PnPD: Plug-and-Play Dual-Path Framework for Computer-Aided Detection with Decoupled Localization and Subtype Classification

Yung-Han Chen[1] HAVENCHEN.EE11@NYCU.EDU.TW
Jian-Yu Jiang-Lin[2] JIANYU@CMLAB.CSIE.NTU.EDU.TW
Tsung-Hsing Chen[3] ITOCHENYU@GMAIL.COM
Chang-Fu Kuo[3,4] ZANDIS@GMAIL.COM
Hung-Yu Wu [5] MILES.WU@DAIKSO.COM
Wen-Huang Cheng[2] WENHUANG@CSIE.NTU.EDU.TW
Juinn-Dar Huang[1] JDHUANG@NYCU.EDU.TW

[1] *National Yang Ming Chiao Tung University, Hsinchu, Taiwan*
[2] *National Taiwan University, Taipei, Taiwan*
[3] *Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[4] *School of Medicine, Chang Gung University, Taoyuan, Taiwan*
[5] *Daikso, Taipei, Taiwan*

Editors: Under Review for MIDL 2026

Abstract

Colonoscopy-based computer-aided detection (CADe) requires reliable polyp localization alongside clinically meaningful subtype classification. However, modern single-stage detectors optimize both tasks on a shared representation, creating a fundamental feature conflict: localization relies on boundary cues, while subtyping demands intra-lesion fine-grained textural details. To resolve this inherent limitation, we propose **PnPD**, a **P**lug-a**n**d-**P**lay **D**ual-path framework that decouples localization and classification into independent, parallel streams. This modular design allows for the flexible integration of specialized models, utilizing lightweight object detectors exclusively for bounding box regression while reformulating subtype classification as a segmentation task. By replacing the standard classification head with a segmentation module, our approach explicitly forces the model to capture dense, pixel-level mucosal patterns. Extensive experiments demonstrate that our framework effectively resolves the feature misalignment observed in unified models. Remarkably, lightweight PnPD configurations match or surpass the detection F1-scores of large and x-large YOLO baselines while reducing parameters and MACs by over 80%, simultaneously improving average subtype precision and recall by more than 9% and 7%. This offers a scalable solution tailored for efficient deployment in real-time clinical applications.

Keywords: colonoscopy, object detection, semantic segmentation

1. Introduction

Colonoscopy is the gold standard for preventing colorectal cancer (CRC), yet its efficacy varies among endoscopists (Siegel et al., 2025). Computer-aided detection (CADe) systems have successfully addressed this by significantly improving polyp detection rates, a benefit now explicitly endorsed by clinical guidelines (Makar et al., 2025; Bretthauer et al., 2025;

Sultan et al., 2025). However, current clinical needs extend beyond mere detection. The focus is increasingly shifting toward multi-class detection, which simultaneously localizes and subtypes lesions to offer real-time risk stratification. Such capabilities are essential for streamlining workflows and enhancing population-level prevention. Addressing this need, our work introduces a framework designed for efficient, real-time multi-class polyp detection.

Despite this need for real-time multi-class polyp detection, the dominant paradigm in CADe remains *localization-centric*. Built primarily upon one-stage object detectors such as the YOLO family, these systems prioritize the rapid regression of bounding boxes (Pacal and Karaboga, 2021; Lalinia and Sahafi, 2024). Modern variants achieve impressive sensitivity and low latency through lightweight backbones (Zhao et al., 2024; Viet et al., 2025), yet they typically treat all lesions as a single foreground category. By design, these architectures optimize spatial boundary delineation and often treat the fine-grained internal textures required for subtype differentiation as secondary.

In addition to these localization-centric efforts, a separate stream of research concentrates on *polyp classification*. These approaches utilize deep architectures to classify specific polyp subtypes, often relying on high-magnification or cropped images to analyze mucosal patterns (Byrne et al., 2019; Korbar et al., 2017; Younas et al., 2023). To resolve histologically similar lesions, modern classifiers employ complex multi-branch or attention-based designs that capture intricate texture cues (Li et al., 2025b,a). However, while these models excel at feature extraction and subtype discrimination, they typically lack the inherent capability to localize lesions within full video frames under real-time constraints.

While integrating localization and classification into a single unified framework is the logical next step, it presents significant challenges due to a fundamental *feature representation conflict*. Standard multi-task learning relies on a shared backbone (Wu et al., 2020; Standley et al., 2020), yet the tasks demand distinct features: robust box regression requires boundary-oriented cues, whereas polyp subtyping depends on fine-grained intra-lesion textures. Consequently, optimizing these diverging objectives within a shared representation creates an inherent trade-off between spatial delineation and subtype reasoning. Beyond this theoretical conflict, monolithic architectures also suffer from *operational rigidity*. Any update to the subtyping criteria or architectural improvements in one task necessitates retraining the entire network, consuming excessive computational resources and potentially destabilizing the performance of the other task. This creates a need for a modular paradigm where localization and classification can be maintained and optimized independently.

To address these limitations, we propose **PnPD**, a **P**lug-a**n**d-**P**lay **D**ual-path framework that structurally resolves this conflict by decoupling localization and classification into parallel streams. Crucially, instead of treating subtyping as a standard classification task, we reframe it as a *dense segmentation problem*. This design ensures that the total inference latency is constrained only by the slower branch rather than the sum of both, thereby preserving real-time performance. Our main contributions are as follows:

- We propose a decoupled architecture that delegates localization to a dedicated detection network and subtyping to a specialized semantic segmentation model, effectively bypassing the optimization bottlenecks of shared-backbone designs.

- We reframe polyp subtyping as a *dense segmentation task*. By utilizing pixel-wise supervision instead of image-level labels, our approach explicitly preserves the fine-grained texture evidence essential for accurate histologic classification.
- We introduce a resolution-agnostic, voting-based fusion mechanism that integrates bounding boxes and segmentation masks at the output level. This plug-and-play design enables flexible combinations of lightweight detectors and segmentation networks, achieving superior accuracy–efficiency trade-offs compared with single-path baselines while maintaining real-time feasibility.

2. Related Work

2.1. Localization-Centric Architectures and the Texture Gap

Driven by the need for high-throughput screening, the YOLO family has become the dominant choice for real-time CADe (Pacal et al., 2022; Gündüz and Işık, 2023; Ragab et al., 2024). Recent versions such as YOLOv12 and YOLOv13 improve the trade-off between latency and accuracy through multi-scale aggregation and attention-enhanced feature interaction, including Area Attention and hypergraph-based correlation fusion (Tian et al., 2025; Lei et al., 2025). These designs are highly effective for stabilizing box regression and improving detection robustness. However, they are not explicitly tailored to preserve the subtle intra-lesion mucosal textures that underpin histological subtyping. As a result, YOLO-based CADe systems can be highly sensitive to polyp presence and boundaries while exhibiting a *texture gap* for distinguishing subtypes of lesions.

2.2. Texture Characterization: From Classification to Dense Prediction

To recover subtype cues that may be under-emphasized by real-time detectors, recent polyp classifiers attempt to preserve fine textures through specialized designs, such as fused residual attention in FRAN and dual-branch interactions that combine transformer and CNN-Wavelet features in DMFI-Net (Li et al., 2025b,a). Yet these methods are still trained with image-level supervision, so spatial evidence is eventually compressed into a global prediction, limiting explicit localization of discriminative mucosal patterns. In contrast, medical image segmentation provides dense supervision that encourages high-resolution feature retention across the encoder-decoder hierarchy. Polyp-PVT exemplifies this texture-preserving paradigm in a polyp-specific setting (Dong et al., 2021), while EMCAD offers an efficient multi-scale attention decoder applicable across diverse modalities (Rahman et al., 2024). Moreover, UNeXt suggests that lightweight, transformer-free dense models can still deliver strong detail-aware representations (Valanarasu and Patel, 2022). Motivated by these properties, we repurpose segmentation as a spatially aware subtype classifier to complement detector-centric localization in our dual-path design.

2.3. The Dilemma of Existing Multi-class Strategies

Recent studies explore unified detectors that simultaneously localize and classify polyps, sometimes using lightweight ensembles to maintain real-time speed (Zhao et al., 2024). However, unified designs couple box quality with classification, where localization errors

directly degrade subtype assignment. This coupling also reduces maintainability, as updating subtype features requires retraining the detector. Alternatively, cascaded pipelines like PolyDSS (Saad et al., 2024) operate sequentially, increase latency and complicating multi-polyp handling. Our PnPD framework addresses these limitations by decoupling localization and classification, enabling independent optimization and flexible fusion.

3. Methods

3.1. Preliminary

To motivate our dual-path design, we apply Grad-CAM (Selvaraju et al., 2017) to a YOLOv13 model, as shown in Figure 1. We target layers 4, 6, and 8, corresponding to feature pyramid levels $P3$, $P4$, and $P5$, where P_n denotes a spatial resolution of $1/2^n$ relative to the input image. In panel (a), the Grad-CAM of the classification head often spills outside the polyp and, even within the bounding box, concentrates on the lesion boundary rather than the internal mucosal pattern. In contrast, panel (b) shows that the box regression head consistently focuses along the polyp boundary at all scales. This indicates that localization is driven mainly by edge cues, which conflicts with prior work emphasizing intra-lesion surface and vascular textures for subtype classification (Li et al., 2025b,a).

Next, we complement these qualitative observations with a quantitative analysis of how the shared features feeding the YOLOv13 prediction head are utilized by the two tasks, as shown in Figure 2. For each feature channel, we estimate its normalized importance for the box branch and for the class branch, and plot these values on the x- and y-axes, respectively. Many channels lie clearly below the diagonal, indicating features that strongly influence localization but contribute little to classification. The proportion of channels with higher importance for the box branch than for the class branch reaches 74% in $P3$, 56% in $P4$, and 62% in $P5$, consistently favoring localization across all scales.

Taken together, these results suggest that the shared representation in a standard detector cannot serve localization and subtype classification equally well: under the current training framework, the common features are biased toward localization. Motivated by this imbalance, we design a dual-path framework that keeps detection as a dedicated localization module and relies on a parallel segmentation path to perform subtype classification within the detector boxes.

3.2. Dual-path Framework

The proposed PnPD framework, shown in Figure 3, is founded on a strict **architectural decoupling** designed to resolve the inherent feature conflict in polyp analysis. While accurate localization requires translation variance to capture precise boundary details, robust subtype characterization demands translation invariance to focus on internal texture patterns regardless of spatial position. Conventional multi-task learning approaches force a shared backbone to learn a compromise representation between these opposing objectives. In contrast, our framework imposes zero gradient coupling between the two paths. The interaction occurs solely at the inference level through a standardized bounding box interface.

This design is inherently plug-and-play because the detection branch functions as a universal region proposer while the segmentation branch acts as a specialized dense pixel-

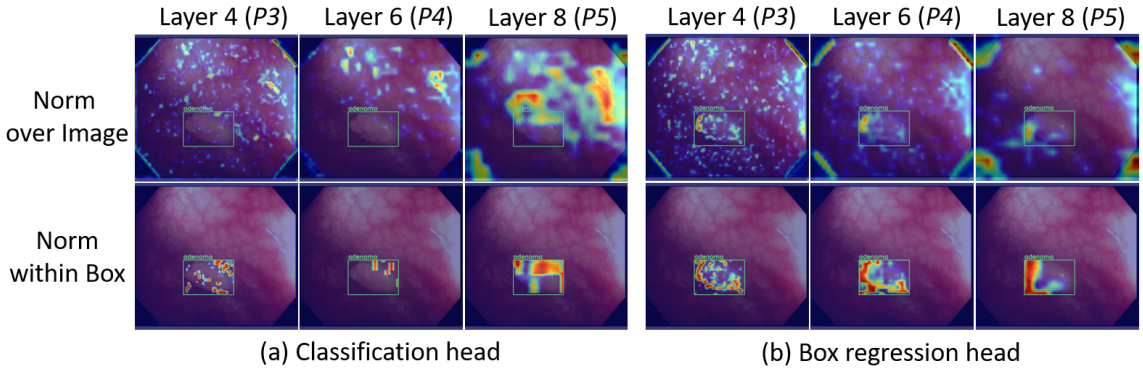

Figure 1: Grad-CAM visualization of YOLOv13. Heatmaps are generated by backpropagating gradients from (a) the classification head and (b) the box regression head, targeting layers 4, 6, and 8. In each panel, the top row displays Grad-CAM values normalized over the entire image, while the bottom row shows values normalized only within the predicted bounding box.

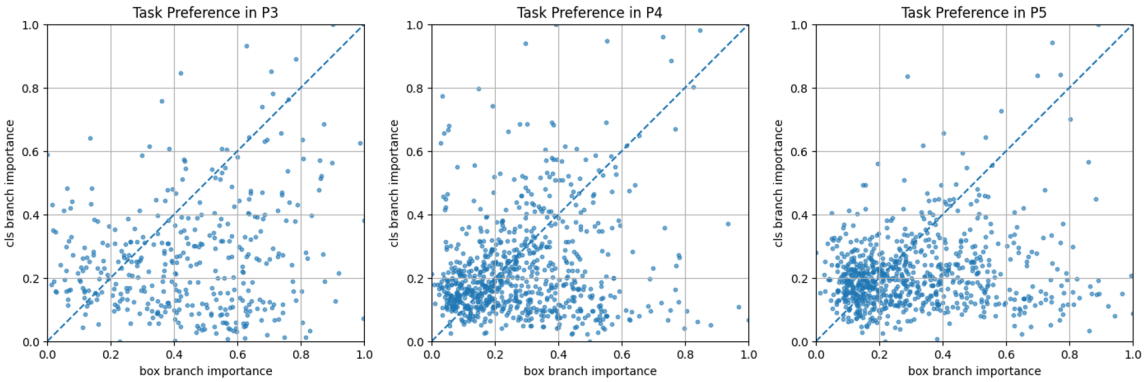

Figure 2: Task preference of shared features in the YOLOv13 prediction head. Each scatter plot shows, for a shared feature channel, its normalized importance for localization (box branch) versus classification (class branch), reflecting how the common representation is used by the two tasks. From left to right, the plots correspond to three different feature-map resolutions.

wise classifier. Consequently, either component can be replaced or upgraded independently. This flexibility allows the system to leverage the latest state-of-the-art real-time detectors or medical segmentation networks without requiring joint architectural redesign or retraining of the entire pipeline.

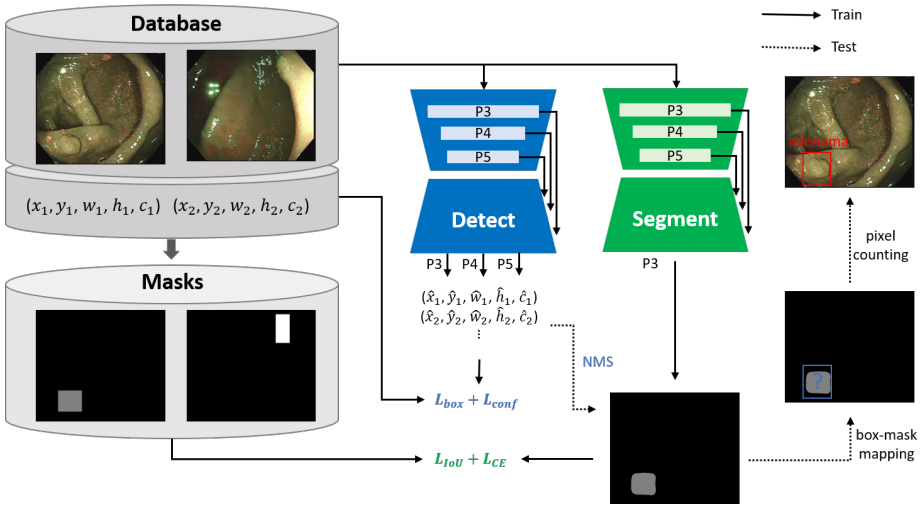

Figure 3: Overview of the proposed P^2D framework. The pipeline decouples the tasks into two parallel streams including a detection path for localization and a segmentation path for subtype characterization. The original bounding box annotations are converted into mask supervision for the segmentation branch. Solid arrows indicate the independent training flows while dashed arrows denote the test-time fusion workflow that integrates outputs via coordinate projection.

3.2.1. Training Phase

During the training phase, the two paths are optimized independently using supervision derived from the same bounding box annotations.

Detection Path. We reformulate the multi-class problem into a binary classification task, distinguishing polyps from the background. Consequently, the classification branch predicts a single objectness confidence score rather than a categorical distribution. For each predicted bounding box parameterized by $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$, the head outputs a scalar logit representing the presence probability after sigmoid activation. This confidence term is optimized via binary cross-entropy and combined with the CIoU loss for bounding box regression to form the total detection loss: $L_{det} = \lambda_{box} L_{box} + \lambda_{conf} L_{conf}$, where λ_{box} and λ_{conf} serve as balancing coefficients.

Segmentation Path. We supervise the segmentation path using coarse-grained pseudo-masks generated directly from bounding box annotations. Pixels falling within the spatial extent of a ground-truth box are assigned the corresponding subtype label, while those outside are treated as background. This formulation results in a $(C + 1)$-class dense classification task where C denotes the number of polyp subtypes. The network estimates a pixel-wise softmax distribution and is optimized via a compound objective that balances classification accuracy with structural consistency: $L_{seg} = \lambda_{iou} L_{IoU} + \lambda_{ce} L_{CE}$, where λ_{iou} and λ_{ce} are weighting coefficients.

Crucially, this independent training regime resolves the optimization conflict by decoupling the learning objectives. The detection path is unconstrained in learning translation variance for boundary delineation, while the segmentation path focuses purely on translation invariance for texture recognition. This ensures that feature extraction is fully specialized for each specific task, effectively bypassing the shared-head bottleneck.

3.2.2. Test Phase

In the test phase, the detection and segmentation models process the input image in parallel. First, the detection path generates candidate bounding boxes, which are filtered via non-maximum suppression (NMS) to retain only high-confidence proposals (Hosang et al., 2017). To address potential resolution mismatches between the two paths, we employ a **resolution-agnostic fusion** mechanism based on dynamic coordinate projection. Let the detection input space be $H_D \times W_D$ and the segmentation output map be $H_S \times W_S$. For a predicted box B_i in the detection space, we project it to the segmentation space B_i' via linear scaling:

$$B_i' = B_i \times \frac{(W_S, H_S)}{(W_D, H_D)} \tag{1}$$

This linear projection ensures that lightweight detectors running at low resolutions can be seamlessly paired with high-fidelity segmentation models running at native resolution without manual alignment.

Once the region of interest is projected, we assign a specific subtype via **vote-based class assignment**. Given the segmentation probability map M, the score for each foreground class k is calculated as the sum of probabilities for all pixels p strictly within B_i':

$$Score_k = \sum_{p \in B_i'} M(p, k) \tag{2}$$

The final label is determined by the maximum score among foreground classes. Crucially, we explicitly exclude the background class from this voting process. This design relies on the detection branch acting as a rigorous *objectness gatekeeper*. Since the localized boxes have already passed high-confidence filtering, the fusion stage effectively models the conditional probability of the subtype given the presence of a lesion. This prevents background noise within the bounding box from dominating the vote, ensuring the decision is driven solely by discriminative textural evidence.

4. Experiments and Results

4.1. Dataset

In this study, we constructed a more reliable dataset to demonstrate the benefits of our approach. The dataset consists of clinical images collected during patient examinations at Linkou Chang-Gung Memorial Hospital (CGMH), Taoyuan, Taiwan, comprising 311 patients' examination videos. The images were annotated by professional physicians. After gathering all the data, we divided it into training and test datasets. The training dataset comprises 280 patients, totaling 6,442 images, with a total of 2,803 hyperplastic and 3,708 adenoma instances. The test dataset consists of 31 patients, totaling 704 images, with a total

of 323 hyperplastic and 411 adenoma instances. In contrast to the public dataset(Bernal et al., 2017; Mesejo et al., 2016), our CGMH polyp dataset addresses a broader range of scenarios: (a) multiple polyps appearing in a single frame, (b) different categories of polyps present in the same frame, and (c) polyps before closing up to make the training and test phases more representative of real-world conditions.

4.2. Comparison Results

Table 1: Comparison of detection metrics for single-path and dual-path frameworks.

Framework	Model	$AP_{50:95}$	AP_{50}	F1-score	Model	$AP_{50:95}$	AP_{50}	F1-score
v13								
Single-path	v13n	47.22	77.48	73.25	v13l	55.78	87.13	82.24
	v13s	52.60	84.19	79.31	v13x	56.07	87.74	82.39
Dual-path	v13n+seg	55.16	87.26	82.57	v13l+seg	**57.22**	**89.50**	**86.58**
	v13s+seg	55.86	87.54	84.26	v13x+seg	57.12	89.39	84.89
v12								
Single-path	v12n	53.75	85.04	80.48	v12l	55.66	87.41	83.49
	v12s	57.08	89.20	85.10	v12x	56.85	87.99	83.06
	v12m	57.56	89.41	84.50				
Dual-path	v12n+seg	56.01	87.59	84.72	v12l+seg	57.97	89.66	86.30
	v12s+seg	57.83	90.76	87.24	v12x+seg	58.26	90.07	86.41
	v12m+seg	**59.00**	**91.49**	**87.84**				
v11								
Single-path	v11n	54.00	84.76	79.64	v11l	56.51	87.27	81.56
	v11s	55.12	86.43	81.61	v11x	56.96	88.70	83.63
	v11m	56.08	88.97	84.05				
Dual-path	v11n+seg	53.68	86.27	84.21	v11l+seg	57.37	89.64	86.18
	v11s+seg	57.39	89.07	85.11	v11x+seg	57.96	**91.05**	**87.24**
	v11m+seg	**58.44**	89.83	85.40				

4.2.1. Detection performance

Table 1 presents the quantitative comparison between the proposed dual-path framework and the single-path baselines across YOLOv11, v12, and v13 architectures (Khanam and Hussain, 2024; Tian et al., 2025; Lei et al., 2025). A defining feature of our framework is its **plug-and-play** nature; the localization task is fully decoupled from the subsequent classification stage. Consequently, the detection performance is determined solely by the chosen object detector's capability along the detection path.

This structural independence yields a significant performance advantage. The dual-path models consistently outperform their single-path counterparts across all model scales (nano through x-large) and versions. For instance, the lightweight **v13n** in the dual-path setting achieves an F1-score of 82.57%, significantly surpassing the single-path **v13n** (73.25%) and even **v13x** (82.39%). This improvement stems from the simplification of the detector's objective: by offloading the complex subtype classification to the segmentation branch, the detection model functions as a class-agnostic locator (binary detection). This allows

the network to focus its capacity entirely on precise boundary regression and objectness, effectively resolving the feature conflict between localization and classification observed in single-stage baselines.

Table 2: Subtype classification metrics for single-path and dual-path configurations.

Detect	Segment	Hyperplastic		Adenoma		Average	
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
v13							
v13l	– (Single-path)	57.88	52.32	64.86	61.07	61.37	56.70
v13x		61.29	52.94	70.35	58.88	65.82	55.91
v13n	EMCAD_b0	70.61	53.56	63.56	69.59	67.09	61.58
v13s		71.25	52.94	65.23	69.83	68.24	61.39
v13n	UNeXt	76.73	58.20	64.67	70.80	70.70	64.50
v13s		76.10	59.13	68.53	71.53	**72.32**	**65.33**
v12							
v12l	– (Single-path)	65.81	55.42	65.90	63.02	65.86	59.22
v12x		61.92	53.87	68.23	63.75	65.08	58.81
v12n	EMCAD_b0	69.20	56.35	62.99	70.80	66.10	63.58
v12s		73.88	56.04	69.58	71.78	71.73	63.91
v12m		74.58	55.42	70.28	72.51	72.43	63.97
v12n	UNeXt	73.58	60.37	64.78	72.51	69.18	66.44
v12s		77.51	59.75	70.95	72.51	74.23	66.13
v12m		78.31	60.37	72.77	73.48	**75.54**	**66.93**
v11							
v11l	– (Single-path)	56.72	58.82	63.85	60.58	60.29	59.70
v11x		59.35	51.08	68.22	64.23	63.79	57.66
v11n	EMCAD_b0	71.90	53.87	67.21	70.32	69.56	62.10
v11s		69.80	55.11	66.29	71.29	68.05	63.20
v11m		68.87	54.80	65.26	75.43	67.07	65.12
v11n	UNeXt	76.45	57.28	67.91	71.05	72.18	64.17
v11s		77.69	60.37	67.26	72.99	72.48	66.68
v11m		74.05	60.06	65.53	74.94	**69.79**	**67.50**

4.2.2. Subtype classification performance

In Table 2, we present the subtype classification results. We employed the lightweight nano (n), small (s), and medium (m) variants of the YOLO series as the detection path, integrated with two efficient segmentation models, UNeXt and EMCAD (Rahman et al., 2024; Valanarasu and Patel, 2022), for the parallel classification path. We report the subtype-wise precision and recall for hyperplastic polyps and adenomas, along with their averages. Overall, dual-path configurations using UNeXt achieve higher average precision and recall than those using EMCAD_b0 across all YOLO generations, indicating that UNeXt provides stronger per-polyp classification features.

Although the nano detectors possess weaker localization capabilities compared to the large and x-large baselines, combining them with a segmentation branch yields significantly

better subtype metrics. The most substantial gains are observed in the YOLOv12m+UNeXt configuration, which surpasses the single-path v12x by more than 10% in average precision and 8% in average recall. These results confirm that the dedicated segmentation path serves as a much stronger subtype classifier than the conventional detector's classification head.

Table 3: Complexity and inference time of single- and dual-path configurations.

Detect	**Segment**	**Param (M)**	**MACs (G)**	**Infer$_{\text{det}}$** (ms)	**Infer$_{\text{seg}}$** (ms)	**Post (ms)**
v12l	– (Single-path)	26.445	41.470	23.682	–	0.9
v12x		59.323	92.704	24.102	–	
v12s	EMCAD_b0	13.013	11.379	13.585	13.476	1.2
v12m		23.484	31.193	13.663		
v12s	UNeXt	10.569	10.875	13.585	3.852	1.2
v12m		21.040	30.689	13.663		

4.2.3. Computational efficiency

Table 3 compares parameter counts, MACs, and inference time between the single-path YOLOv12 baselines and the dual-path configurations. By pairing lightweight detectors with lightweight segmentation models, the dual-path variants use substantially fewer parameters and computations than the large and x-large single-path detectors, while still achieving higher detection and subtype classification performance. For example, v12s+UNeXt reduces parameters from 59.3 M to 10.6 M (an 82.2% reduction) and MACs from 92.7 G to 10.9 G (an 88.3% reduction) compared with v12x, and improve average subtype precision and recall by 9.4% and 7.3%

On an NVIDIA A100 GPU, the detection and segmentation branches can be executed in parallel, so the effective inference latency is dominated by the slower of the two forward passes plus a small increase in post-processing time (from 0.9 ms to 1.2 ms) due to box–mask mapping and pixel counting. Even with this extra post-processing, the critical-path latency of v12s+UNeXt ($\approx$ 14.8 ms) remains noticeably lower than that of the x-large single-path baseline v12x ($\approx$ 25.0 ms).

5. Conclusion

In this work, we addressed the inherent limitations of unified multi-class detectors, where conflicting feature requirements for localization and classification hinder overall performance. We introduced PnPD, a Plug-and-Play Dual-path framework that strategically decouples these tasks to ensure that each is optimized with its most relevant features. Our results confirm that separating the detection path (focused on boundaries) from the segmentation path (focused on internal textures) yields significant performance gains. Crucially, the plug-and-play nature of PnPD enables the flexible pairing of efficient, state-of-the-art models, allowing lightweight configurations to surpass the capabilities of large-scale single-path baselines. This architecture not only enhances diagnostic accuracy but also provides a versatile foundation for future CADe systems, where detection and classification components can be independently upgraded to adapt to evolving clinical needs.

References

Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

Michael Bretthauer, Jabed Ahmed, Giulio Antonelli, Hanneke Beaumont, Sabina Beg, Ariel Benson, Raf Bisschops, Elena De Cristofaro, Eimear Gibbons, Michael Häfner, et al. Use of computer-assisted detection (cade) colonoscopy in colorectal cancer screening and surveillance: European society of gastrointestinal endoscopy (esge) position statement. *Endoscopy*, 57(06):667–673, 2025.

Michael F Byrne, Nicolas Chapados, Florian Soudan, Clemens Oertel, Milagros Linares Pérez, Raymond Kelly, Nadeem Iqbal, Florent Chandelier, and Douglas K Rex. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*, 68(1):94–100, 2019.

Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

Mehmet Şirin Gündüz and Gültekin Işık. A new yolo-based method for real-time crowd detection from video and performance analysis of yolo models. *Journal of Real-Time Image Processing*, 20(1):5, 2023.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515, 2017.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics*, 8:30, 2017.

Mehrshad Lalinia and Ali Sahafi. Colorectal polyp detection in colonoscopy images using yolo-v8 network. *Signal, Image and Video Processing*, 18(3):2047–2058, 2024.

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhu Zheng, Guiguang Ding, Shaoyi Du, Zongze Wu, and Yue Gao. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint arXiv:2506.17733*, 2025.

Sheng Li, Bo Cao, Xiaoheng Tang, Xiongxiong He, Shufang Ye, and Yujun Rao. Dmfi-net: Dual-branch multi-scale feature interaction network integrating transformer and cnn-wavelet for image classification of colorectal polyps. *Biomedical Signal Processing and Control*, 106:107753, 2025a.

Sheng Li, Xinran Guo, Beibei Zhu, Shufang Ye, Jietong Ye, Yongwei Zhuang, and Xiongxiong He. Multi-classification of colorectal polyps with fused residual attention. *Signal, Image and Video Processing*, 19(2):144, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Jonathan Makar, Jonathan Abdelmalak, Danny Con, Bilal Hafeez, and Mayur Garg. Use of artificial intelligence improves colonoscopy performance in adenoma detection: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 101(1):68–81, 2025.

Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging*, 35(9):2051–2063, 2016.

Ishak Pacal and Dervis Karaboga. A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*, 134:104519, 2021.

Ishak Pacal, Ahmet Karaman, Dervis Karaboga, Bahriye Akay, Alper Basturk, Ufuk Nalbantoglu, and Seymanur Coskun. An efficient real-time colonic polyp detection with yolo algorithms trained by using negative samples and large datasets. *Computers in biology and medicine*, 141:105031, 2022.

Mohammed Gamal Ragab, Said Jadid Abdulkadir, Amgad Muneer, Alawi Alqushaibi, Ebrahim Hamid Sumiea, Rizwan Qureshi, Safwan Mahmood Al-Selwi, and Hitham Alhussian. A comprehensive systematic review of yolo for medical object detection (2018 to 2023). *IEEE Access*, 12:57815–57836, 2024.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024.

Abdelrahman I Saad, Fahima A Maghraby, and Osama M Badawy. Polydss: computer-aided decision support system for multiclass polyp segmentation and classification using deep learning. *Neural Computing and Applications*, 36(9):5031–5057, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Rebecca L Siegel, Tyler B Kratzer, Angela N Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *Ca*, 75(1):10, 2025.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

Shahnaz Sultan, Dennis L Shung, Jennifer M Kolb, Farid Foroutan, Cesare Hassan, Charles J Kahi, Peter S Liang, Theodore R Levin, Shazia Mehmood Siddique, and Benjamin Lebwohl. Aga living clinical practice guideline on computer-aided detection–assisted colonoscopy. *Gastroenterology*, 168(4):691–700, 2025.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, pages 23–33. Springer, 2022.

Hang Dao Viet, Tung Thanh Nguyen, Hoa Ngoc Lam, Binh Phuc Nguyen, Trung Quoc Vu, Hien Minh Nguyen, Vinh Tuan Pho, Hieu Huy Dang, Dinh Viet Sang, and Thuy Thi Nguyen. Validation of yolov8 algorithm in detecting colon polyps in endoscopy videos. *Journal of Medical Artificial Intelligence*, 8:35, 2025.

Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020.

Farah Younas, Muhammad Usman, and Wei Qi Yan. A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*, 53(2):2410–2433, 2023.

Luqing Zhao, Nan Wang, Xihan Zhu, Zhenyu Wu, Aihua Shen, Lihong Zhang, Ruixin Wang, Dianpeng Wang, and Shengsheng Zhang. Establishment and validation of an artificial intelligence-based model for real-time detection and classification of colorectal adenoma. *Scientific Reports*, 14(1):10750, 2024.

Appendix A. Training and implementation details

A.1. YOLO series Baselines

For the YOLO series baselines, we train the large (l) and x-large (x) models from scratch using the official Ultralytics implementation and YOLOv13 hyperparameters (Jocher et al., 2023), unless otherwise noted. All models are optimized with stochastic gradient descent (SGD) with an initial learning rate of 0.01, batch size of 32, and an input resolution of 640×640. Training is performed for 300 epochs with mosaic augmentation enabled (mosaic = 1.0) and scale jitter set to 0.9 for all configurations. Following the defaults, we use mixup

and copy-paste augmentations and slightly differentiate the settings for the large and x-large models: for models whose name ends with "l" we set `mixup = 0.15` and `copy_paste = 0.5`, while for all other scales (including "x") we use `mixup = 0.20` and `copy_paste = 0.6`. Apart from these scale-dependent augmentation parameters and the fixed batch size, all remaining optimization and regularization hyperparameters (weight decay, momentum, warmup schedule, etc.) follow the original YOLOv13 training configuration. All detectors are trained on a single NVIDIA A100 GPU.

A.2. Detection path in Dual-path Framework

For the detection path in the proposed dual-path framework, we also follow the default YOLOv13 training configuration and optimizer settings, and only adjust the data augmentation hyperparameters according to the detector scale. Specifically, we train the nano (n), small (s), and medium (m) models from scratch with SGD (initial learning rate 0.01, batch size 32, input resolution 640×640, 300 epochs, mosaic = 1.0, scale jitter enabled), while keeping all other optimization hyperparameters identical to the single-path baselines. The only differences lie in the values of `scale`, `mixup`, and `copy_paste`: for the nano detector we set `scale = 0.5`, `mixup = 0.0`, and `copy_paste = 0.1`; for the small detector we use `scale = 0.9`, `mixup = 0.05`, and `copy_paste = 0.15`; and for the medium detector we adopt `scale = 0.9`, `mixup = 0.20`, and `copy_paste = 0.60`. All detection-path models are trained on a single NVIDIA A100 GPU.

A.3. Segmentation path in Dual-path Framework

All segmentation models in the dual-path framework are trained with the AdamW optimizer (learning rate $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$) for 100 epochs, using a batch size of 16. We do not apply any data augmentation in the segmentation path. The training objective is a weighted combination of pixel-wise cross-entropy loss and IoU loss,

$$L_{\text{seg}} = 0.3\, L_{\text{CE}} + 0.7\, L_{\text{IoU}},$$

and the gradient norm is clipped to 0.5 to stabilize optimization. For transformer-based backbones, we compare three architectures: Polyp-PVT, EMCAD, and U-Net v2. Unless otherwise stated, these models use a PVTv2-B2 encoder. For CNN-based segmentation, we additionally evaluate UNeXt. The only exception is the EMCAD entry in Table 2, where we adopt a lighter PVTv2-B0 backbone (consistent with the original EMCAD implementation) for a fair comparison of efficiency.

Appendix B. Failure case: segmentation-only multi-class polyp detection

Before designing the dual-path framework, we attempted to perform multi-class polyp detection using only a segmentation model. As illustrated in Figure 4, the network was trained to produce pixel-wise masks with separate labels for hyperplastic and adenoma polyps, and then each connected region in the mask was converted into a predicted bounding box. Blue pixels/boxes denote regions predicted as hyperplastic and green pixels/boxes denote adenomas. Although the segmentation masks occasionally covered the lesion regions, they also contained many small, fragmented blobs scattered across the mucosal surface.

When these tiny components were transformed into boxes, they produced a large number of false positives with highly irregular sizes and aspect ratios. We experimented with several post-processing strategies (area thresholds, morphological operations, and merging nearby components), but any aggressive filtering that suppressed spurious fragments also removed true small lesions or broke apart partially segmented polyps. As a consequence, the segmentation-only pipeline could not simultaneously maintain reasonable recall and precision, and its box-level performance was markedly worse than YOLO-based detectors. This negative result motivated our final design choice: keep detection as a dedicated localization task and use segmentation primarily for subtype classification inside detector boxes, rather than relying on segmentation masks to define boxes from scratch.

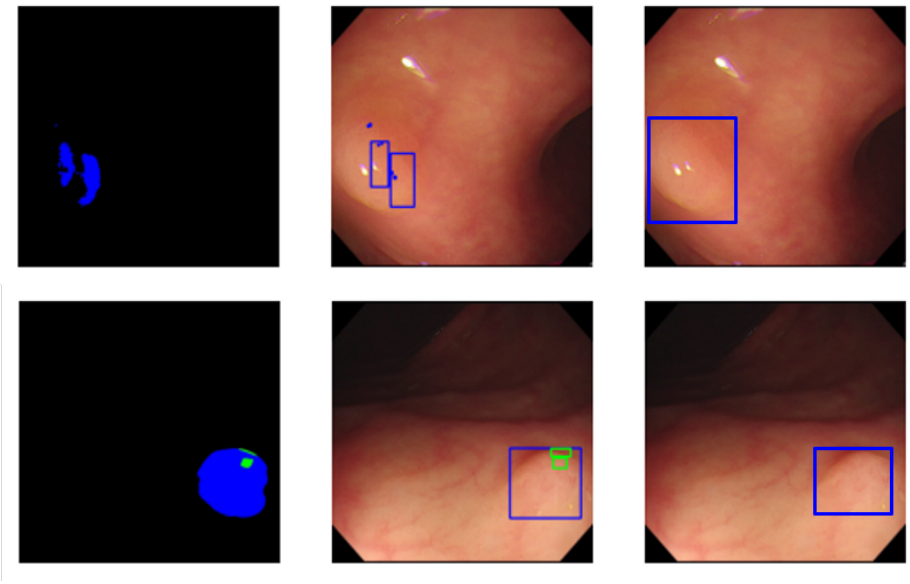

Figure 4: The left column displays the segmentation masks, where blue pixels denote regions predicted as 'hyperplastic' by the model, while green pixels represent 'adenoma.' In the middle column, the segmentation masks are transformed into predicted boxes. The boxes derived from the blue areas are depicted in blue, and those from the green areas are shown in green. On the right, the original image is displayed with ground truth boxes, whose colors correspond to the categories in the same way as in the middle column.

Appendix C. Dataset Comparison

The COCO dataset contains 80 different object classes in total (Lin et al., 2014), far more than the two classes used for polyp detection. Surprisingly, the performance of the YOLO series in classifying objects in the COCO dataset surpasses that in polyp detection. Intu-

itively, objects in the COCO dataset often possess distinct recognizable features, making them easily distinguishable despite a large number of classes.

In Table 4, we analyze the predicted boxes from the Polyp and COCO datasets, and also their categories, based on their rounded confidence values. We observe a generally higher proportion of intermediate values in the polyp dataset compared to the COCO categories. The results are depicted in Figure 5. The distribution of the COCO dataset tends to exhibit a U-shaped pattern, suggesting a lower proportion of uncertain predictions compared to the polyp dataset, resulting in lower uncertainty.

Table 4: Distribution of Confidence Level of Different Categories and Datasets.

Category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
hp	0.251	0.085	0.078	0.094	0.087	0.139	0.150	0.110	0.004	0.000
ad	0.153	0.055	0.075	0.080	0.078	0.125	0.165	0.210	0.059	0.000
Polyp	0.199	0.069	0.076	0.087	0.083	0.132	0.158	0.163	0.033	0.000
person	0.260	0.110	0.071	0.062	0.055	0.063	0.073	0.102	0.192	0.023
bicycle	0.307	0.149	0.101	0.074	0.084	0.041	0.048	0.046	0.106	0.022
car	0.332	0.122	0.078	0.068	0.058	0.060	0.077	0.101	0.096	0.007
motor	0.277	0.112	0.082	0.047	0.044	0.047	0.093	0.068	0.217	0.021
airplane	0.154	0.033	0.044	0.049	0.027	0.038	0.066	0.099	0.396	0.093
COCO	0.327	0.119	0.078	0.062	0.054	0.055	0.060	0.082	0.139	0.024

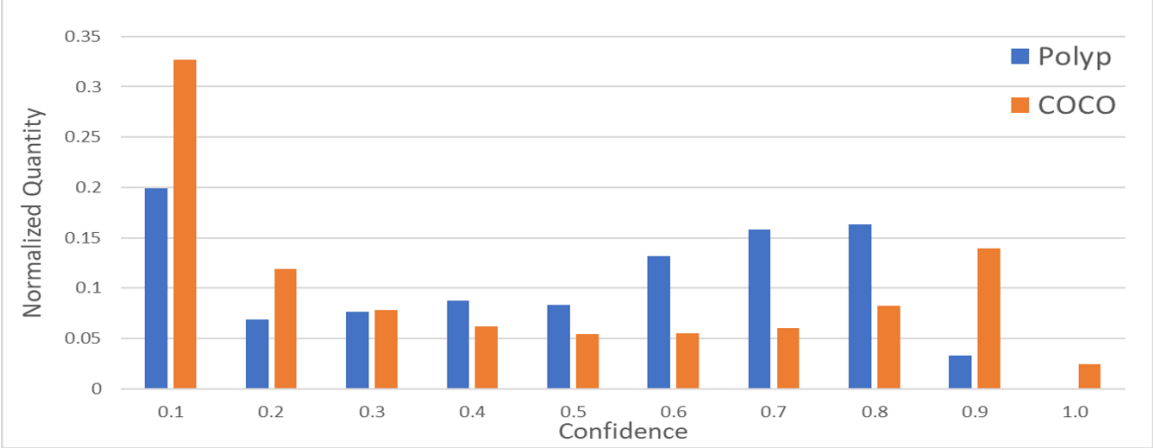

Figure 5: The bar chart presents the distribution of confidence values for all predicted boxes from the COCO and polyp dataset. The confidence levels of the predicted boxes are rounded, and the quantities are normalized.

Appendix D. Classification of detection vs. segmentation models

Table 5 compares the intrinsic subtype classification capability of single-path detection models and segmentation models, independent of their detection quality. For the detection baselines (YOLOv11/12/13, large and x-large), we first lower the confidence threshold to 0.001 and match all predicted boxes to ground-truth boxes using the same IoU criterion as in the main experiments. The classification accuracy is then computed only on the true-positive detection boxes that match, i.e., assuming the lesion has already been correctly localized. In contrast, for the segmentation models, we bypass detection entirely. Each ground-truth box is mapped onto the segmentation probability map, and a majority vote over pixel-wise predictions inside the box is used to assign a hyperplastic or adenoma label. As shown in Table 5, even under this oracle-detection setting, large and x-large YOLO detectors achieve average subtype accuracies of only 50–59%, whereas segmentation models deliver substantially higher values (e.g., UNeXt reaches 78.29% average accuracy, and EMCAD-B0 75.21%). This gap confirms that, given the same region of interest, segmentation networks provide much stronger per-polyp subtype classification than detector heads, motivating our dual-path design where detection is used purely for localization and segmentation is responsible for classification.

Table 5: Classification comparison between detection and segmentation models.

Model	**Size**	**$\mathbf{Acc_{hp}}$**	**$\mathbf{Acc_{ad}}$**	**$\mathbf{Acc_{avg}}$**
v11	l	48.92	51.09	50.01
	x	53.56	64.48	59.02
v12	l	49.23	55.96	52.59
	x	51.39	51.82	51.61
v13	l	50.46	55.47	52.97
	x	52.01	52.55	52.28
UNeXt	–	69.97	86.62	**78.29**
EMCAD_B0	–	65.02	85.40	75.21

Appendix E. Dataset statistics

Table 6: Image- and box-level statistics of the training and test splits.

Split	**#Images**	**BG-only**	**>1 box**	**Both**	**#Boxes**	**#HP**	**#AD**
Train	6442	509	374	53	6511	2803	3708
Test	704	25	45	8	734	323	411

For completeness, we also report summary statistics of box areas (as a percentage of image area). In the training split, the mean box area is 6.33% with a standard deviation of 8.71%; the median (p50) is 3.29%, with p75 = 7.42%, p90 = 14.89%, and p95 = 22.62%. In

Table 7: Box area distribution (as % of image area) in the training and test splits.

Box area range	**Train**	**Test**
[0.0, 0.1)%	0 (0.00%)	0 (0.00%)
[0.1, 0.5)%	304 (4.67%)	35 (4.77%)
[0.5, 1.0)%	740 (11.37%)	79 (10.76%)
[1.0, 2.0)%	1120 (17.20%)	147 (20.03%)
[2.0, 5.0)%	1972 (30.29%)	213 (29.02%)
[5.0, 10.0)%	1211 (18.60%)	123 (16.76%)
[10.0, 25.0)%	900 (13.82%)	114 (15.53%)
[25.0, 100.0]%	264 (4.05%)	23 (3.13%)

the test split, the mean is 6.12% and the standard deviation is 7.78%, with p50 = 3.16%, p75 = 7.51%, p90 = 15.49%, and p95 = 20.89%. Both splits are therefore dominated by small to medium polyps (median area $\approx$3% of the image), which makes the detection and subtype classification tasks non-trivial.