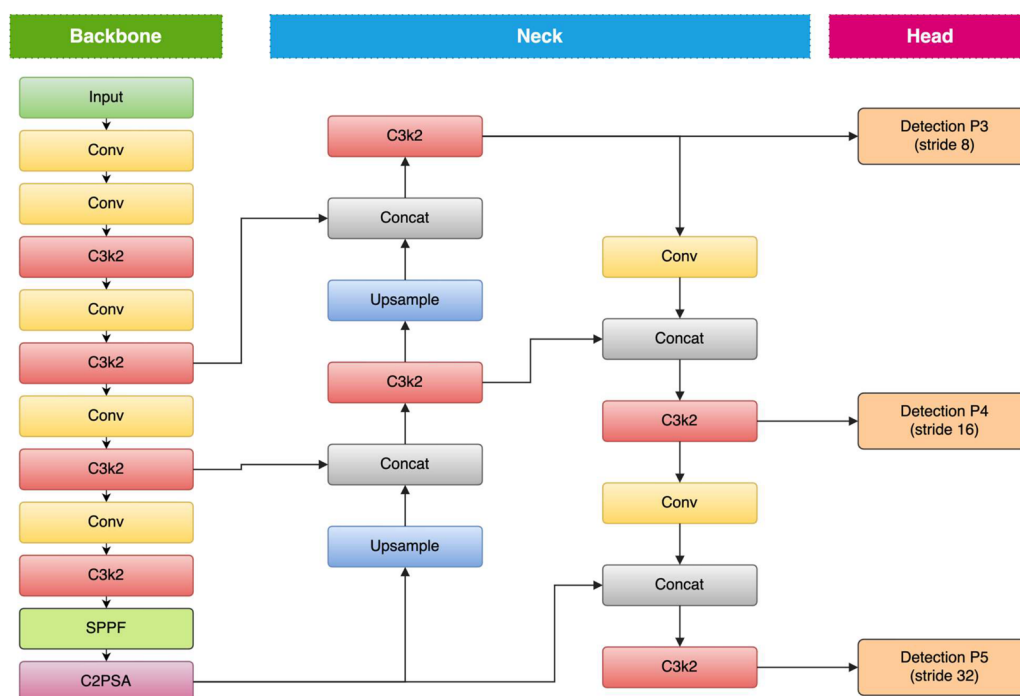


Figure 2. General architecture of the YOLO26-based detector used for colorectal polyp detection. The detector follows a single-stage design composed of a feature extraction backbone, a multi-scale feature aggregation neck, and a detection head that predicts polyp bounding boxes and confidence scores. The figure provides an architectural overview of the YOLO26 model family evaluated in this benchmark.

2.4. Transfer learning, training protocol, and data augmentation

All models were trained for single-class colorectal polyp detection using bounding-box annotations in YOLO format. Transfer learning was performed by initializing each detector with its corresponding pretrained checkpoint, followed by full-network fine-tuning on colonoscopy images.⁴⁰ No layers were frozen during training, allowing both low-level visual features and higher-level lesion representations to adapt to the endoscopic domain.

Training was conducted on a workstation equipped with an NVIDIA GeForce RTX 5090 GPU, 64 GB RAM, and an Intel Core Ultra 7 processor. The experiments were implemented using the Ultralytics framework. Deterministic training was enabled with a fixed random seed of 0 to improve reproducibility across model runs.

The input image size was set to 640×640 pixels, and the batch size was 16. No manual image enhancement, denoising, contrast correction, artifact removal, or dataset-specific preprocessing was applied before training. Images were processed through the standard YOLO input pipeline, with aspect-ratio-preserving letterbox resizing applied when required. Pixel-value normalization and tensor conversion were handled internally by the implementation. Training was allowed to continue for up to 400 epochs, with early stopping patience set to 100 epochs. Validation was performed during training using the predefined validation split.

The optimizer was selected automatically by the training framework. The initial learning rate was 0.01, the final learning-rate factor was 0.01, momentum was 0.937, and weight decay was 0.0005. Warm-up training was applied for three epochs, with warm-up momentum of 0.8 and warm-up bias learning rate of 0.1. Mixed precision training was enabled, eight data-loading workers were used, and the maximum number of detections per image was set to 300. The IoU threshold used for non-maximum suppression during validation was set to 0.7. The loss coefficients were kept fixed as described in Section 2.3.1 and were not separately optimized for individual model families.

Data augmentation was applied only to the training images after the dataset split stage. Validation and test images were not augmented and were kept unchanged apart from standard model-input formatting. The augmentation policy included moderate hue, saturation, and value adjustments at 0.015, 0.7, and 0.4, respectively. Translation was set to 0.1, scale augmentation to 0.5, and horizontal flipping to 0.5. Mosaic augmentation was enabled during training and disabled during the final ten epochs to

stabilize late-stage optimization. RandAugment was used as the automatic augmentation policy, and random erasing was set to 0.4.

Rotation, vertical flipping, shear, perspective transformation, MixUp, CutMix, and copy-paste augmentation were not used. These operations were excluded because aggressive geometric or compositional transformations may distort lesion boundaries or create anatomically implausible polyp-background relationships in colonoscopy images. The augmentation strategy was kept moderate to improve visual diversity while preserving clinically plausible lesion appearance.

2.5. Evaluation metrics

Model performance was evaluated using detection accuracy metrics and computational efficiency indicators. This combined evaluation was necessary because a clinically useful polyp detector should not only localize lesions accurately but also operate with feasible latency and computational cost. Precision was used to measure the proportion of predicted polyp detections that were correct, as shown in **Equation 4**:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

where TP denotes true-positive detections and FP denotes false-positive detections. High precision indicates that the detector produces fewer false alarms. Recall was used to measure the proportion of true polyps that were detected, as shown in **Equation 5**:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

where FN denotes false-negative detections. Recall is particularly important in colonoscopy because a false-negative result corresponds to a missed polyp. Average precision was calculated from the precision-recall curve, as shown in **Equation 6**:

$$AP = \int_0^1 P(R) dR \quad (6)$$

where $P(R)$ denotes precision as a function of recall. Since this was a single-class detection task, the mean average precision corresponded to the average precision of the polyp class. mAP50 was defined as average precision at an IoU threshold of 0.50, as shown in **Equation 7**:

$$mAP50 = AP_{0.50} \quad (7)$$

mAP50-95 was calculated by averaging average

precision values across IoU thresholds from 0.50 to 0.95 with a step size of 0.05, as shown in Equation 8:

$$mAP_{50-95} = \frac{1}{10} \sum_{t \in \{0.50, 0.55, \dots, 0.95\}} AP_t \quad (8)$$

mAP50–95 was used as the primary ranking metric because it evaluates localization quality under increasingly strict overlap thresholds. This is important for polyp detection because a model should not only identify suspicious regions but also provide spatially reliable bounding boxes around the lesion. Computational efficiency was assessed using parameter count, GFLOPs, and inference time per image. Parameter count reflected model size, GFLOPs estimated computational complexity, and inference time indicated practical speed under the evaluation setting. These indicators were interpreted together with detection metrics to distinguish the best accuracy-oriented detector from the most efficient lightweight alternative.

3. Results

3.1. Benchmark setting and evaluated models

The experimental results were analyzed across two independent test settings to assess both detection accuracy and generalization behavior. The evaluated detectors included YOLO12, YOLOv13, and YOLO26 model families. Within these families, available model scales from lightweight to high-capacity versions were included, allowing both architecture-level and scale-level comparisons. YOLO12 was evaluated with nano, small, medium, large, and x-large variants. YOLOv13 was evaluated with nano, small, large, and x-large variants, because the medium checkpoint was not included in the evaluated model set. YOLO26 was evaluated with nano, small, medium, large, and x-large variants.

The final evaluation was performed on the ETIS-Larib external test set and the PICCOLO held-out test set. ETIS-Larib was not used during training, validation, early stopping, threshold inspection, or model selection, and therefore provided a strict external test setting. The PICCOLO test subset was also excluded from model development and was used according to its predefined patient-level split. Detection performance was assessed using precision, recall, mAP50, and mAP50–95, while model efficiency was examined using parameter count, GFLOPs, and inference time per image. Since mAP50–95 evaluates localization quality across multiple intersection-over-union thresholds, it was used as the primary ranking metric.

3.2. Detection performance on the ETIS-Larib external test set

The ETIS-Larib dataset provided the most stringent test condition because it was kept completely separate from model development. This dataset is visually challenging because polyps may appear under variable illumination, low contrast, mucosal folds, bubbles, reflections, and heterogeneous background texture. The quantitative results on ETIS-Larib are presented in Table 2.

On ETIS-Larib, YOLO26x achieved the highest mAP50–95 value of 0.6475, indicating the strongest localization performance under strict IoU-based evaluation. The same model also obtained the highest recall, 0.7736, which is important for polyp screening because missed lesions represent a clinically relevant failure. These findings suggest that the x-large YOLO26 model was more sensitive to externally tested lesions and preserved better spatial localization across multiple IoU thresholds.

YOLO26n showed a different but equally important profile. It achieved the highest precision (0.8792) and the highest mAP50 (0.8507), while also having the smallest parameter count (2.3750 M), the lowest computational cost (5.1894 GFLOPs), and the fastest inference time (2.8929 ms/img). This result is notable because the smallest YOLO26 model outperformed larger models in several practical indicators. It suggests that compact detector design can remain highly competitive when the target class is well-defined, and the training data are sufficiently focused.

The lowest mAP50–95 value on ETIS-Larib was obtained by YOLOv13x at 0.5805. This is an important observation because YOLOv13x was one of the largest models in the benchmark. The result indicates that a larger capacity does not automatically translate into better external test behavior. In polyp detection, cross-dataset performance is influenced not only by model size but also by how feature representations respond to subtle lesion boundaries, mucosal texture, illumination changes, and background artifacts.

3.3. Detection performance on the PICCOLO held-out test set

The PICCOLO test subset provided a complementary evaluation setting. Unlike ETIS-Larib, PICCOLO includes predefined patient-level train, validation, and test partitions. The test subset was not used during model development and therefore served as an independent held-out evaluation within the same benchmark source. The results on the PICCOLO test set are reported in Table 3.

Table 2. Detection performance of You Only Look Once models on the ETIS-Larib external test set

Model	Precision	Recall	mAP50	mAP50–95	Params (M)	GFLOPs	Time (ms/img)
YOLO12n	0.8424	0.7217	0.8100	0.6117	2.5569	6.3199	3.5111
YOLO12s	0.7814	0.7075	0.7915	0.5886	9.1944	21.1518	4.1464
YOLO12m	0.8690	0.6557	0.7973	0.6245	20.1057	67.1149	3.8531
YOLO12l	0.8140	0.6811	0.7815	0.6143	26.3398	88.5459	6.2124
YOLO12x	0.8643	0.7212	0.8316	0.6358	59.0445	198.5223	7.5016
YOLOv13n	0.7117	0.7102	0.7578	0.5834	2.4481	6.2021	5.6854
YOLOv13s	0.8142	0.7264	0.8162	0.6210	9.5297	21.2898	5.7336
YOLOv13l	0.7120	0.7075	0.7739	0.5903	27.5136	88.0929	8.6126
YOLOv13x	0.7881	0.6842	0.7695	0.5805	63.8863	198.7226	9.0252
YOLO26n	0.8792 ^a	0.7211	0.8507 ^a	0.6364	2.3750 ^a	5.1894 ^a	2.8929 ^a
YOLO26s	0.8010	0.6415	0.7713	0.6103	9.4656	20.5225	2.9779
YOLO26m	0.8603	0.6679	0.7902	0.6205	20.3502	67.8484	3.3063
YOLO26l	0.8278	0.7075	0.8078	0.6356	24.7465	86.0955	4.5278
YOLO26x	0.8333	0.7736 ^a	0.8470	0.6475 ^a	55.6347	193.3677	5.9053

Note: ^a indicates the best result in each column. For Params, GFLOPs, and Time, lower values are better.

Abbreviations: GFLOPs: Giga floating-point operations; mAP: Mean average precision; Params: Parameters; YOLO: You Only Look Once.

The PICCOLO results followed a different pattern from ETIS-Larib. YOLOv13x achieved the highest mAP50–95 (0.6630) and the highest mAP50 (0.8286). This indicates that YOLOv13x localized PICCOLO test polyps more accurately than the other evaluated models. YOLOv13n achieved the highest recall (0.7238), showing that a smaller YOLOv13 variant was more sensitive to lesion retrieval in this dataset.

YOLO26x reached the highest precision on PICCOLO, with 0.9118, but its recall was lower than that of several other models. This suggests a more conservative detection behavior. The model produced fewer false-positive predictions, but this selectivity came with reduced lesion retrieval. In screening-oriented colonoscopy, this trade-off should be interpreted carefully, because recall is critical when the main objective is to reduce missed polyps.

Efficiency results again favored YOLO26n. It required only 2.3750 M parameters and 5.1894 GFLOPs, with an inference time of 2.6825 ms/img. In contrast, YOLOv13x achieved the best PICCOLO mAP50–95 but required 63.8863 M parameters and 198.7226 GFLOPs. This

contrast emphasizes that the best detection score and the most deployable model may not be the same. For resource-limited or real-time settings, compact models such as YOLO26n may offer a more practical balance.

3.4. Cross-dataset ranking based on mAP50–95

Since the best-performing model differed between ETIS-Larib and PICCOLO, a cross-dataset ranking was calculated to avoid selecting a model based on a single test set. The mAP50–95 values from both test sets were averaged for each model. This ranking is presented in [Table 4](#) and provides a more balanced estimate of external and held-out test performance.

The cross-dataset ranking identified YOLO26x as the most balanced detector, with the highest average mAP50–95 of 0.6462. This result supports its selection for detailed curve-based and qualitative analysis. Although YOLOv13x was the strongest model on PICCOLO, its weaker ETIS-Larib result reduced its average ranking. This confirms that a single test dataset can lead to a different conclusion than a cross-dataset evaluation.

Table 3. Detection performance of YOLO models on the PICCOLO held-out test set

Model	Precision	Recall	mAP50	mAP50–95	Params (M)	GFLOPs	Time (ms/img)
YOLO12n	0.8034	0.7023	0.7758	0.6102	2.5569	6.3199	3.2992
YOLO12s	0.8251	0.6771	0.7823	0.6166	9.2313	21.2108	3.3361
YOLO12m	0.8906	0.6536	0.8040	0.6480	20.1057	67.1149	3.6232
YOLO12l	0.8127	0.7057	0.7882	0.6197	26.3398	88.5459	5.9319
YOLO12x	0.8674	0.6458	0.7824	0.6274	59.0445	198.5223	6.5566
YOLOv13n	0.8176	0.7238 ^a	0.8026	0.6507	2.4481	6.2021	5.4562
YOLOv13s	0.8502	0.6693	0.7858	0.6238	9.5297	21.2898	5.4445
YOLOv13l	0.8693	0.6582	0.7960	0.6258	27.5136	88.0929	8.4023
YOLOv13x	0.8743	0.6901	0.8286 ^a	0.6630 ^a	63.8863	198.7226	8.5792
YOLO26n	0.8710	0.6327	0.7667	0.6265	2.3750 ^a	5.1894 ^a	2.6825 ^a
YOLO26s	0.8625	0.6207	0.7807	0.6276	9.4656	20.5225	2.7534
YOLO26m	0.8489	0.6615	0.7889	0.6526	20.3502	67.8484	3.0749
YOLO26l	0.8419	0.6641	0.7780	0.6358	24.7465	86.0955	4.1674
YOLO26x	0.9118 ^a	0.6380	0.7885	0.6449	55.6347	193.3677	5.6760

Note: ^a indicates the best result in each column. For Params, GFLOPs, and Time, lower values are better.

Abbreviations: GFLOPs: Giga floating-point operations; mAP: Mean average precision; Params: Parameters; YOLO: You Only Look Once.

YOLO26m and YOLO12m followed YOLO26x with average mAP50–95 values of 0.6366 and 0.6363, respectively. YOLO26l also showed stable behavior, with nearly identical values on ETIS-Larib and PICCOLO. These results suggest that medium and large variants may provide reliable localization without necessarily requiring the largest model in every family.

YOLO26n deserves separate attention despite ranking sixth by average mAP50–95. It provided the most favorable efficiency profile on both test sets and remained competitive in detection performance. In practical terms, YOLO26x appears more suitable when maximum cross-dataset localization performance is prioritized, whereas YOLO26n is more appropriate when real-time deployment and computational efficiency are the main constraints.

3.5. Training and validation dynamics of the selected You Only Look Once 26x model

After the cross-dataset ranking, YOLO26x was selected for additional training-dynamics analysis because it achieved the highest average mAP50–95 across the ETIS-Larib and PICCOLO test sets. The training and validation curves

of the selected model are presented in Figure 3. These curves provide complementary evidence on optimization behavior and help determine whether the final model performance was obtained through stable convergence rather than abrupt late-stage fluctuations.

Figure 3 shows the evolution of training loss, validation loss, precision, recall, mAP50, and mAP50–95 during model optimization. The solid green curves represent the recorded epoch-wise values, whereas the dotted burgundy curves indicate the smoothed trends. Training box loss, classification loss, and distribution focal loss decreased progressively, showing that the model learned both lesion localization and class-specific representations during fine-tuning. The mAP50 and mAP50–95 curves increased during the early and middle training stages and then stabilized in the later epochs. Validation losses showed moderate fluctuations after the initial convergence phase, which is expected in heterogeneous colonoscopy images because polyp size, illumination, mucosal texture, lesion boundary clarity, and background artifacts vary considerably across frames.

The training curves indicate that YOLO26x followed a

Table 4. Cross-dataset ranking based on mAP50–95

Rank	Model	ETIS-Larib mAP50–95	PICCOLO mAP50–95	Average mAP50–95	Interpretation
1	YOLO26x	0.6475	0.6449	0.6462	Best cross-dataset model
2	YOLO26m	0.6205	0.6526	0.6366	Strong balanced detector
3	YOLO12m	0.6245	0.6480	0.6363	Competitive mid-size model
4	YOLO26l	0.6356	0.6358	0.6357	Highly consistent across datasets
5	YOLO12x	0.6358	0.6274	0.6316	High-capacity but less efficient
6	YOLO26n	0.6364	0.6265	0.6315	Best lightweight alternative
7	YOLOv13s	0.6210	0.6238	0.6224	Moderate balanced performance
8	YOLOv13x	0.5805	0.6630	0.6218	Best on PICCOLO but weaker on ETIS-Larib
9	YOLO26s	0.6103	0.6276	0.6190	Efficient small model
10	YOLOv13n	0.5834	0.6507	0.6171	Sensitive on PICCOLO but less stable externally
11	YOLO12l	0.6143	0.6197	0.6170	Moderate cross-dataset performance
12	YOLO12n	0.6117	0.6102	0.6110	Lightweight baseline
13	YOLOv13l	0.5903	0.6258	0.6081	Limited cross-dataset balance
14	YOLO12s	0.5886	0.6166	0.6026	Lowest average mAP50–95

Abbreviations: mAP: Mean average precision; YOLO: You Only Look Once.

stable optimization pattern. The training box, classification, and distribution focal losses declined steadily, suggesting that the detector continued to improve localization and polyp-specific feature learning throughout training. In parallel, precision and mAP values increased rapidly during the early epochs and reached a stable range in the later phase. Recall showed a mild decline after its peak, which may reflect the model becoming more selective as optimization progressed. This behavior is consistent with the high precision observed on the PICCOLO test set and the strong cross-dataset mAP50–95 obtained by YOLO26x.

The validation curves should be interpreted together with the independent test results rather than in isolation. Although validation box, classification, and distribution focal losses fluctuated and increased mildly in the late training stage, the detection metrics remained stable, and

no collapse was observed in mAP50 or mAP50–95. Such divergence between validation loss and detection metrics can occur in polyp detection because small changes in bounding-box regression loss may not always translate into lower average precision, particularly when lesions are small, low contrast, or boundary ambiguous. The final selection of YOLO26x was therefore based on its cross-dataset test performance rather than validation loss alone. This choice is important because the model achieved the best average mAP50–95 across the independent ETIS-Larib and PICCOLO test sets.

3.6. Precision–recall behavior of the selected You Only Look Once 26x model

The cross-dataset ranking favored YOLO26x, and its precision–recall behavior was therefore examined on

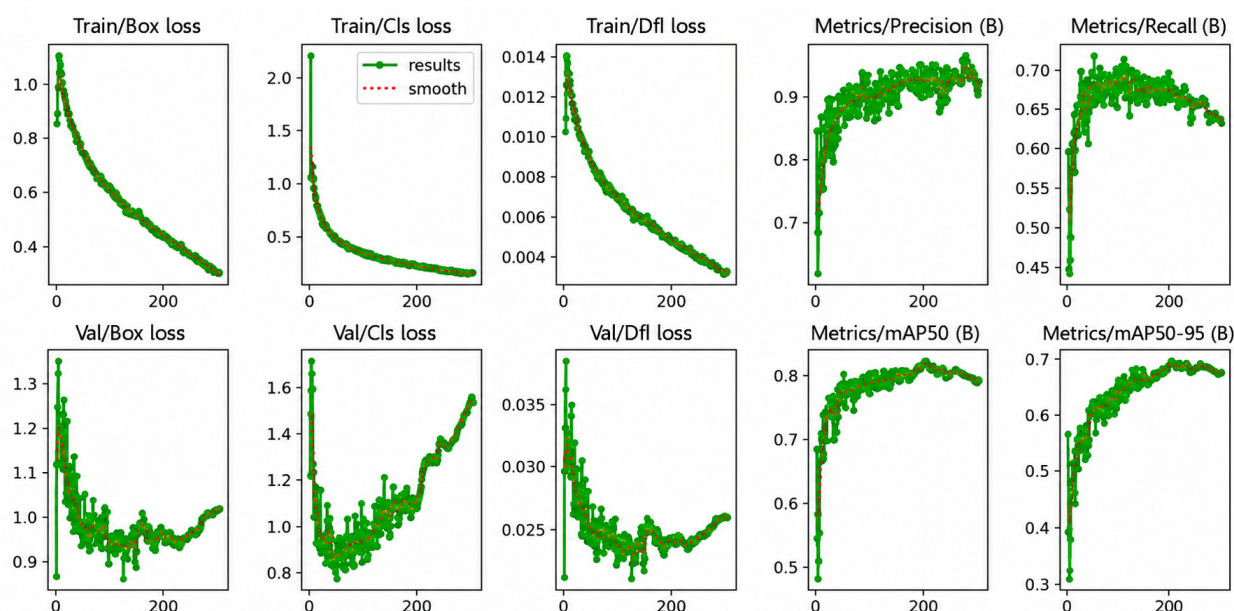


Figure 3. Training and validation dynamics of the selected YOLO26x model. The curves show the evolution of training losses, validation losses, precision, recall, mAP50, and mAP50-95 during fine-tuning. The model showed progressive reduction in training losses and stabilization of detection metrics in the later training stage, indicating stable optimization without abrupt metric collapse.

both independent test settings. This analysis helps clarify how the selected model balances false-positive control and lesion retrieval at different operating thresholds. The corresponding curves are shown in Figure 4.

On ETIS-Larib, YOLO26x achieved an mAP50 of approximately 0.847, consistent with the value reported in Table 2. The curve remained in the high-precision region over a broad recall interval, indicating that the model preserved reliable predictions across a range of confidence thresholds. The decline at higher recall values reflects the difficulty of detecting subtle or ambiguous lesions without accepting additional uncertain predictions.

On PICCOLO, the mAP50 of YOLO26x was approximately 0.789. The curve was consistent with the quantitative result in Table 3, where the model achieved very high precision but lower recall. This behavior indicates that YOLO26x was more selective on the PICCOLO test set. Such a profile may be useful when false alarms must be minimized, but threshold calibration would be required if higher screening sensitivity is preferred.

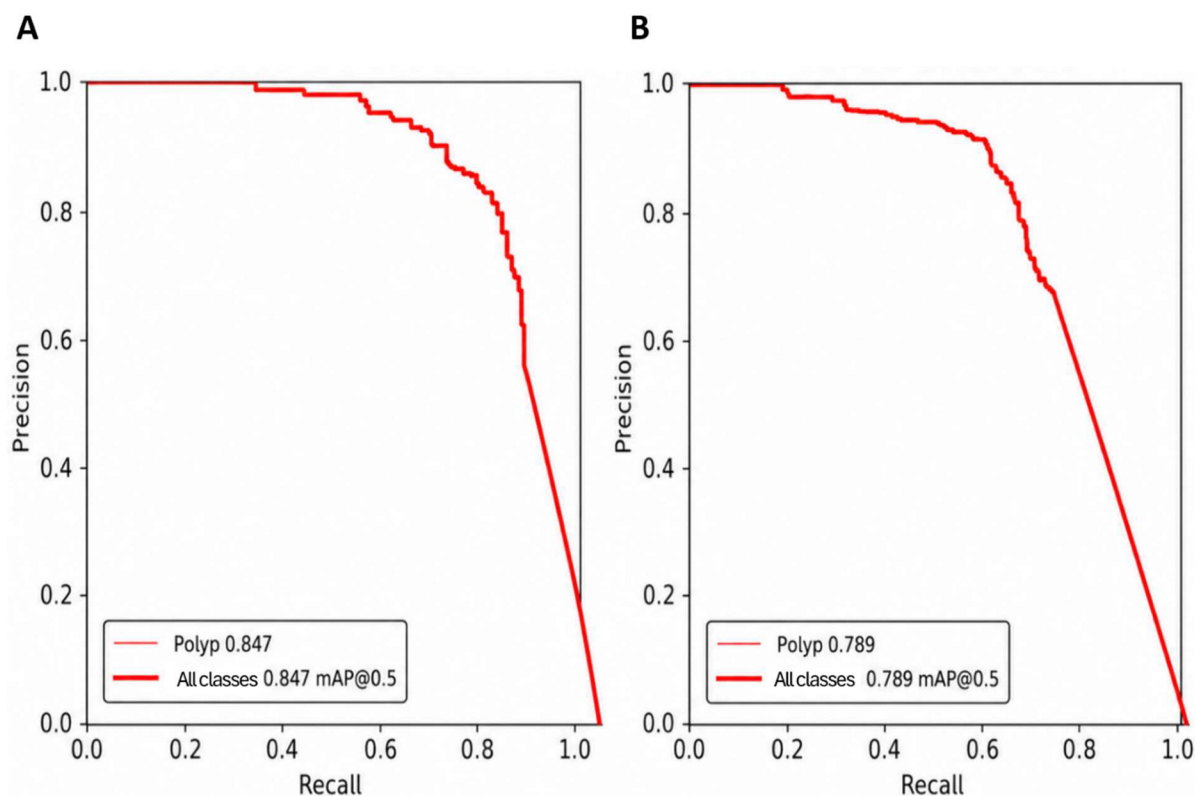
3.7. Qualitative detection results

Quantitative metrics are necessary for benchmarking, but visual assessment is also important for interpreting

localization quality in medical object detection. Because mAP50-95 evaluates detection performance across increasingly strict IoU thresholds, representative examples were examined with both ground-truth boxes and YOLO26x predictions. Figure 5 shows paired overlays from ETIS-Larib and PICCOLO, allowing spatial agreement and boundary mismatch to be assessed visually.

The qualitative examples indicate that YOLO26x localized visible polyp regions across heterogeneous appearances, including variable illumination, partial boundary ambiguity, mucosal background similarity, and differences in viewing angle and contrast. In well-defined lesions, the predicted boxes closely followed the annotated regions. In more difficult images, partial boundary mismatch was observed, particularly when lesion margins were indistinct or merged with folds, shadows, specular reflections, or low-contrast mucosa.

These examples help explain why mAP50 alone may be insufficient for colorectal polyp detection. A predicted box may overlap the lesion at an IoU threshold of 0.50 while still failing to capture the lesion boundary precisely at stricter thresholds. The ground-truth and prediction overlays therefore provide a visual counterpart to the mAP50-95 analysis and support a stricter interpretation of localization performance.



.Figure 4. Precision–recall curves of YOLO26x on the independent test sets. Precision–recall curves are shown for (A) ETIS-Larib and (B) PICCOLO. The curves illustrate how YOLO26x balances false-positive control and lesion retrieval across confidence thresholds, with different operating behavior across the external and held-out test distributions.

3.8. Failure case analysis

A separate failure analysis was included to avoid an overly favorable interpretation based only on successful detections. In medical imaging, aggregate metrics may obscure clinically relevant errors, especially when false negatives involve small or subtle lesions. Representative challenging cases are shown in [Figure 6](#).

False-positive detections were mainly associated with mucosal folds, specular reflections, bubbles, residual content, and local texture changes that resembled polyp-like structures. These patterns are expected in colonoscopy because background anatomy and imaging artifacts may share similar visual cues with true lesions. In such cases, increasing model capacity alone may not resolve the problem unless the model learns more discriminative lesion-background representations.

False negatives and weak detections were more common in small, flat, low-contrast, or partially visible

polyps. Boundary mismatch was also observed when lesion margins were indistinct or when the polyp extended into shadowed or saturated regions. These findings indicate that future improvement should focus on small-object sensitivity, boundary-aware localization, and robustness to illumination artifacts.

The failure analysis provides a realistic interpretation of the benchmark. YOLO26x achieved the best cross-dataset mAP50–95, but the model still showed clinically relevant limitations in difficult frames. For deployment-oriented colonoscopy systems, performance should therefore be evaluated not only by aggregate mAP values but also by subgroup behavior in small lesions, flat lesions, low-contrast frames, and artifact-heavy scenes.

3.9. Practical interpretation

The benchmark results indicate that model selection should depend on the intended use case. YOLO26x is the strongest candidate when the goal is maximum cross-

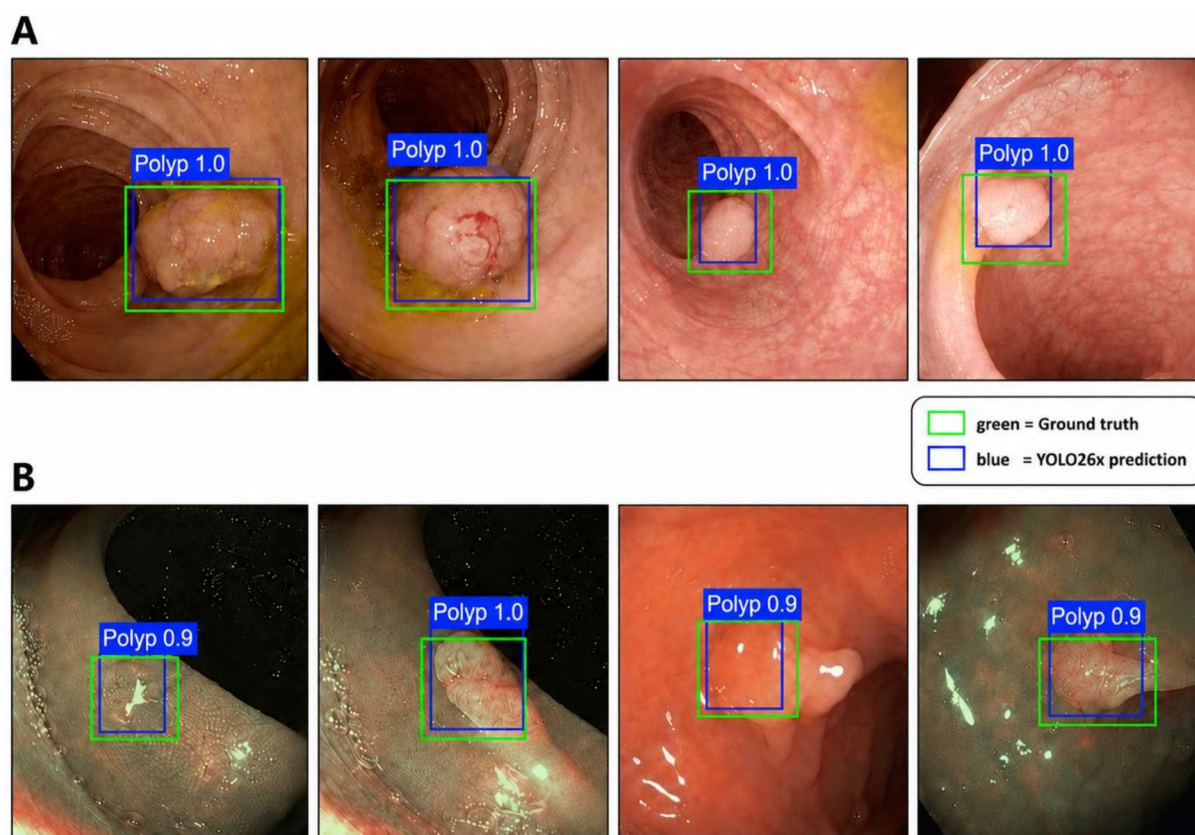


Figure 5. Representative ground-truth and YOLO26x prediction overlays on (A) ETIS-Larib and (B) PICCOLO. Green bounding boxes denote ground-truth polyp annotations, whereas blue bounding boxes denote YOLO26x predictions with confidence scores. The examples illustrate spatial agreement and boundary mismatch across heterogeneous endoscopic appearances, including variable illumination, low contrast, mucosal folds, specular reflections, and differences in lesion size and boundary visibility.

dataset localization performance. It achieved the highest average mAP50–95 across ETIS-Larib and PICCOLO and also produced the best mAP50–95 and recall on the ETIS-Larib external test set.

YOLO26n is the most practical lightweight alternative. It had the smallest parameter count, the lowest GFLOPs, and the fastest inference time on both test sets, while still maintaining competitive detection performance. This profile is important for real-time colonoscopy systems, where latency and hardware constraints may influence whether a model can be integrated into clinical workflow.

The comparison also shows that model family and model size should not be interpreted in isolation. YOLOv13x achieved the strongest PICCOLO mAP50–95, but it did not generalize as well to ETIS-Larib. YOLO26x provided the best cross-dataset balance, while YOLO26n delivered the best efficiency. This separation between

accuracy-oriented and deployment-oriented choices is one of the main findings of the benchmark.

4. Discussion

4.1. Cross-dataset behavior of You Only Look Once model families

The present study compared recent YOLO12, YOLOv13, and YOLO26 model families for colorectal polyp detection using three public datasets with distinct experimental roles. Rather than evaluating a single detector or a single model scale, the analysis considered both architectural family differences and within-family capacity changes from lightweight to high-capacity variants. This design is important for polyp detection because model performance may depend not only on the detector generation but also on the balance between feature capacity, localization ability, and computational cost.

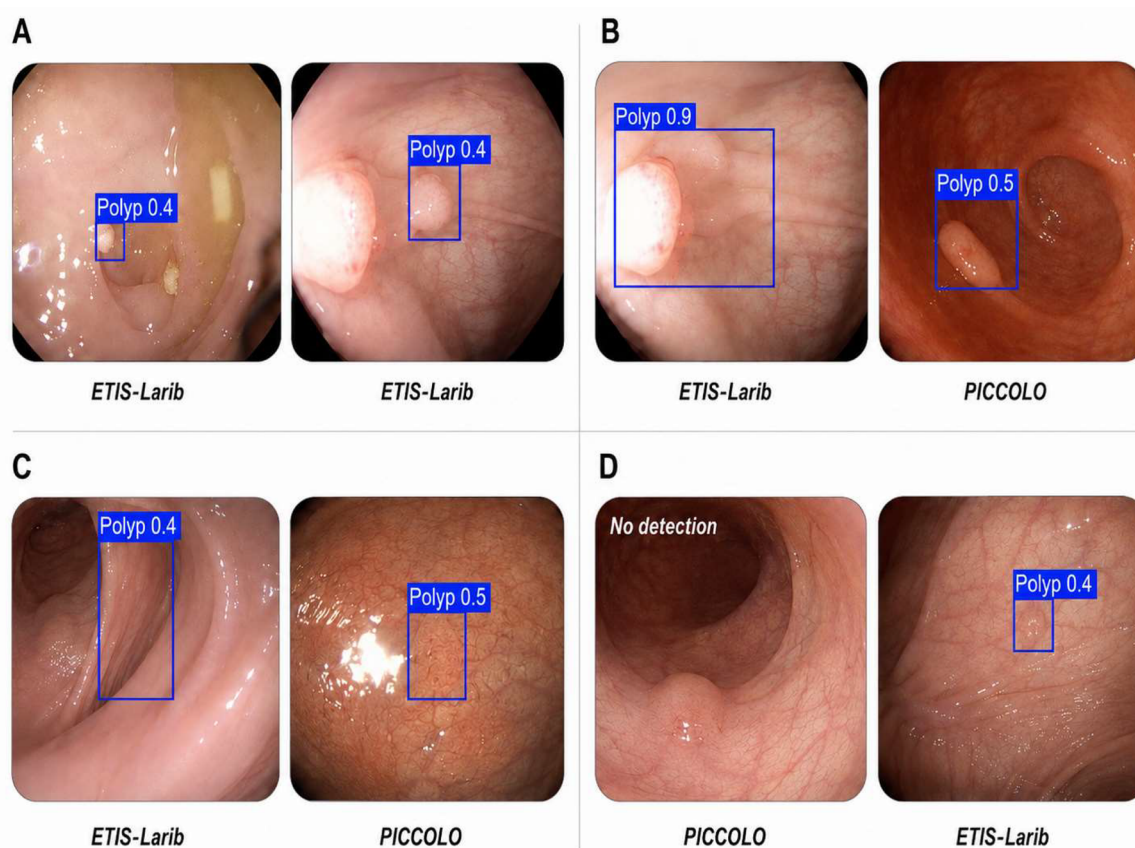


Figure 6. Representative challenging cases and failure patterns. Representative examples illustrate false-positive detections, weak localization, and missed-detection patterns observed during qualitative assessment. Common sources of error included mucosal folds, specular reflections, bubbles, residual content, low-contrast lesions, partial visibility, and ambiguous lesion boundaries. (A) Low-confidence detections; (B) boundary mismatch cases; (C) false-positive detections; and (D) missed or weakly localized lesions.

The results showed that the best-performing model was not identical across the two test sets. On the ETIS-Larib external test set, YOLO26x achieved the highest mAP50–95 and recall, indicating stronger localization and lesion retrieval under a more challenging external setting. On the PICCOLO held-out test set, YOLOv13x achieved the highest mAP50–95 and mAP50, whereas YOLO26x produced the highest precision. This difference suggests that model ranking is sensitive to dataset characteristics, including illumination, lesion morphology, background texture, image sharpness, and annotation properties.

The cross-dataset ranking provided a more balanced interpretation. YOLO26x obtained the highest average mAP50–95 across ETIS-Larib and PICCOLO, making it the most stable accuracy-oriented model in this benchmark. However, the performance gap between

models was not uniform across datasets. For example, YOLOv13x performed strongly on PICCOLO but showed weaker generalization on ETIS-Larib. This finding supports the use of more than one independent test setting in polyp detection studies. A detector that performs well on a familiar test distribution may not show the same behavior when image appearance and lesion presentation change.

The results also indicate that increasing model size alone is not a sufficient strategy for robust polyp detection. Some large models produced strong results on one dataset but did not maintain the same advantage externally. In contrast, several mid-size and lightweight models remained competitive, especially when efficiency was considered. This is relevant for colonoscopy applications, where deployment conditions may favor models that are fast and stable rather than models that only provide marginal gains in aggregate accuracy.

4.2. Accuracy, efficiency, and clinical interpretation

The primary ranking metric in this study was mAP50–95 because it evaluates localization quality across multiple IoU thresholds. This is stricter than mAP50 and more appropriate for medical object detection, where a predicted box should not merely overlap a lesion but should also localize it with acceptable spatial precision. The difference between mAP50 and mAP50–95 in the present results confirms that relying only on mAP50 may overestimate the practical quality of a detector.

From an accuracy perspective, YOLO26x was the strongest overall model. Its highest average mAP50–95 across the two test sets indicates that it provided the most balanced localization performance. Its high recall on ETIS-Larib is also clinically meaningful because false-negative detections correspond to missed polyps. In screening-oriented colonoscopy, sensitivity is particularly important, since diminutive, flat, pale, or low-contrast lesions may be overlooked during visual inspection.

However, the efficiency results point to a different model choice. YOLO26n had the lowest parameter count, the lowest GFLOPs, and the fastest inference time on both test sets, while maintaining competitive detection performance. This makes YOLO26n a strong lightweight alternative for real-time or hardware-constrained environments. Therefore, the model choice depends on the intended use. YOLO26x is more suitable when maximum cross-dataset detection performance is required, whereas YOLO26n is more appropriate when low latency, small model size, and computational efficiency are prioritized.

The precision–recall curves and qualitative examples provided additional insight beyond the tabular results. YOLO26x maintained strong precision over a broad recall range on ETIS-Larib, but the curve declined when recall was pushed to higher values. This reflects the expected trade-off in polyp detection: recovering more subtle lesions often increases the number of uncertain predictions. On PICCOLO, YOLO26x showed a more conservative pattern, with very high precision but lower recall. Such behavior may be acceptable in some decision-support settings, but threshold calibration would be needed if the system is intended primarily for screening assistance.

The failure cases were consistent with known difficulties in colonoscopy-based detection. False-positive predictions were often associated with mucosal folds, specular reflections, bubbles, residual content, or local texture changes that resembled polyp-like structures. Missed or weakly localized lesions were more often small, flat, low-contrast, partially visible, or located near visually complex regions. These errors suggest that future improvement

should focus on small-lesion sensitivity, boundary-aware localization, and robustness to illumination artifacts rather than simply increasing network capacity.

4.3. Implications for polyp detection benchmarking

The findings have several implications for future benchmark studies. First, polyp detection models should be assessed using both held-out and external test sets whenever possible. Internal validation alone may not reveal how a model behaves under dataset shift. The difference between ETIS-Larib and PICCOLO rankings in this study illustrates why a single test set can lead to an incomplete conclusion.

Second, benchmark studies should report efficiency metrics together with detection metrics. Parameter count, GFLOPs, and inference time are not secondary technical details in colonoscopy applications. They directly affect whether a model can be deployed in real-time clinical environments. A model with slightly lower mAP50–95 but substantially faster inference may be more useful in practice than a larger model with only marginal accuracy gains.

Third, qualitative analysis remains necessary. Aggregate metrics summarize detection behavior but do not reveal whether errors occur in clinically important cases. For this reason, successful detections and failure cases should both be presented. In the present study, the failure analysis showed that the remaining errors were mainly related to visually difficult lesion presentations and background artifacts. This type of evidence is important for interpreting the clinical relevance of model performance.

4.4. Limitations and future work

This study has several limitations that should be considered when interpreting the benchmark results. First, the experiments were conducted using public still-image datasets. Although PICCOLO and ETIS-Larib provide useful visual diversity, they cannot fully reproduce the variability of routine colonoscopy. Real clinical deployment may involve different endoscopy systems, imaging modes, bowel preparation quality, patient populations, lesion prevalence, frame sampling patterns, withdrawal technique, and operator-dependent acquisition conditions. Multi-center prospective validation is therefore required before the benchmark results can be translated into clinical decision-support use.

Second, the task was formulated as single-class polyp detection. This design is appropriate for computer-aided detection systems whose primary function is real-time localization of suspicious polyp regions, but it does not address lesion characterization. The included datasets

did not provide a consistent lesion-level taxonomy for macroscopic morphology or histological diagnosis across all images. As a result, the present analysis cannot determine whether detector performance differs among pedunculated, sessile, flat, depressed, serrated, adenomatous, hyperplastic, dysplastic, or malignant lesions. Future datasets should include verified lesion-level endoscopic and histopathological metadata so that localization performance can be examined together with morphology, histology, size, anatomical location, and malignancy risk.

Third, lesion-level subgroup analysis was not performed because harmonized subtype labels were not available across the included datasets. This is clinically relevant because errors may occur disproportionately in diminutive, flat, low-contrast, serrated, partially visible, or artifact-obscured lesions. Without reliable lesion-level metadata, these categories could not be evaluated statistically. Future benchmarks should report performance separately by lesion size, macroscopic morphology, histological category, anatomical location, image quality, and artifact burden. Such analyses would make detection results more interpretable for clinical use, especially when comparing models with similar aggregate mAP values.

Fourth, negative-only images without polyp annotations were not included in the evaluated subsets. The benchmark therefore primarily assesses localization performance on polyp-positive images and does not fully characterize false-positive behavior on completely polyp-free colonoscopy frames. This limitation is important because routine colonoscopy contains long periods without visible polyps, during which false alarms may affect workflow, attention, and user trust. Future work should include negative-only frames and longer video segments to evaluate false alarm rate, false alarm persistence, and threshold behavior under screening-like conditions.

Fifth, the evaluation was based on still images rather than continuous colonoscopy videos. Still-image analysis is useful for controlled benchmarking, but real-time colonoscopy involves temporal continuity, camera motion, repeated lesion views, transient blur, changing insufflation, and frame-level redundancy. Video-based evaluation would allow clinically relevant endpoints such as temporal stability, detection delay, missed-lesion duration, frame-to-frame flickering, and persistent false alarms to be assessed more directly. Future models may benefit from temporal smoothing, tracking, or video-level aggregation, but these additions should be validated without introducing delayed or unstable alerts.

Sixth, the comparisons were interpreted as benchmark-

based numerical comparisons rather than formal claims of statistical superiority. Formal confidence intervals and paired p -values were not reported for mAP50 or mAP50–95 because valid statistical inference for these threshold-integrated detection metrics requires image-level prediction outputs and repeated metric recomputation under paired resampling. Aggregate table values alone are not sufficient for a statistically valid paired test. For this reason, small differences between closely ranked models should be interpreted cautiously, particularly when the performance gaps are narrow. Future benchmark studies should retain image-level prediction outputs and report paired bootstrap confidence intervals for the primary localization metric.

Seventh, a unified training configuration was used to ensure a fair comparison across model families and scales. This protocol supports direct benchmarking, but it may not represent the optimal setting for every architecture. Some detectors may benefit from model-specific augmentation policies, learning-rate schedules, confidence thresholds, loss-weight tuning, or longer optimization. Future work should compare fixed-protocol benchmarking with separately optimized model-specific training strategies, while keeping independent test sets unchanged.

Finally, confidence calibration, uncertainty estimation, and reader interaction were not evaluated. In clinical use, confidence scores should be reliable because overconfident false positives and false negatives may affect trust and workflow integration. The practical value of a detector also depends on how endoscopists respond to alerts, whether false positives interrupt inspection, and whether threshold selection can be adjusted for screening-oriented sensitivity. Future studies should evaluate calibration, uncertainty-aware detection, threshold selection, reader studies, and prospective deployment under real colonoscopy conditions.

5. Conclusion

This study presented a unified benchmark of YOLO12, YOLOv13, and YOLO26 model families for colorectal polyp detection using PolypGen2021, PICCOLO, and ETIS-Larib datasets. The experimental design included patient-level splitting for PolypGen2021, preservation of the predefined patient-level PICCOLO split, and external testing on ETIS-Larib. By evaluating both family-level differences and model-scale variations from nano to x-large, the study provided a practical comparison of detection accuracy, localization quality, computational cost, and inference speed. The results showed that model performance varied across test settings. On the ETIS-Larib external test set, YOLO26x achieved the best mAP50–95 of 0.6475 and the

highest recall of 0.7736, indicating the strongest external localization and lesion retrieval performance. On the PICCOLO held-out test set, YOLOv13x obtained the highest mAP50–95 of 0.6630 and the highest mAP50 of 0.8286, while YOLO26x reached the highest precision of 0.9118. When the two test sets were considered together, YOLO26x achieved the highest average mAP50–95 of 0.6462, making it the most balanced cross-dataset detector in this benchmark. The comparison between lightweight and large models revealed an important deployment trade-off. YOLO26x provided the strongest overall detection performance, but YOLO26n achieved the most favorable efficiency profile on both test sets, with only 2.3750 M parameters, 5.1894 GFLOPs, and inference times of 2.8929 ms/img on ETIS-Larib and 2.6825 ms/img on PICCOLO.

Acknowledgments

The author gratefully acknowledges the developers and maintainers of the publicly available PolypGen2021, PICCOLO, and ETIS-Larib datasets, which made this research possible. The author also acknowledges the computational resources used to train and evaluate the deep learning models.

Funding

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1507 SME R&D Start-up Support Program (project number 7250224, call period 1507-2025-1). The funder had no role in the study design, data collection, analysis, interpretation of results, manuscript preparation, or the decision to submit the article for publication.

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this manuscript.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

This study used publicly available, anonymized colorectal polyp image datasets and did not involve the recruitment of new human participants, collection of new clinical data, or direct interaction with patients. Therefore, additional institutional ethics approval and informed consent were not required for this secondary analysis of public datasets.

Consent for publication

Not applicable. This study used publicly available anonymized image datasets and did not include identifiable

personal information.

Availability of data

The datasets analyzed in this study are publicly available from their respective official repositories or original dataset providers. PolypGen2021, PICCOLO, and ETIS-Larib were used in accordance with their public availability and dataset usage conditions. The processed training, validation, and test split information can be made available by the author upon reasonable request, where permitted by the dataset licenses.

The code used for model training and evaluation was implemented using the Ultralytics YOLO framework. The experiment configuration, model settings, and evaluation protocol are described in the Materials and Methods section to support reproducibility. Additional implementation details can be made available by the author upon reasonable request.

Further disclosure

AI-assisted tools were used only for language editing and readability improvement. The study design, data analysis, interpretation, and conclusions were developed and verified by the author, who takes full responsibility for the content of the manuscript.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin.* 2024;74(1):12-49.
doi: 10.3322/caac.21820
2. Morgan E, Arnold M, Gini A, *et al.* Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut.* 2023;72(2):338-344.
doi: 10.1136/gutjnl-2022-327736
3. Winawer SJ, Zauber AG, Ho MN, *et al.* Prevention of colorectal cancer by colonoscopic polypectomy. *N Engl J Med.* 1993;329(27):1977-1981.
doi: 10.1056/NEJM199312303292701
4. Zauber AG, Winawer SJ, O'Brien MJ, *et al.* Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N Engl J Med.* 2012;366(8):687-696.
doi: 10.1056/NEJMoa1100370
5. Corley DA, Jensen CD, Marks AR, *et al.* Adenoma detection rate and risk of colorectal cancer and death. *N Engl J Med.* 2014;370(14):1298-1306.
doi: 10.1056/NEJMoa1309086
6. Rex DK, Schoenfeld PS, Cohen J, *et al.* Quality indicators for colonoscopy. *Gastrointest Endosc.* 2015;81(1):31-53.
doi: 10.1016/j.gie.2014.07.058

7. van Rijn JC, Reitsma JB, Stoker J, Bossuyt PM, van Deventer SJ, Dekker E. Polyp miss rate determined by tandem colonoscopy: a systematic review. *Am J Gastroenterol*. 2006;101(2):343-350.
doi: 10.1111/j.1572-0241.2006.00390.x
8. Soetikno RM, Kaltenbach T, Rouse RV, *et al*. Prevalence of nonpolypoid colorectal neoplasms in asymptomatic and symptomatic adults. *JAMA*. 2008;299(9):1027-1035.
doi: 10.1001/jama.299.9.1027
9. Leggett B, Whitehall V. Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology*. 2010;138(6):2088-2100.
doi: 10.1053/j.gastro.2009.12.066
10. Wang P, Liu P, Glissen Brown JR, *et al*. Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. *Gastroenterology*. 2020;159(4):1252-1261.e5.
doi: 10.1053/j.gastro.2020.06.023
11. Schauer C, van Rijnsoever M, Jafer A, Walmsley R, Wang MTM, Atkinson N. Man plus machine: a randomized control trial of artificial intelligence including the impact of adjunctive polyp detection techniques. *J Gastroenterol Hepatol*. 2026;41(1):117-127.
doi: 10.1111/jgh.70145
12. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg*. 2014;9(2):283-293.
doi: 10.1007/s11548-013-0926-3
13. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans Med Imaging*. 2016;35(2):630-644.
doi: 10.1109/TMI.2015.2487997
14. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. *Comput Biol Med*. 2020;126:104003.
doi: 10.1016/j.compbiomed.2020.104003
15. Amirmohammadi E, Shalbf A, Esteki A, *et al*. Application of artificial intelligence in colonoscopy imaging for polyp analysis: a systematic review. *Comput Methods Programs Biomed*. 2026;277:109239.
doi: 10.1016/j.cmpb.2026.109239
16. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput Med Imaging Graph*. 2015;43:99-111.
doi: 10.1016/j.compmedimag.2015.02.007
17. Jha D, Smedsrud PH, Riegler MA, *et al*. Kvasir-SEG: a segmented polyp dataset. In: *MultiMedia Modeling*. Heidelberg, Germany: Springer; 2020:451-462.
doi: 10.1007/978-3-030-37734-2_37
18. Sánchez-Peralta LF, Pagador JB, Picón A, *et al*. PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. *Appl Sci*. 2020;10(23):8501.
doi: 10.3390/app10238501
19. Ali S, Jha D, Ghatwary N, *et al*. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Sci Data*. 2023;10(1):75.
doi: 10.1038/s41597-023-01981-y
20. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NY, USA: IEEE; 2016:779-788.
doi: 10.1109/CVPR.2016.91
21. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection. *arXiv*. Preprint posted online 2020.
doi: 10.48550/arXiv.2004.10934
22. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York, NY, USA: IEEE; 2023:7464-7475.
doi: 10.1109/CVPR52729.2023.00721
23. Urban G, Tripathi P, Alkayali T, *et al*. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*. 2018;155(4):1069-1078.e8.
doi: 10.1053/j.gastro.2018.06.037
24. Lee JY, Jeong J, Song EM, *et al*. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. *Sci Rep*. 2020;10(1):8379.
doi: 10.1038/s41598-020-65387-1
25. Pacal I, Karaboga D. A robust real-time deep learning based automatic polyp detection system. *Comput Biol Med*. 2021;134:104519.
doi: 10.1016/j.compbiomed.2021.104519
26. Pacal I, Karaman A, Karaboga D, *et al*. An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets. *Comput Biol Med*. 2022;141:105031.
doi: 10.1016/j.compbiomed.2021.105031
27. Karaman A, Karaboga D, Pacal I, *et al*. Hyper-parameter

- optimization of deep learning architectures using artificial bee colony algorithm for high performance real-time automatic colorectal cancer polyp detection. *Appl Intell.* 2023;53(12):15660-15685.
doi: 10.1007/s10489-022-04299-1
28. Karaman A, Pacal I, Basturk A, *et al.* Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony. *Expert Syst Appl.* 2023;221:119741.
doi: 10.1016/j.eswa.2023.119741
29. Zare O, Beigzadeh M, Bzuayene AA, *et al.* LightAttn-YOLO-V8: an efficient colorectal polyp detection with lightweight attention mechanism based on YOLO. *Sci Rep.* 2026;16(1):18100.
doi: 10.1038/s41598-026-35390-z
30. Sureban NB, Ajish AC, Abhinav Y, Prusty MR, Patra SR. A channel-efficient colorectal polyp detection using wavelet-enhanced reverse attention and custom neck YOLO architecture. *Biomed Signal Process Control.* 2026;119:109909.
doi: 10.1016/j.bspc.2026.109909
31. Du J, He Z, Zhang S, Wu G, Liang C. Gastrointestinal polyp detection method based on the improved RT-DETR. *Sci Rep.* 2026;16(1):7020.
doi: 10.1038/s41598-026-38617-1
32. Lin Y, Huang C, Tian H, Yang B, Deng T, Pan Y, Wang H, Li X. Improving generalization of polyp detection via conditional StyleGAN augmented training. *npj Digit Med.* 2026;9(1):113.
doi: 10.1038/s41746-025-02293-8
33. Liao H, Hu X, Zhang X, Tang T, Xu X. An efficient object detection network for early polyp detection. *Biomed Signal Process Control.* 2026;113:108860.
doi: 10.1016/j.bspc.2025.108860
34. Elkarazle K, Raman V, Chua C, Then P, Prabhavathy M. EfficientPolySeg: a fast and accurate network for complex polyp segmentation. *Biomed Signal Process Control.* 2026;112:108449.
doi: 10.1016/j.bspc.2025.108449
35. Tesema FB, Guerra-Manzanara A, Cui T, Zhang Q, Solomon MM, He X. LGPS: a lightweight GAN-based approach for polyp segmentation in colonoscopy images. *Biomed Signal Process Control.* 2026;118:109777.
doi: 10.1016/j.bspc.2026.109777
36. Li F, Xu L, Ma Z, Zhao Y, Li X. APU-Net: a U-Net enhanced network with dynamic feature fusion and pyramid cross-attention mechanism for polyp segmentation. *Digit Signal Process.* 2026;172:105879.
doi: 10.1016/j.dsp.2025.105879
37. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. *arXiv.* Preprint posted online 2025.
doi: 10.48550/arXiv.2502.12524
38. Lei M, Li S, Wu Y, *et al.* YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv.* Preprint posted online 2025.
doi: 10.48550/arXiv.2506.17733
39. Jocher G, Qiu J, Liu M, Lyu S, Akyon FC, Kalfaoglu ME. Ultralytics YOLO26: unified real-time end-to-end vision models. *arXiv.* Preprint posted online 2026.
doi: 10.48550/arXiv.2606.03748
40. Lin TY, Maire M, Belongie S, *et al.* Microsoft COCO: Common Objects in Context. In: *Lecture Notes in Computer Science.* Cham, Switzerland: Springer International Publishing; 2014:740-755.
doi: 10.1007/978-3-319-10602-1_48