



UDC 004.89; 004.93

PACS 07.05.Mh 07.05.Tp 42.62.Be

DOI: 10.22363/2658-4670-2026-34-2-260-273

EDN: JMNMPUP

Efficient Polyp Segmentation and Classification with YOLO11–SAM3 and LoRA

Eugene Yu. Shchetinin¹, Leonid A. Sevastianov², Anastasiia A. Tiutiunnik²

¹ Sevastopol State University, 33 Universitetskaya St, Sevastopol, 299053, Russian Federation

² RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

(received: February 22, 2026; revised: March 12, 2026; accepted: March 20, 2026)

Abstract. *Purpose* To develop and experimentally evaluate an autonomous computer-aided diagnosis (CADx) system for colonoscopy that integrates polyp segmentation with binary histological classification (adenoma/non-adenoma) while maintaining computational efficiency for near real-time clinical deployment. *Materials and Methods* We propose a cascaded YOLO11–SAM3–LoRA architecture wherein a YOLO11–medium detector generates region proposals (bounding boxes), followed by SAM 3 performing box-prompted segmentation with parameter-efficient domain adaptation via Low-Rank Adaptation (LoRA) and a semantic anchoring mechanism based on text prompt ensembling. Histological classification is implemented through a dual-stream classifier that fuses region-of-interest (ROI) features extracted by EfficientNet–B0 with morphological mask descriptors. Training was conducted on the combined Kvasir-SEG and CVC–ClinicDB datasets (detection/segmentation) and CVC–HDClassif (classification). Generalization was evaluated on the external ETIS–Larib dataset under an external zero-shot protocol. *Results* On the primary test set, the system achieved a mean Dice coefficient (mDice) of 0.962 for segmentation, adenoma sensitivity of 92.3%, and area under the receiver operating characteristic curve (AUROC) of 0.942 for binary classification. Under external zero-shot evaluation on ETIS–Larib, the system maintained segmentation quality (mDice = 0.884), indicating robustness to domain shift. End-to-end processing throughput reached 38.2 FPS on an NVIDIA A100 GPU. *Conclusions* The results demonstrate that the combination of cascaded localization, foundation model segmentation with LoRA adaptation, and text prompt ensembling achieves high segmentation quality and competitive CADx accuracy while preserving computational efficiency. Clinical deployment requires further prospective multicenter validation and error analysis stratified by lesion subtypes.

Key words and phrases: polyp segmentation, polyp classification, computer-aided diagnosis, CADx, colonoscopy, YOLO11, SAM 3, LoRA, prompt ensembling, deep learning

For citation: E. Y. Shchetinin, L. A. Sevastianov, A. A. Tiutiunnik “Efficient Polyp Segmentation and Classification with YOLO11–SAM3 and LoRA,” *Discrete and Continuous Models and Applied Computational Science*, vol. 34, no. 2, pp. 260–273, 2026. DOI: 10.22363/2658-4670-2026-34-2-260-273. EDN: JMNMPUP.

© 2026 E. Y. Shchetinin, L. A. Sevastianov, A. A. Tiutiunnik



This work is licensed under a Creative Commons “Attribution-NonCommercial 4.0 International” license.

1. Introduction

The substantial mortality burden associated with colorectal cancer (CRC) necessitates improvements in early detection quality. Although colonoscopy remains the primary screening modality, its effectiveness is constrained by inter-operator variability: systematic reviews indicate that the adenoma miss rate (AMR) ranges from approximately 14 to 30% [1]. This observation motivates the development of computer-aided systems capable of enhancing lesion detection reliability (CADE) and supporting optical stratification (CADx) under real-time constraints. When appropriately validated according to clinical criteria, CADx approaches may potentially support resect-and-discard strategies, thereby reducing unnecessary histopathological examinations and associated costs [2].

The evolution of automated endoscopic analysis has progressed from specialized convolutional architectures to the adaptation of foundation models. Early segmentation solutions based on U-Net [3] and its variants (UNet++ [4], ResUNet [5]) achieved high accuracy on images with well-defined boundaries; however, their limited effective receptive field and sensitivity to local textures could lead to errors on typical endoscopic artifacts (mucosal folds, specular reflections, air bubbles). Subsequent developments incorporated attention mechanisms and multi-scale contextual aggregation (e.g., PraNet [6]) alongside transformer-based architectures (e.g., Polyp-PVT [7]), which better capture global dependencies but frequently require trade-offs between computational complexity, spatial resolution, and video-rate inference speed.

At the current stage, two complementary research directions can be identified, each addressing only part of the clinical requirements. First, object-centric detectors of the YOLO family [8, 9] provide high throughput suitable for video streams; however, their bounding box outputs are insufficient for precise boundary delineation and morphological assessment—both critical for measurement and subsequent texture-based or morphological classification. Second, promptable foundation segmentation models (SAM/MedSAM/SAM2/SAM3) can produce detailed masks and demonstrate transferability across diverse visual conditions, yet their application typically requires external prompts (points, boxes, or text queries) and may exhibit sensitivity to domain shift in the specialized endoscopic environment [10–13]. This complicates the construction of fully autonomous pipelines without additional localization mechanisms and domain adaptation, particularly for low-contrast and morphologically ambiguous lesions including flat and sessile serrated polyps.

Existing attempts to combine detectors with foundation segmentation models (e.g., YOLO-SAM hybrids) are often limited to naive cascading, which in certain implementations leads to: (i) redundant computation when multiple localization mechanisms operate simultaneously, and (ii) underutilization of semantic prompts and textual concepts. Accordingly, the scientific and applied objective of this study is to develop an autonomous and computationally efficient mechanism for localization and semantic adaptation that transfers the advantages of foundation segmentation to the specialized colonoscopy domain while minimizing manual prompting requirements and preserving suitability for near real-time deployment.

1.1. Scientific Novelty and Contributions

The principal contributions of this work are as follows:

1. **Autonomous cascade architecture for colonoscopy detection and foundation-based segmentation.** We propose a computationally efficient YOLO11-SAM3-LoRA architecture that combines rapid candidate localization with subsequent semantic boundary refinement. This design jointly satisfies video-stream requirements (throughput) and clinically meaningful segmentation (precise masks) within a unified automatic pipeline.

2. **Semantic stabilization via text prompt ensembling.** We develop a semantic anchoring mechanism based on an ensemble of clinical terminology and aggregated textual representations, which enhances prompting robustness to formulation variability and domain shift. Ablation analysis demonstrates a measurable contribution of this component to mDice compared with single-prompt baselines.
3. **Integration of segmentation with CADx classification and external generalization assessment.** We implement a dual-stream classification scheme that fuses ROI features with morphological mask descriptors and conduct external zero-shot evaluation on ETIS-Larib. The obtained results (mDice = 0.884 on ETIS-Larib; AUROC = 0.942 for binary histological classification) indicate practical transferability of the proposed configuration to conditions differing from the training distribution.

2. Materials and Methods

2.1. Datasets

The following datasets were employed for training, evaluation, and generalization assessment.

2.1.1. Training and Validation (Detection and Segmentation)

Kvasir-SEG [14] and CVC-ClinicDB [15] were used for training detection and segmentation tasks. The combined volume comprises 1,612 images with binary polyp masks.

2.1.2. Training (Classification)

CVC-HDClassif [16] was employed for training the binary classifier, as it contains histologically verified labels (adenoma/non-adenoma). The class ratio (adenomas/non-adenomas) was 1.32:1; class imbalance was addressed through weighted loss functions.

2.1.3. In-Domain Testing

For in-domain evaluation, a held-out test partition comprising 20% of the training datasets was utilized, allocated at the image level with a fixed seed (seed = 5) and stratified by dataset to maintain proportional representation. Images from the external ETIS-Larib dataset were excluded from the test set.

2.1.4. External Zero-Shot Evaluation

ETIS-Larib Polyp DB [17] contains 196 images and exhibits pronounced domain shift relative to the training data (lighting conditions, contrast, equipment, and scene characteristics). External evaluation followed a zero-shot protocol: ETIS data were not used for training or hyperparameter tuning. All images were resized to an input resolution of 512×512 pixels and normalized identically to the training pipeline. Augmentations were applied exclusively to the training partition and were not employed during metric computation on ETIS.

2.2. Model Architecture

The system implements a three-stage processing pipeline as illustrated in Figure 1.

Table 1

Configuration of trainable parameters in the YOLO11-SAM3-LoRA model

Component	Status	Method
YOLO11 Backbone	Frozen	COCO pretraining
SAM 3 Image Encoder	Frozen	SA-Co pretraining
SAM 3 Attention Projections	Trainable	LoRA ($r = 16$)
SAM 3 Text Prompt Encoder	Frozen	Built-in SAM 3 encoder
Classification Module (EfficientNet-B0 + MLP)	Trainable	Full training

2.2.1. Coarse Detection (Region Proposal Generation)

YOLO11-medium [18] serves as the region proposal network. It performs initial background filtering and generates bounding boxes $B_i = [x, y, w, h]$ with associated confidence scores C_d . Unlike standard SAM deployment, the built-in DETR detector in SAM 3 is not utilized to avoid computational redundancy and prediction conflicts.

2.2.2. Semantic Refinement with Prompt Ensembling

For each candidate B_i , the SAM 3 model [19] is applied in box-prompted mode. To enhance semantic selectivity, instead of a single prompt, we employ an ensemble of clinical concepts T_{ens} :

$$T_{\text{ens}} = \{\text{“colorectal polyp”}, \text{“sessile lesion”}, \text{“adenoma”}, \text{“mucosal anomaly”}\}$$

For semantic prompting, we utilize the built-in text encoder of SAM 3. We construct an ensemble of K textual class descriptions (prompt ensemble), obtain corresponding embeddings, normalize them, and aggregate via averaging (with subsequent normalization), which reduces sensitivity to single-prompt formulation. Notably, this approach enables confident detection of low-contrast flat neoplasms that are systematically missed by the model when using a unitary “polyp” prompt.

2.2.3. Low-Rank Adaptation (LoRA)

For domain adaptation to medical imaging, trainable low-rank matrices ($r = 16$, $\alpha = 32$) are injected into the attention layers of the SAM 3 encoder. Weights are updated according to:

$$W' = W_{\text{frozen}} + \Delta W = W_{\text{frozen}} + BA$$

where $A \in \mathbb{R}^{r \times k}$, $B \in \mathbb{R}^{d \times r}$, and W_{frozen} denotes the frozen pretrained weights [20]. At rank $r = 16$, only 2.03% of parameters are trainable, providing a balance between plasticity (adaptation to intestinal textures) and stability (preservation of boundary detection capabilities). The parameter distribution is summarized in Table 1.

2.2.4. Dual-Path Classifier

Traditional approaches utilize mask features for classification; however, histological type is determined by surface texture (pit pattern) and vascularization, which may be lost in binary masks. The YOLO11-SAM3-LoRA classifier integrates two streams:

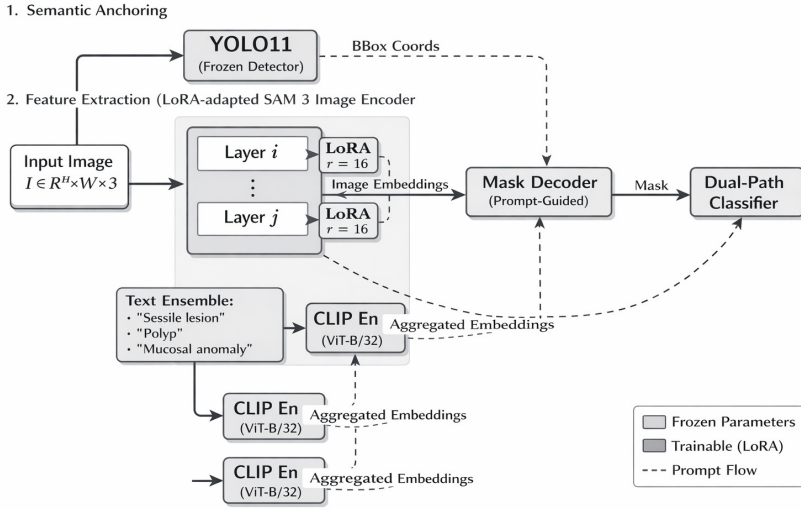


Figure 1. Block diagram of the proposed CADx system architecture: candidate generation using YOLO11, semantic refinement based on SAM 3, and dual-path classification

Visual Path (V_{vis}). Feature extraction within the image ROI is performed using the lightweight convolutional network EfficientNet-B0, enabling texture and vascular pattern analysis.

Morphological Path (V_{morph}). Geometric mask descriptors are analyzed via a multilayer perceptron (MLP), including area, perimeter, shape factor, and boundary irregularity.

The final class probability P_{cls} is computed by applying the softmax function to the output of a fully connected layer over the concatenated feature vectors V_{vis} and V_{morph} :

$$P_{\text{cls}} = \text{Softmax}(\text{MLP}(V_{\text{vis}} \oplus V_{\text{morph}})).$$

3. Results

3.1. Comparative Segmentation Performance Analysis

Table 2 presents comparative analysis results on the primary test set derived from Kvasir-SEG and CVC-ClinicDB. These datasets are characterized by high image quality and standardized lighting conditions, enabling assessment of maximal architectural potential. Evaluation on standard datasets (Kvasir/CVC) was conducted using 5-fold cross-validation.

The proposed YOLO11-SAM3-LoRA architecture demonstrates superior performance across all metrics on the combined Kvasir-SEG and CVC-ClinicDB test set, achieving mDice of 0.962 and mIoU of 0.927. This confirms the effectiveness of the hierarchical approach integrating YOLO11 for candidate generation with SAM 3 featuring LoRA adaptation and prompt ensembling, providing a balance between segmentation precision and recall.

Table 2

Comparative segmentation performance of the proposed architecture against state-of-the-art methods on the combined Kvasir-SEG and CVC-ClinicDB test set

Model	mDice	mIoU	FPS (A100)
U-Net [3]	0.678	0.530	55.1
PraNet [6]	0.899	0.845	45.0
ColonSegNet [21]	0.820	0.756	140
UNet++ [4]	0.821	0.710	35
ResUNet++ [22]	0.842	0.740	32.0
YOLOv8-SAM2 [23]	0.908	0.830	44.0
Polyp-PVT [7]	0.917	0.846	28
MedSAM [11]	0.951	0.907	31
YOLO11-SAM3 (Ours)	0.962	0.927	38.2

Compared to U-Net (mDice = 0.678, mIoU = 0.530), the improvement in mDice is 28.4 percentage points (pp) and in mIoU is 39.7 pp. ColonSegNet (mDice = 0.820, mIoU = 0.756) achieves high throughput (140 FPS), surpassing U-Net by 14.2 pp in mDice, but underperforming the proposed system by 14.2 pp in mDice and 17.1 pp in mIoU. PraNet (mDice = 0.899, mIoU = 0.845) exceeds U-Net by 22.1 pp in mDice but falls short of the proposed system by 6.3 pp in mDice and 8.2 pp in mIoU at 45.0 FPS.

UNet++ variants, including the original (mDice = 0.821, mIoU = 0.710) and ResUNet++ (mDice = 0.842, mIoU = 0.740), demonstrate improvements through nested structures and residual connections, enhancing recall and F1-score relative to baseline U-Net; however, they still underperform the proposed system by 12.0–14.1 pp in mDice and 18.7–21.7 pp in mIoU at lower processing speeds (32–35 FPS versus 38.2 FPS). The hybrid YOLOv8-SAM2 model (mDice = 0.908, mIoU = 0.830) represents a strong competitor, demonstrating good performance on CVC-ClinicDB but with notable degradation on Kvasir-SEG, indicating reduced robustness to inter-dataset variation; the proposed model surpasses it by 9.7 pp in mIoU, although YOLOv8-SAM2 achieves higher throughput (44.0 FPS).

Transformer-based and SAM-derived approaches, including Polyp-PVT (mDice = 0.917, mIoU = 0.846) and MedSAM (mDice = 0.951, mIoU = 0.907), achieve high performance through pyramidal structures and semantic refinement but underperform the proposed system by 1.1–4.5 pp in mDice and 2.0–8.1 pp in mIoU at lower throughput (28–31 FPS). The overall improvement underscores the system’s potential for minimizing polyp miss rates in real-time “optical biopsy” scenarios; however, validation on prospective clinical data and additional datasets is necessary to confirm generalizability and reduce false-positive rates.

Note: “pp” (percentage points) denotes the absolute difference between metric values: $\Delta_{pp} = (x_{new} - x_{base}) \times 100$.

3.2. Generalization Capability Assessment

For objective evaluation of the proposed architecture’s effectiveness, comparison was conducted against a broad spectrum of methods, including classical CNNs (UNet, UNet++, ResUNet++),

Table 3

Comparative zero-shot generalization performance on the independent ETIS–Larib dataset

Model	mDice	mIoU	FPS	Notes
U-Net [3]	0.398	0.248	55.1	Severe degradation on small objects and artifacts
ColonSegNet [21]	0.658	0.583	140	Moderate robustness, but misses flat lesions
PraNet [6]	0.709	0.628	45.0	Good boundary accuracy, but degradation under complex lighting
UNet++ [4]	0.762	0.616	35	Significant drop on small objects
ResUNet++ [22]	0.793	0.666	32.0	Improved boundaries, but limited generalization to flat morphologies
YOLOv8–SAM2 [23]	0.840	0.724	44.0	Unstable boundaries
Polyp–PVT [7]	0.865	0.762	28	Good generalization, but misses on artifacts
MedSAM [11]	0.815	0.688	31	Misses flat lesions
YOLO11–SAM3 (Ours)	0.884	0.792	38.2	Robust to artifacts and flat morphologies

specialized attention-based models (PraNet, Polyp–PVT), and contemporary hybrid approaches (MedSAM, YOLOv8–SAM2). To assess system robustness to challenging cases (small and flat polyps), testing was performed on the ETIS–Larib dataset without additional fine-tuning (zero-shot). Neither the YOLO11 detector nor the SAM 3 segmentor underwent fine-tuning or hyperparameter optimization on ETIS images. For each ETIS image, the detector generated bounding boxes; the box with maximum confidence was passed to SAM 3. If no box exceeded the confidence threshold τ , the algorithm returned an empty mask, and metrics for that image were recorded as zero (e.g., Dice = 0). Segmentation metrics were computed at the pixel level per image and subsequently averaged (macro-average). For robust uncertainty estimation, 95% confidence intervals were additionally calculated via bootstrap resampling across images.

Table 3 presents generalization capability (zero-shot) evaluation results on the independent ETIS–Larib dataset. This dataset was selected as a stress test due to its high proportion of small (<6 mm) and flat (sessile) polyps, as well as challenging lighting conditions, substantially differentiating it from the training distribution (Kvasir/CVC).

Analysis of generalization metrics on the independent ETIS–Larib dataset (Table 3) yields the following observations:

Performance under domain shift. Despite the expected decline in absolute metrics compared to in-domain evaluation, the proposed YOLO11–SAM3–LoRA system demonstrated the highest segmentation quality on ETIS–Larib: mDice = 0.884. The advantage over MedSAM (mDice = 0.815) was +6.9 pp, and over the nearest hybrid competitor YOLOv8–SAM2 (mDice = 0.840) was +4.4 pp. The observed superiority may be attributed to the combination of cascaded detection and domain adaptation of the segmentor (LoRA), which potentially enhances robustness to low-contrast

boundaries under varying illumination and endoscopic equipment characteristics. Relative to classical CNN models U-Net (mDice = 0.398) and UNet++ (mDice = 0.762), the advantage is +48.6 pp and +12.2 pp respectively, underscoring the limited transferability of certain CNN architectures in zero-shot scenarios.

Robustness to quality degradation. Classical convolutional networks (e.g., ResUNet++: mDice = 0.793; PraNet: mDice = 0.709) exhibit notable quality degradation on the external dataset. In contrast, the proposed system maintains metrics above 0.88, indicating more robust target class representation under domain shift. For confirmation of statistical significance of these differences, confidence intervals and/or significance tests across ETIS images should additionally be reported.

Role of prompts and clinical interpretation. Results on ETIS-Larib are consistent with the hypothesis that text prompt ensembling may reduce sensitivity to formulation variability and help focus the model on subtle mucosal changes. In particular, inclusion of formulations such as “sessile lesion” and “mucosal anomaly” in the ensemble may facilitate better localization of flat/low-contrast lesions; however, claims regarding specific effectiveness on subtypes (e.g., SSL) require separate quantitative analysis by subgroups with corresponding annotations.

Overall, the mDice = 0.884 result on ETIS-Larib demonstrates the competitiveness of the proposed system under domain shift conditions. Nevertheless, further validation on additional independent datasets and, where possible, prospective clinical data is necessary to confirm generalizability and robustness to false-positive detections.

3.3. Classification Results

Table 4 presents quantitative diagnostic performance metrics for the dual-stream classifier integrating visual (ROI) and morphological segmentation features. On the test set, the classifier achieved Accuracy = 0.891 (95% CI: 0.862–0.920). For CADx/optical diagnosis applications, the key safety metric is sensitivity to adenomatous (neoplastic) lesions, which reached Sensitivity = 0.923 (95% CI: 0.884–0.962), indicating a low false-negative rate within this sample. Specificity was 0.847 (95% CI: 0.812–0.882), reflecting a moderate tendency of the algorithm to classify borderline cases as adenomas. Such a bias toward enhanced sensitivity is often considered preferable from the perspective of minimizing clinically significant lesion misses; however, it is accompanied by increased false-positive decisions and may potentially lead to additional interventions and costs.

The integrated metric AUROC = 0.942 (95% CI: 0.923–0.961) indicates high discriminative ability across the full range of classification thresholds. The relatively narrow confidence intervals suggest estimation stability on this test set (under the chosen CI computation method). Confusion matrix analysis revealed that a substantial proportion of false-positive detections within the binary labeling scheme corresponded to sessile serrated lesions (SSL), which may morphologically mimic adenomas and represent a known diagnostic challenge even for expert endoscopists [24]. Integration of the morphological stream partially reduces the frequency of such errors; further improvement may be achieved through extension to multiclass classification with SSL as a separate class, potentially improving specificity without substantial sensitivity loss.

3.4. Ablation Study

The conducted ablation analysis, results of which are presented in Table 5, enabled quantitative assessment of each architectural component’s contribution to overall system performance. Comparisons were performed across 5 mDice values (one per fold); paired Wilcoxon tests ($\alpha = 0.05$) were employed. Differences between the full model and ablation configurations were evaluated at the fold level of 5-fold cross-validation. For each configuration, mDice was computed separately per fold,

Table 4

Performance metrics of the dual-stream classifier for binary histological polyp classification. $N_{\text{total}} = 470$, $N_{\text{adenoma}} = 272$, $N_{\text{non-adenoma}} = 198$, TP = 251, FN = 21, Sens = $251/272 = 0.923$, TN = 168, FP = 30, Spec = $168/198 = 0.848$. Confidence intervals were estimated via clustered bootstrap by polyps/patients (2,000 iterations): at each iteration, cluster resampling with replacement was performed, followed by metric recalculation across all frames of selected clusters; 95% CI defined as 2.5th and 97.5th percentiles.

Metric	Value	95% CI
Accuracy	0.891	[0.862; 0.920]
Sensitivity (Adenoma)	0.923	[0.884; 0.962]
Specificity (Non-adenoma)	0.847	[0.812; 0.882]
AUROC	0.942	[0.923; 0.961]

after which pairwise differences (Full – Ablation) were compared using the Wilcoxon signed-rank test (two-sided). p -values were corrected via the Holm method for multiple comparisons. Effect size is reported as rank-biserial correlation (r_{rb}). Additionally, 95% confidence intervals for mean ΔmDice were calculated via bootstrap across folds. Sequential component exclusion revealed the following significance hierarchy:

Role of the specialized detector (YOLO11). The greatest quality reduction was observed when replacing the external YOLO11 detector with the built-in SAM 3 localization mechanism (DETR): mDice decreased from 0.962 to 0.933 ($\Delta\text{mDice} = -0.029$). This indicates that under endoscopic conditions, candidate localization quality is a critical factor for stable segmentation: the specialized detector better rejects typical artifacts (mucosal folds, specular reflections, air bubbles) and provides the segmentor with more accurate spatial prompts.

Significance of semantic ensembling. Replacing the prompt ensemble with a single “polyp” prompt led to mDice reduction of 0.009 ($0.962 \rightarrow 0.953$). Despite the small absolute magnitude, this component enhances robustness to formulation variability and, according to qualitative error analysis observations, particularly affects “difficult” cases (flat/low-contrast lesions). For quantitative confirmation of this effect, stratified error analysis by subgroups (flat/SSL, etc.) and/or external dataset results should additionally be reported.

Comparison of LoRA and full fine-tuning. Replacing LoRA adaptation with full fine-tuning yielded similar mDice values (0.962 vs. 0.959; $\Delta\text{mDice} = -0.003$; NS), indicating comparable effectiveness of parameter-efficient adaptation within this protocol. Practically, this means LoRA can achieve comparable quality at substantially reduced training computational costs. For claims regarding domain shift robustness, additional comparison of LoRA, full fine-tuning, and no adaptation on an external dataset (e.g., ETIS-Larib) with confidence intervals is recommended.

Thus, the proposed YOLO11 + SAM 3 (LoRA) + Prompt Ensemble configuration represents an optimal balance wherein each component addresses a specific challenge: YOLO ensures localization, LoRA provides domain adaptation, and the prompt ensemble ensures semantic stability.

The ablation analysis (Table 5) demonstrates that the specialized localization stage contributes most substantially to final segmentation accuracy: replacing the external YOLO11 detector with the built-in SAM 3 localization mechanism (DETR) leads to mDice reduction of 0.029. Abandoning the prompt ensemble in favor of a single “polyp” prompt reduces mDice by 0.009, consistent with the role of semantic stabilization in challenging/low-contrast cases. Finally, replacing LoRA adaptation with full fine-tuning yields similar mDice values ($\Delta = -0.003$), indicating comparable quality with more

Table 5

Ablation study results: quantitative assessment of architectural component contributions and statistical significance of changes. * indicates values to be computed; NS = not significant

Configuration	mDice (mean)	Δ mDice vs Full	p (Wilcoxon)	p_{adj} (Holm)	95% CI (Δ mDice)	r_{rb}
Full Model (YOLO11 + SAM3 + LoRA + Prompt Ensemble)	0.962	—	—	—	—	—
SAM3 native localization (DETR) instead of YOLO11	0.933	−0.029	*	*	*	*
Single prompt (“polyp”) instead of ensemble	0.953	−0.009	*	*	*	*
Full fine-tuning instead of LoRA	0.959	−0.003	NS	NS	*	*
No adaptation (SAM3 frozen)	0.951	−0.011	*	*	*	*

parameter-efficient tuning; statistical conclusions are confirmed by fold-level tests with multiple comparison correction (p_{adj} , Holm) and effect sizes (r_{rb}).

4. Discussion

In this study, we presented a cascaded YOLO11–SAM3–LoRA system for endoscopic CAD that combines a rapid candidate localization stage with subsequent boundary refinement using a foundation approach augmented by parameter-efficient domain adaptation. Experimental evaluation on open datasets demonstrates high segmentation metrics on in-domain testing and quality preservation under external assessment on ETIS–Larib, as well as high discriminative ability in binary histological classification. Collectively, results indicate that the combination of specialized localization, LoRA adaptation, and semantic prompts can enhance system robustness to endoscopic condition variability while preserving computational efficiency—critical for near real-time decision support scenarios.

Comparison with baseline architectures (U–Net, U–Net++, ResU–Net++, PraNet) and hybrid competitors (YOLOv8–SAM2), presented in Tables 2–3, demonstrates the advantage of the proposed approach. Specifically, the difference relative to YOLOv8–SAM2 was +1.8 pp in mDice on combined in-domain evaluation (Kvasir–SEG + CVC–ClinicDB) and +4.4 pp in mDice on ETIS–Larib (external zero-shot). These differences may be attributed to cascaded localization characteristics and subsequent mask refinement at the individual frame level, as well as parameter-efficient domain adaptation of the segmentation component (LoRA). For rigorous causal inference, additional comparison of SAM2 and SAM3 under identical localization and inference protocols is required. These findings align with recent comprehensive reviews of intelligent colonoscopy systems [25].

4.1. Generalization Capability and Semantic Anchoring

A key indicator of system transferability is external evaluation results on ETIS–Larib (Table 3), where under domain shift conditions the model maintained high values (mDice = 0.884; mIoU = 0.792). Ablation analysis (Table 5) demonstrates that prompt ensemble usage provides measurable contribution to the final metric (Δ mDice relative to single prompt). The observed advantage is consistent with the hypothesis that prompt ensembling reduces sensitivity to formulation variability

and may be particularly useful for low-contrast/flat lesions. Compared to baseline models, the advantage reaches +6.9–48.6 pp in mDice (depending on baseline), underscoring the limited capability of certain CNN architectures in zero-shot scenarios and the potential role of prompt engineering in enhancing foundation approach transferability.

4.2. Error Analysis and Clinical Interpretation

Confusion matrix analysis reveals that a substantial proportion of false-positive detections (FP) in the current binary formulation is associated with sessile serrated lesions (SSL), which may morphologically mimic adenomas and represent a challenging differential diagnosis even for expert endoscopists [24]. The proportion of SSL among FP should be specified as X% (Y/Z) to avoid formulation ambiguity. High adenoma sensitivity (Sensitivity = 92.3%; Table 4) reduces false-negative risk within this test sample, while specificity of 84.7% reflects bias toward “cautious” classification. However, compliance with clinical criteria for resect-and-discard strategies (e.g., PIVI-oriented NPV/high-confidence metrics and surveillance interval agreement) was not evaluated in this work; therefore, clinical implications should be interpreted as preliminary. For comparisons with “typical AUROC values,” references to relevant CADx studies should be provided.

4.3. Computational Efficiency and Deployment

Inference throughput of 38.2 FPS on NVIDIA A100 indicates potential for integration into high-performance endoscopic systems under comparable settings (input resolution, batch = 1, FP16/FP32 mode, and inclusion/exclusion of post-processing). Preliminary acceleration assessment using TensorRT INT8 for Jetson Orin indicates potential performance of 12–15 FPS under specified conditions; final throughput depends on specific device variant, input resolution, and whether timing is end-to-end or neural network forward-pass only.

4.4. Limitations and Regulatory Considerations

This study is retrospective in nature and based on open de-identified data; limitations include absence of prospective clinical validation in real workflow, potential frame correlation, and binary CADx formulation without SSL designation as a separate class. Future work involves multicenter validation and assessment of impact on clinically meaningful endpoints (e.g., ADR), as well as preparation for SaMD regulatory evaluation (FDA/CE), including reproducibility, calibration, interpretability, and risk management considerations.

5. Conclusions

In this work, we developed and experimentally evaluated a hybrid CAD architecture for endoscopy that integrates polyp localization, segmentation, and binary histological stratification. The proposed approach is based on hierarchical integration of the YOLO11 detector with SAM 3 foundation segmentation, augmented by parameter-efficient LoRA domain adaptation, achieving high segmentation metrics on in-domain testing and quality preservation under external evaluation on ETIS-Larib (mDice = 0.884).

Ablation analysis demonstrates measurable contribution of semantic text prompts: prompt ensembling reduces model sensitivity to formulation variability and promotes more robust behavior under domain shift. Collectively, results indicate that the combination of specialized localization, LoRA adaptation, and semantic prompts can enhance practical applicability of foundation approaches for endoscopic analysis, including processing of low-contrast and morphologically ambiguous cases.

From an applied perspective, achieved adenoma sensitivity (92.3%) and computational throughput (38.2 FPS on NVIDIA A100) demonstrate potential for integration into near real-time decision support workflows. However, clinical conclusions (including resect-and-discard scenarios) require separate prospective validation on multicenter data with reporting of metrics directly related to clinical endpoints (e.g., NPV/high-confidence mode and surveillance interval agreement), as well as analysis of impact on detection rates in real workflow (including ADR) and error profiles stratified by subtypes (including SSL).

Author Contributions: Conceptualization, methodology, software, Shchetinin Eu. Yu.; validation, formal analysis, supervision, Sevastianov L. A.; writing—original draft preparation, review and editing, Tiutiunnik A. A. All authors have read and agreed to the published version of the manuscript.

Funding: The work was carried out with the financial support of the Sevastopol State University, project 42-01-09/319/2025-1.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: The authors thank the creators of the open Kvasir, CVC, and ETIS-Larib datasets.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

- [1] J. C. van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. van Deventer, and E. Dekker, “Polyp miss rate determined by tandem colonoscopy: a systematic review,” *The American journal of gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006. DOI: 10.1111/j.1572-0241.2006.00390.x
- [2] D. K. Rex, C. Kahi, M. O’Brien, et al., “The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps,” *Gastrointestinal endoscopy*, vol. 73, no. 3, pp. 419–422, 2011. DOI: 10.1016/j.gie.2011.01.023
- [3] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351, Springer, Cham, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28
- [4] Z. Zhou, M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS 2018*, ser. Lecture Notes in Computer Science, vol. 11045, Springer, Cham, 2018, pp. 3–11. DOI: 10.1007/978-3-030-00889-5_1
- [5] F. 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. DOI: 10.1016/j.isprsjprs.2020.01.013
- [6] D.-P. Fan and et al., “PraNet: Parallel Reverse Attention Network for Polyp Segmentation,” in *Proceedings of the 2020 International Conference on Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 12266, Springer, Cham, 2020, pp. 263–273. DOI: 10.1007/978-3-030-59725-2_26
- [7] B. Dong and et al., *Polyp-PVT: Polyp segmentation with pyramid vision transformers*, 2021.

- [8] C. Wang, I. Yeh, and H.-Y. Liao, *YOLOv9: Learning what you want to learn using programmable gradient information*. 2024.
- [9] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, *YOLOv10: Real-time end-to-end object detection*. 2024.
- [10] A. Kirillov and et al., *Segment Anything*, 2023.
- [11] J. Ma, Y. He, F. Li, and et al., “Segment anything in medical images,” *Nature Communications*, vol. 15, pp. 1–9, 2024. DOI: 10.1038/s41467-024-44824-z
- [12] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, and et al., *SAM 2: Segment anything in images and videos*. 2024.
- [13] E. Shchetinin and A. Tiutiunnik, “On Segmentation of Polyps Using Segment Anything Model,” *Research result. Information technologies*, vol. 10, no. 3, pp. 55–63, 2025. DOI: 10.18413/2518-1092-2025-10-3-0-5
- [14] D. Jha, P. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. Johansen, “Kvasir-SEG: A segmented polyp dataset,” in *MultiMedia Modeling*, ser. Lecture Notes in Computer Science, vol. 11962, Springer International Publishing, 2020, pp. 451–462. DOI: 10.1007/978-3-030-37734-2_37
- [15] J. Bernal, F. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “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. DOI: 10.1016/j.compmedimag.2015.02.007
- [16] Y. Tudela, M. Majó, N. de la Fuente, and et al., “A complete benchmark for polyp detection, segmentation and classification in colonoscopy images,” *Frontiers in Oncology*, vol. 14, pp. 1–19, 2024. DOI: doi.org/10.3389/fonc.2024.1417862
- [17] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 283–293, 2014. DOI: doi.org/10.1007/s11548-013-0926-3
- [18] G. Jocher and J. Qiu, *Ultralytics YOLO11 [Internet]. Version 11.0.0. GitHub*, 2024.
- [19] N. Carion, L. Gustafson, Y. Hu, and et al., *SAM 3: Segment anything with concepts*. 2025.
- [20] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, and et al., *LoRA: Low-rank adaptation of large language models*. 2022.
- [21] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning,” *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021. DOI: 10.1109/ACCESS.2021.3063716
- [22] D. Jha, P. Smedsrud, M. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. Johansen, “ResUNet++: An advanced architecture for medical image segmentation,” in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, ser. Lecture Notes in Computer Science, IEEE, 2019, pp. 225–230. DOI: 10.1109/ISM46123.2019.00049
- [23] M. Mansoori, S. Shahabodini, J. Abouei, K. Plataniotis, and A. Mohammadi, *Self-prompting polyp segmentation in colonoscopy using hybrid YOLO-SAM 2 model*. 2024.
- [24] D. Rex, D. Ahnen, J. Baron, and et al., “Serrated lesions of the colorectum: Review and recommendations from an expert panel,” *American Journal of Gastroenterology*, vol. 107, no. 9, pp. 1315–1329, 2012. DOI: 10.1038/ajg.2012.161
- [25] G. Ji, J. Liu, P. Xu, N. Barnes, F. Khan, S. Khan, and D. Fan, *Frontiers in intelligent colonoscopy*. 2025.

Information about the authors

Shchetinin, Eugene Yu.—Doctor of Physical and Mathematical Sciences, Professor of the Department of Information Technology and Systems of the Sevastopol State University (e-mail: riviера-molto@mail.ru, ORCID: 0000-0003-3651-7629, ResearcherID: O-8287-2017, Scopus Author ID: 16408533100)

Sevastianov, Leonid A.—Doctor of Physical and Mathematical Sciences, Professor of the Department of Mathematical Modeling and Artificial Intelligence of the RUDN University (e-mail: sevastianov-la@rudn.ru, ORCID: 0000-0002-1856-4643, ResearcherID: B-8497-2016, Scopus Author ID: 8783969400)

Tiutyunnik, Anastasia A.—PhD in Physics and Mathematics, Associate Professor of the Department of Mathematical Modeling and Artificial Intelligence of the RUDN University (e-mail: tyutyunnik-aa@rudn.ru, ORCID: 0000-0002-4643-327X, Scopus Author ID: 55978487200)

RUSSIAN METADATA

УДК 004.89; 004.93

PACS 07.05.Mh 07.05.Tp 42.62.Be

DOI: 10.22363/2658-4670-2026-34-2-260-273

EDN: JMNMPUP

Эффективная сегментация и классификация полипов с помощью YOLO11–SAM3 и LoRA

Е. Ю. Щетинин¹, Л. А. Севастьянов², А. А. Тютюнник²

¹ Севастопольский государственный университет, ул. Университетская, д. 33, Севастополь, 299053, Российская Федерация

² Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

Аннотация. Цель Разработать и экспериментально оценить автономную систему компьютерной поддержки для колоноскопии, объединяющую сегментацию полипов и бинарную гистологическую классификацию (аденома/не-аденома) с сохранением вычислительной эффективности для near real-time применения. Материал и методы Предложена каскадная архитектура YOLO11–SAM3–LoRA: детектор YOLO11–medium формирует кандидаты (bounding boxes), после чего SAM3 выполняет box-prompted сегментацию с параметро-эффективной доменной адаптацией (LoRA) и механизмом семантического якорения на основе ансамбля текстовых промптов. Гистологическая классификация реализована двухпоточным классификатором, объединяющим признаки ROI (EfficientNet–B0) и морфологические дескрипторы маски. Обучение проведено на объединённом наборе Kvasir–SEG и CVC–ClinicDB (детекция/сегментация) и CVC–HDClassif (классификация). Переносимость оценивалась на внешнем датасете ETIS–Larib в режиме external zero-shot. Результаты На основной тестовой выборке достигнуты mDice = 0,962 для сегментации, чувствительность к аденомам (Sensitivity) = 92,3% и AUROC = 0,942 в задаче бинарной классификации. Во внешнем тестировании на ETIS–Larib (zero-shot) система сохраняет качество сегментации (mDice = 0,884), что указывает на устойчивость к доменному сдвигу. Скорость end-to-end обработки составляет 38,2 FPS на NVIDIA A100. Заключение Результаты показывают, что сочетание каскадной локализации, foundation-сегментации с LoRA-адаптацией и ансамблирования текстовых промптов обеспечивает высокое качество сегментации и конкурентоспособную точность CADx при сохранении вычислительной эффективности. Для клинического внедрения необходима дальнейшая проспективная многоцентровая валидация и анализ ошибок по подтипам поражений.

Ключевые слова: сегментация полипов, классификация полипов, YOLO11, SAM 3, LoRA, ансамбль промптов, CAD, CADx, колоноскопия, глубокое обучение