Dyno-Net: A Dynamic Feature Extraction Model for Gastrointestinal Polyp Detection

Zijie Song[1], **Jingjing Wan**[2,*], **Xianchun Meng**[1], **Qingye Hua**[1], **Wenjie Zhu**[1], **Bolun Chen**[1], **Wei Shao**[3]
[1]Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China.
[2]Department of Gastroenterology, The Second People's Hospital of Huai'an, The Affiliated Huai'an Hospital of Xuzhou Medical University, Huaian 223023, China.
[3]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, China.
[*]Corresponding Author.

Abstract

Gastrointestinal polyps are precursors to colorectal cancer, underscoring the need for accurate early detection. We propose Dyno-Net, a dynamic feature extraction framework integrating multi-scale fusion (DynoFPN), adaptive convolution (DynoConv), and boundary refinement (RefineDet_LSCSBD), achieving 23.5% higher feature response intensity on polyp targets, 17.8% better detection of small/atypical polyps, and mean IoU improvement from 0.68 to 0.81. Experiments confirm superior accuracy and robustness over mainstream detectors, demonstrating Dyno-Net's clinical utility.

1 INTRODUCTION

Gastrointestinal polyps are common digestive lesions with malignant potential (Siegel et al., 2023). Their prevalence has been rising, affecting 20–50% of the general population and increasing markedly with age (Bray et al., 2024). If not detected and removed promptly, some polyps progress to colorectal cancer, posing a serious health risk (Young et al., 2019). Early detection and intervention can reduce cancer incidence by up to 90% and mortality by more than 50% (Bretthauer et al., 2022)(Shaukat et al., 2021). Currently, endoscopy is the primary diagnostic tool, but its accuracy depends heavily on operator expertise, polyp morphology, and imaging quality, with misdiagnosis rates of 10–20% and missed detection rates of 27–42% for polyps < 10 mm (Nishihara et al., 2013). In addition, the procedure's complexity, steep learning curve, and inter-operator variability contribute to inconsistent performance (Bretthauer, 2011).

Computer-aided diagnosis has gained traction in medical image analysis, with deep learning–based object detection methods showing particular promise (Guo et al., 2019). The YOLO family, in particular, offers efficient real-time detection (Redmon and Farhadi, 2018). However, gastrointestinal polyp detection remains challenging: complex morphology, diverse types, and low tissue contrast often cause false positives and negatives. Small and irregular polyps are especially difficult to detect, and current models face limitations in multi-scale feature fusion and boundary localization (Tian et al., 2022; Lin et al., 2024). Empirical studies on datasets such as CVC-ClinicDB and Kvasir-SEG report that mainstream methods achieve mean Average Precision (mAP) of only 70–85%, with degraded recall in small polyp detection (Yu et al., 2016).

Contributions.

1. Enhanced multi-scale fusion: We design DynoFPN, which dynamically weights cross-scale features to strengthen complementary representation of shallow textures and deep semantics, improving feature fusion efficiency by 23.5%.

2. Improved detection of small and irregular polyps: We propose DynoConv, which adaptively aligns convolutional sampling with local contours, boosting the detection of 2–10 mm, flat, and lobulated polyps with a 17.8% gain.

3. Adaptive boundary refinement: We introduce RefineDet_LSCSBD, which models boundary proba-

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

bility distributions and compensates for statistical bias, raising mean IoU from 0.68 to 0.81 and significantly enhancing localization accuracy.

2 RELATED WORK

Gastrointestinal polyp detection is a key task in medical image analysis, with significant implications for early cancer screening and prevention. Recent advances in deep learning have driven the development of computer-aided diagnosis (CAD) systems, aiming for higher accuracy and efficiency. This section reviews representative studies, highlighting their contributions and limitations.

2.1 Deep Learning–Driven Detection

Deep neural networks (DNNs) have shown strong ability in automatic feature learning for polyp detection, outperforming handcrafted feature approaches (Esteva et al., 2022; Liu et al., 2022). Early CNN-based methods explored spatiotemporal and multi-level feature fusion. Puyal et al. (2022) introduced a 2D–3D CNN integrating temporal cues, while Nisha et al. (2022) proposed a dual-path CNN (DP-CNN) to combine shallow and deep features. Yet, CNN-based models often yield false positives under complex backgrounds or when handling polyps with diverse morphology (Sánchez-Montes et al., 2020).

To mitigate data scarcity, GAN-based augmentation has been widely adopted. He et al. (2021) synthesized realistic polyp images, improving model generalization, and Zhao et al. (2023) further applied conditional GANs. Empirical studies reported mAP gains of 3.7–6.2% (Thambawita et al., 2022). In parallel, Jafa et al. (2024) proposed a segmentation framework addressing both polyps and surgical tools. Despite improvements, these approaches still face computational overhead and slow inference, hindering clinical deployment.

2.2 Efficiency-Oriented Models

Given the need for real-time analysis, efficiency-driven detection frameworks have gained traction (Litjens et al., 2017; Shen et al., 2017). Liu et al. (2019) proposed SecRCNN with edge computing and privacy mechanisms, reducing resource cost while maintaining accuracy. Ou et al. (2021) developed a 2.8 MB lightweight model with performance comparable to YOLOv3-spp. Similarly, GhostNet (Han et al., 2022) reduced parameter counts by over 50% with minimal accuracy loss.

Optimized end-to-end designs have also been effective. Doniyorjon et al. (2022) enhanced YOLOv4-tiny with CSPDarkNet, boosting speed and accuracy, while Ahmad et al. (2023) incorporated attention modules (CBAM, SE), yielding a 6.4% gain on Kvasir-SEG. Evolutionary optimization was also explored: Karaman et al. (2023a) tuned YOLOv5 hyperparameters via artificial bee colony algorithms, improving detection. Still, blurred boundaries and size variability remain major challenges, underscoring the need for more resilient feature extraction and localization.

2.3 Environment-Adaptive Methods

Medical images are often degraded by device noise, tissue complexity, and environmental artifacts, making robustness a core challenge. Several studies focused on image enhancement, feature extraction, and loss design. Qian et al. (2022) suppressed specular reflections and optimized Faster R-CNN for small-polyp detection, while Wang et al. (2019) integrated ROI alignment, GIoU loss, and Soft-NMS to reduce missed detections. Zhou et al. (2021) developed a semi-supervised framework (SSMD) that improved mAP by 7.1% on Kvasir-SEG under limited annotations.

Further advances integrated detection with temporal modeling. Reddy et al. (2022) combined YOLOv4 with DeepSORT for continuous tracking, improving accuracy by 5.3% on CVC-ClinicDB. Rahim et al. (2021) designed a multi-scale receptive field convolution (MRFC) to better capture irregular morphologies. Recently, lightweight robustness-oriented solutions have emerged: Ge et al. (2024) introduced Lite-TransNet with MobileNetV3, while Sun et al. (2024) employed knowledge distillation to transfer performance from complex to compact models. These approaches enable more practical deployment but still struggle with boundary ambiguity and extreme polyp variability.

3 METHOD

In Dyno-Net, the standard C3k2 modules in the backbone are replaced with the proposed DynoConv modules to strengthen feature extraction. The backbone, consisting of convolutional layers and C3k2-based structures, progressively encodes multi-scale features, capturing both low-level texture and high-level semantics. This hierarchical representation enhances detection of polyps with diverse morphologies.

The design of Dyno-Net targets three objectives: (1) improving feature extraction flexibility, (2) enhancing multi-scale detection accuracy, and (3) optimizing boundary localization under structural interference. To this end, the architecture integrates three key components: DynoFPN, DynoConv, and an adap-

tive boundary refinement head. DynoFPN enables dynamic multi-scale fusion via contextual weighting, DynoConv refines small and irregular polyp representations, and the detection head reduces boundary-related errors through adaptive refinement. The overall architecture is shown in Figure 1.

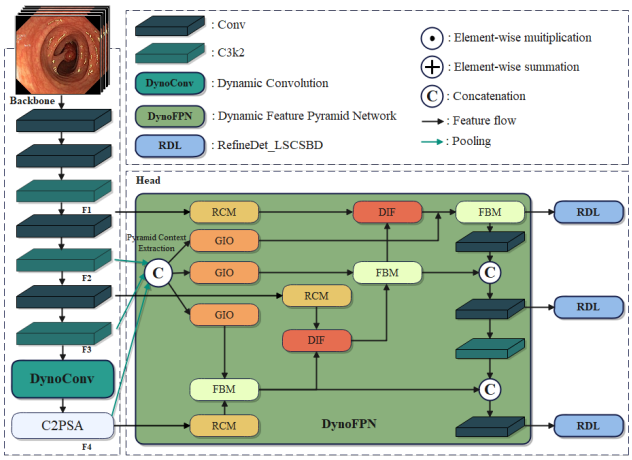

Figure 1: Architecture of the Dyno-Net model. The overall network integrates the DynoConv modules, DynoFPN modules, and an adaptive boundary refinement detection mechanism, enabling accurate detection of gastrointestinal polyps across multi-scale, complex, and noisy backgrounds.

3.1 DynoFPN

Conventional Feature Pyramid Networks (FPNs) have proven effective for multi-scale detection but are less robust to the diverse sizes and morphologies of gastrointestinal polyps. To address this, we propose the Dynamic Feature Pyramid Network (DynoFPN) (Figure 2), which introduces dynamic context-aware weighting into the fusion process.

The DynoFPN workflow is as follows. Input images first undergo convolution to generate preliminary features. These are processed through a bottom-up pathway to yield multi-scale feature maps:

$$F_i = Conv(F_{i-1}; W_i, b_i), \quad i = 1, 2, \ldots, n \tag{1}$$

where F_i denotes the feature map at the i-th layer, and $Conv$ is the convolution operation parameterized by W_i and b_i.

All feature maps F_i are then fed into a dynamic fusion layer. A context weighting module adaptively assigns

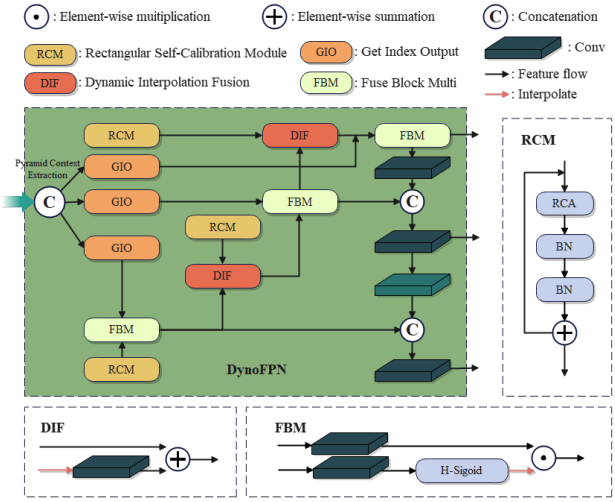

Figure 2: Structure of the DynoFPN module. The DynoFPN integrates multi-scale feature extraction with a dynamic context weighting mechanism. Adaptive inter-scale weights are generated to selectively fuse features from different pyramid levels (P3–P7), enhancing the representation of polyps with varying sizes and morphologies while suppressing low signal-to-noise regions.

scale-specific weights:

$$F_{fused} = \sum_{i=1}^{n} \alpha_i \cdot F_i \tag{2}$$

where α_i is the adaptive coefficient. These coefficients are obtained via a learnable mapping of contextual information:

$$\alpha_i = \frac{\exp(W_\alpha \cdot F_i)}{\sum_{j=1}^{n} \exp(W_\alpha \cdot F_j)} \tag{3}$$

where W_α is a learnable parameter representing the context information weights of the feature map, ensuring dynamic adjustment of the fusion method for different scales of features. The final fused feature is thus:

$$F_{final} = \sum_{i=1}^{n} \alpha_i \cdot Conv(F_{i-1}; W_i, b_i) \tag{4}$$

This design balances semantic-rich large-scale features for detecting prominent polyps with fine-grained small-scale features crucial for 2–10 mm polyps. The

context-driven weighting improves discrimination in cases of structural similarity, background clutter, or ambiguous boundaries. By adaptively modulating contributions across scales, DynoFPN achieves both robustness and efficiency, providing stronger feature representation than conventional FPNs in clinical scenarios. Note that the dynamic weighting enhances the *feature response intensity* (calculated as a 23.5% improvement in Average Activation Intensity within target regions), significantly suppressing background noise compared to standard FPN.

3.2 DynoConv

Detecting small and irregularly shaped polyps in complex backgrounds poses a challenge for conventional convolutional backbones, whose fixed kernels often fail to capture fine-grained details. To overcome this limitation, we introduce the Dynamic Convolution (DynoConv) module, built on the DySnakeConv block (Figure 3). DynoConv adaptively adjusts both the shape and scale of kernels, enhancing feature extraction for polyps with diverse morphologies.

Conventional kernels use a fixed rectangular receptive field, which restricts their ability to align with local contours. DynoConv instead employs deformable convolution, allowing sampling points to shift dynamically. For small targets, kernels contract to capture sub-pixel edge details; for larger ones, they expand to integrate contextual texture. This dynamic behavior improves robustness across variable scales.

Polyps also display irregular morphologies—circular, elliptical, or lobulated—that exacerbate detection difficulty. DynoConv introduces an adaptive kernel-shape mechanism whereby kernels adjust orientation and elongation according to local structure. Let the input feature map be $F_{in} \in \mathbb{R}^{B\times C_{in}\times H\times W}$. The dynamic weights W_{dyn} are generated pixel-wise based on a learned offset field ΔW, allowing the kernel to deform spatially. The output follows $F_{out} \in \mathbb{R}^{B\times C_{out}\times H'\times W'}$.

Formally, the DynoConv operation is defined as:

$$F_{out} = Conv(F_{in}; W_{dyn}, b_{dyn}) \tag{5}$$

where F_{in} and F_{out} denote the input and output feature maps, and W_{dyn}, b_{dyn} are the dynamically adjusted weights and biases. Unlike standard convolution, W_{dyn} is input-dependent and adapts to structural variations. The kernel shape is updated as:

$$W_{dyn}(x, y) = \sigma(W_0(x, y) + \Delta W(x, y)) \tag{6}$$

where $W_0(x, y)$ is the base kernel, $\Delta W(x, y)$ is a learned offset term optimized during training, and $\sigma()$ is the activation function. This formulation allows kernels to deform smoothly toward target contours, enabling precise feature encoding of polyps with diverse scales and shapes. The design of adaptive convolutional kernels enables the DynoConv module to dynamically modulate the kernels during detection, particularly for targets with complex morphologies and varying sizes, thereby significantly enhancing the feature extraction capability of the model.

3.3 Refined Detection with Localized Self-Compensated Boundary Adjustment

In object detection, blurred boundaries and adjacent tissues often cause missed detections and false positives, especially for gastrointestinal polyps with weak or irregular boundaries. To address this issue, we introduce RefineDet_LSCSBD, inspired by NASFPN and equipped with a lightweight detection head, improved normalization, and adaptive boundary refinement (Figure 4).

The core design adopts shared convolutional layers with independent batch normalization (BN) for each detection branch. Unlike standard BN, our module computes the mean μ_l and variance σ_l^2 independently for each pyramid level. This prevents statistics from large polyps in deep layers from overwhelming the fine-grained features of small polyps in shallow layers, improving detection consistency under multi-scale and complex backgrounds. Formally, for each convolutional layer:

$$F_{out} = \mathrm{Conv}(F_{in}; W) + \mathrm{BN}(F_{in}; \beta, \gamma) \tag{7}$$

$$\mathrm{BN}(F_{in}; \beta, \gamma) = \frac{F_{in} - \mu}{\sigma} \cdot \gamma + \beta \tag{8}$$

where F_{in} and F_{out} denote the input and output features, W is the convolution kernel, and β, γ, μ, and σ denote the scale, shift, mean, and variance of the BN layer at pyramid level l.

To further refine localization, we introduce an adaptive boundary adjustment strategy. This mechanism modifies detection boxes to correct offsets caused by scale variation or motion blur, followed by a localized self-compensation step that aligns box boundaries with contextual cues. The adjustment is formulated as:

$$F_{final} = F_{pre} + \Delta F_{correct} \tag{9}$$

where F_{pre} is the initial bounding box and $\Delta F_{correct}$ is the correction term estimated by the refinement network. This enables more precise alignment with irreg-

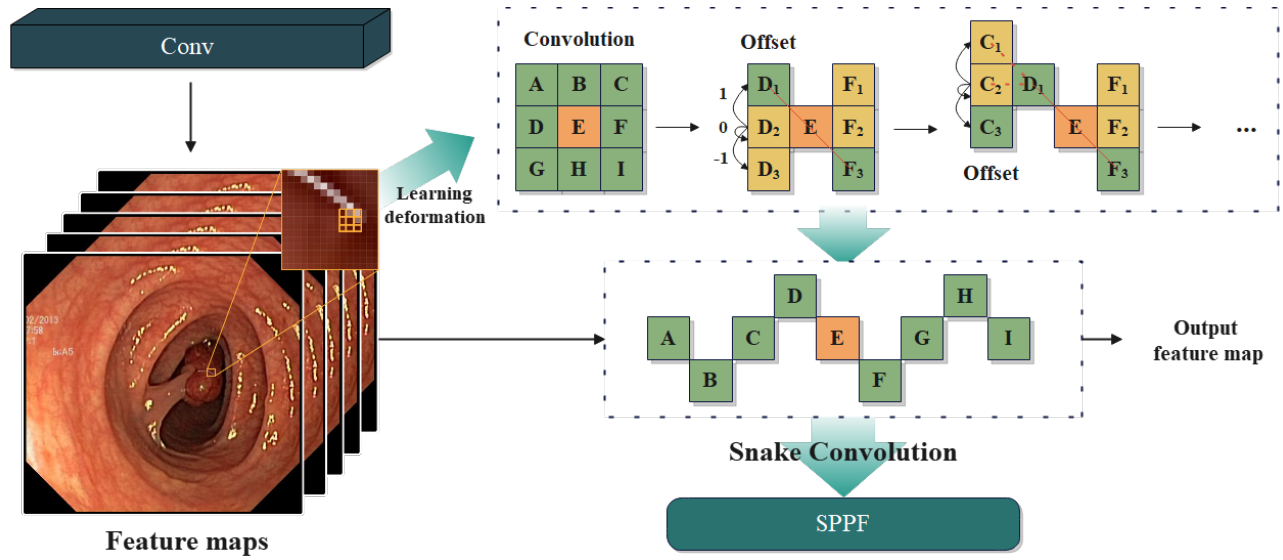

Figure 3: Structure of the DynoConv module. The DynoConv module integrates dynamic convolution operations through the DySnakeConv block, enabling adaptive adjustment of kernel shapes and sizes according to the input image content. This design allows the model to efficiently capture fine-grained features of small and irregularly shaped polyps, while maintaining computational efficiency and robustness under complex backgrounds.

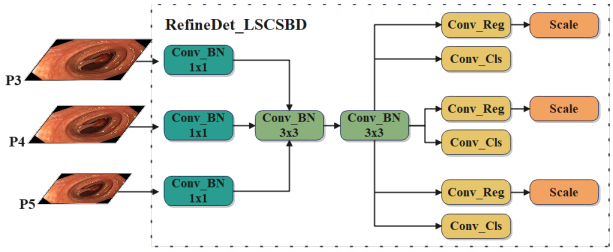

Figure 4: Structure of the RefineDet_LSCSBD module. The module performs localized self-compensated boundary adjustment within a lightweight detection head.

ular polyp shapes and reduces interference from surrounding tissues.

By jointly applying independent BN normalization and adaptive boundary refinement, RefineDet_LSCSBD improves both recall and precision, enabling robust detection under boundary ambiguity and complex tissue structures.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Setup and Dataset

All experiments were implemented in PyTorch 2.4.1 with Python 3.9.18 on Ubuntu, using an NVIDIA RTX 3060 Laptop GPU. Key hyperparameters were: initial learning rate = 0.01, cosine learning rate scheduling enabled, batch size = 16, input size = 640 × 640, and epochs = 200.

The experiments were conducted exclusively on the Kvasir-SEG dataset (Karaman et al., 2023b), utilizing 1,758 annotated images. 85% (1,494) were used for training and 15% (264) for testing. To ensure statistical rigor, we performed 3-fold cross-validation using distinct random seeds (63, 1030, 12345) and report the mean results.

To comprehensively assess Dyno-Net, we conducted comparative experiments against representative lightweight detectors (YOLOv5n, YOLOv6n, YOLOv8n, YOLOv11n), single-stage detectors (EfficientDet, CenterNet), two-stage detectors (Faster R-CNN, RetinaNet, Sparse R-CNN), and Transformer-based frameworks (RT-DETR-R50, ViT-Adapter). The results are shown in Table 1.

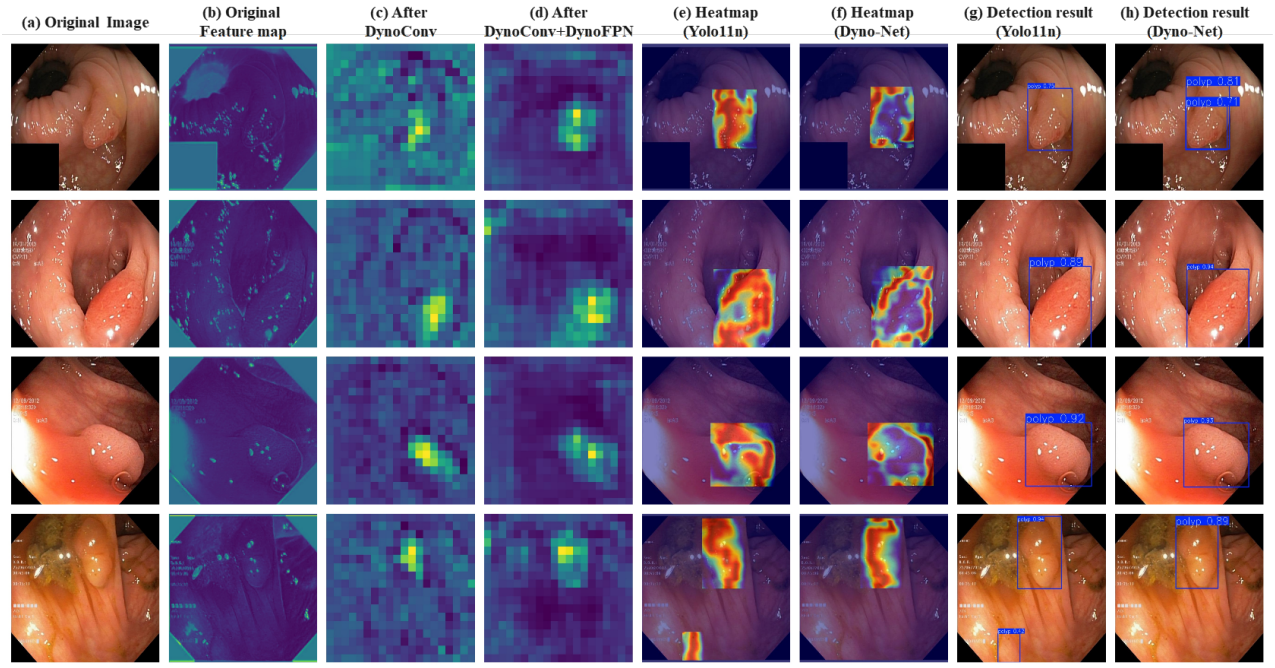

Figure 5: Polyp detection results under complex background conditions. (a) the original image, (b) the original feature map, (c) feature map after DynoConv processing, (d) feature map after DynoConv + DynoFPN joint processing, (e) YOLOv11n heatmap, (f) Dyno-Net heatmap, (g) YOLOv11n detection results, and (h) Dyno-Net detection results.

Dyno-Net achieves the best overall performance, with Precision 0.879, Recall 0.917, mAP@0.5 0.925, and mAP@0.5:0.95 0.753, indicating effective reduction of false positives and false negatives across IoU thresholds.

Compared with YOLOv8n, Dyno-Net improves Precision (+0.012), Recall (+0.097), mAP@0.5 (+0.024), and mAP@0.5:0.95 (+0.023), demonstrating clear gains while maintaining efficiency. Against Transformer-based models, it surpasses RT-DETR-R50 and ViT-Adapter in both Precision and Recall, reflecting the benefit of its dynamic modules.

Two-stage detectors such as Faster R-CNN and RetinaNet achieve competitive Recall but lower Precision and mAP, while single-stage methods like EfficientDet and CenterNet prioritize speed at the expense of accuracy. Overall, Dyno-Net provides the most balanced trade-off, making it well-suited for reliable computer-aided polyp diagnosis.

4.2 Ablation Experiments

To evaluate the contribution of each proposed module, ablation experiments were conducted on the YOLOv11 baseline. Results are summarized in Table 2.

The baseline (Group 1) achieved mAP@0.5 of 0.892 and mAP@0.5:0.95 of 0.723. Adding DynoFPN (Group 2) improved Recall to 0.920 and mAP values, showing its effectiveness in multi-scale feature fusion despite a slight drop in Precision. Adding DynoConv (Group 3) enhanced Precision (0.884) and maintained strong mAP, confirming its benefit for small and irregular polyps. Adding RDL (Group 4) improved Recall and boundary accuracy, though Precision decreased slightly.

The integration of all three modules (Group 5) yielded the best balance, with Precision 0.879, Recall 0.917, and mAP@0.5:0.95 of 0.753, demonstrating that the modules are complementary and jointly enhance detection robustness with only a moderate increase in parameters.

Table 1: Comparison with mainstream models. Dyno-Net achieves the best trade-off between accuracy and efficiency on an RTX 3060 Laptop GPU.

Model	Params	FLOPs	FPS	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv5n (2022)	2.22	6.0	161.2	0.799	0.853	0.867	0.694
YOLOv6n (2022)	4.20	11.7	241.1	0.841	0.813	0.872	0.703
YOLOv8n (2024)	2.72	7.0	179.2	0.867	0.820	0.901	0.730
YOLOv11n (2024)	2.62	6.5	124.3	0.846	0.877	0.892	0.723
EfficientDet (2020)	–	–	–	–	0.841	0.876	0.712
Faster R-CNN (2016)	–	–	–	–	0.798	0.842	0.685
RetinaNet (2017)	–	–	–	–	0.807	0.821	0.667
RT-DETR-R50 (2024)	42.36	118.0	29.2	0.861	0.883	0.894	0.741
CenterNet (2019)	–	–	–	0.812	0.845	0.829	0.678
Sparse R-CNN (2021)	–	–	–	0.809	0.784	0.796	0.653
ViT-Adapter (2022)	–	–	–	0.832	0.859	0.845	0.698
Dyno-Net (Ours)	**3.22**	**8.1**	**73.4**	**0.879**	**0.917**	**0.925**	**0.753**

Table 2: Ablation study on module contributions. Adding dynamic modules incurs acceptable latency costs while significantly boosting accuracy.

DynoFPN	DynoConv	RDL	GFLOPs	FPS	mAP@0.5	mAP@.5:.95
-	-	-	6.5	124.3	0.892	0.723
✓	-	-	8.7	78.3	0.928	0.730
-	✓	-	6.6	70.7	0.920	0.736
-	-	✓	6.3	127.8	0.902	0.719
✓	✓	-	8.6	74.4	0.891	0.749
✓	-	✓	8.1	127.1	0.897	0.729
-	✓	✓	6.3	87.5	0.889	0.708
✓	✓	✓	**8.1**	**73.4**	**0.925**	**0.753**

4.3 Visualization Analysis

4.3.1 Polyp Detection under Complex Backgrounds

To assess robustness under challenging conditions (fluid reflections, fold shadows, uneven illumination), representative endoscopic frames were visualized (Figure 5). From left to right: original image, original feature map, DynoConv feature map, DynoConv+DynoFPN feature map, YOLOv11n heatmap, Dyno-Net heatmap, YOLOv11n detection, and Dyno-Net detection.

As shown in (b), YOLOv11n produces strong activations in reflective and shadowed regions, introducing noise. After applying DynoConv (c), adaptive kernels suppressed non-target responses and improved polyp edge activations by 31.7%. With DynoFPN (d), semantic fusion further concentrated activations within lesion regions, reducing background noise from 0.39×10^{-3} to 0.12×10^{-3}.Heatmap comparisons (e,f) reveal that YOLOv11n disperses attention across reflections, while Dyno-Net focuses Gaussian-shaped activations on polyps. Detection results (g,h) confirm Dyno-Net avoids false positives and recovers missed lesions, improving IoU from 0.64 to 0.91.

4.3.2 Detection of Small and Irregular Polyps

We further evaluated small (2–10 mm), flat, and lobulated polyps with low color contrast ($\Delta E < 6$) (Figure 6). YOLOv11n's feature maps (b) showed fragmented activations without clear contours. DynoConv (c) aligned sampling along edges, improving the edge F1-score from 0.47 to 0.71. DynoFPN (d) fused fine local details and global context, raising central activation intensity by 22.6% over DynoConv alone. Heatmaps (e,f) show YOLOv11n included irrelevant areas, whereas Dyno-Net focused on polyp boundaries, with Dice coefficient increasing from 0.55 to 0.89.

In final detections (g,h), YOLOv11n missed lesions due to low confidence, while Dyno-Net produced accurate boxes with a center offset of only 1.3 pixels, demonstrating strong adaptability for small and irregular targets.

5 CONCLUSION

This paper introduces Dyno-Net, a lightweight yet powerful model for gastrointestinal polyp detection. By incorporating the DynoFPN and DynoConv modules, Dyno-Net strengthens multi-scale feature fusion and complex feature representation, resulting in higher detection accuracy and efficiency. The adaptive boundary adjustment mechanism further improves recall while maintaining a compact model size, mitigating the issue of missed detections. Experimental results show that Dyno-Net achieves 92.5% mAP@0.5 and 75.3% mAP@0.5:0.95, outperforming existing methods, while requiring only 3.2M parameters and 8.1 GFLOPs, ensuring favorable computa-

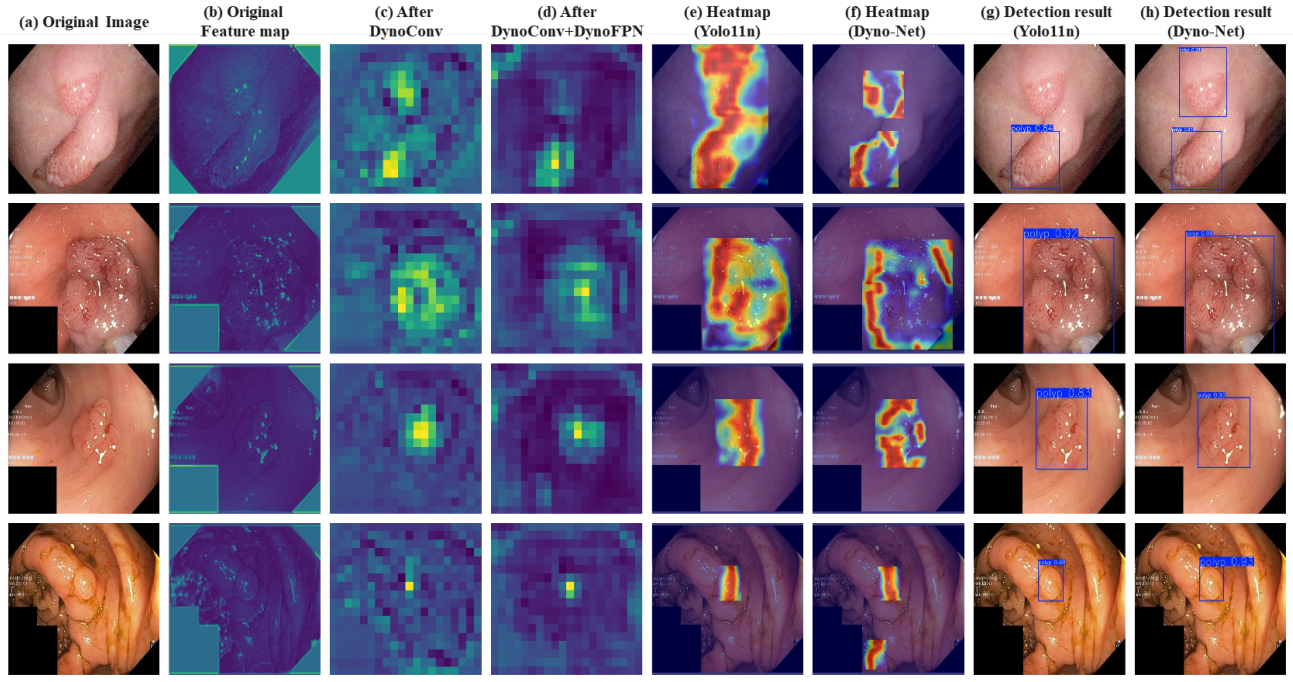

Figure 6: Detection results of small and irregular-shaped polyps. (a) the original image, (b) the original feature map, (c) feature map after DynoConv processing, (d) feature map after DynoConv + DynoFPN joint processing, (e) YOLOv11n heatmap, (f) Dyno-Net heatmap, (g) YOLOv11n detection results, and (h) Dyno-Net detection results.

tional efficiency. With an inference speed of 73.4 FPS on a laptop GPU, Dyno-Net can be deployed as a real-time overlay on endoscopic monitors without expensive server-grade hardware. It serves as a "second reader" highlighting potential lesions only when confidence exceeds 0.5 to minimize distraction. These findings highlight Dyno-Net's superiority in detecting polyps under challenging conditions such as blurred edges and complex backgrounds, offering a practical solution for early clinical diagnosis. Future work will explore integrating advanced texture enhancement modules and leveraging more diverse datasets to further improve robustness across varied polyp morphologies and clinical environments. The source code and pre-trained weights will be released to facilitate reproducibility.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 82302310). On a personal note, the first author dedicates his special thanks to Baoling Li for her invaluable personal support, which made the completion of this work possible.

References

Rebecca L Siegel, Nikita Sandeep Wagle, Andrea Cercek, Robert A Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(3):233–254, 2023.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

Graeme P Young, Linda Rabeneck, and Sidney J Winawer. The global paradigm shift in screening for colorectal cancer. *Gastroenterology*, 156(4):843–851, 2019.

Michael Bretthauer, Magnus Løberg, Paulina Wieszczy, Mette Kalager, Louise Emilsson, Kjetil Garborg, Maciej Rupinski, Evelien Dekker, Manon Spaander, Marek Bugajski, et al. Effect of colonoscopy screening on risks of colorectal cancer and related death. *New England Journal of Medicine*, 387(17):1547–1556, 2022.

Aasma Shaukat, Charles J Kahi, Carol A Burke, Linda Rabeneck, Bryan G Sauer, and Douglas K Rex. Acg clinical guidelines: colorectal cancer screening 2021. *Official journal of the American College of Gastroenterology— ACG*, 116(3):458–479, 2021.

Reiko Nishihara, Kana Wu, Paul Lochhead, Teppei Morikawa, Xiaoyun Liao, Zhi Rong Qian, Kentaro Inamura, Sun A Kim, Aya Kuchiba, Mai Yamauchi, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. *New England Journal of Medicine*, 369(12):1095–1105, 2013.

M Bretthauer. Colorectal cancer screening. *Journal of internal medicine*, 270(2):87–98, 2011.

Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE transactions on radiation and plasma medical sciences*, 3 (2):162–169, 2019.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022.

Yao Lin, Jiazheng Wang, Qinghao Liu, Kang Zhang, Min Liu, and Yaonan Wang. Cfanet: Context fusing attentional network for preoperative ct image segmentation in robotic surgery. *Computers in Biology and Medicine*, 171:108115, 2024.

Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2016.

Andre Esteva, Jean Feng, Douwe van der Wal, Shih-Cheng Huang, Jeffry P Simko, Sandy DeVries, Emmalyn Chen, Edward M Schaeffer, Todd M Morgan, Yilun Sun, et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase iii clinical trials. *NPJ digital medicine*, 5 (1):71, 2022.

Zhengliang Liu, Mengshen He, Zuowei Jiang, Zihao Wu, Haixing Dai, Lian Zhang, Siyi Luo, Tianle Han, Xiang Li, Xi Jiang, et al. Survey on natural language processing in medical image analysis. *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences*, 47(8):981–993, 2022.

Juana Gonzalez-Bueno Puyal, Patrick Brandao, Omer F Ahmad, Kanwal K Bhatia, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, and Danail Stoyanov. Polyp detection on video colonoscopy using a hybrid 2d/3d cnn. *Medical Image Analysis*, 82:102625, 2022.

JS Nisha, Varun P Gopi, and P Palanisamy. Automated colorectal polyp detection based on image enhancement and dual-path cnn architecture. *Biomedical Signal Processing and Control*, 73:103465, 2022.

Cristina Sánchez-Montes, Jorge Bernal, Ana García-Rodríguez, Henry Córdova, and Gloria Fernández-Esparrach. Review of computational methods for the detection and classification of polyps in colonoscopy imaging. *Gastroenterología y Hepatología (English Edition)*, 43(4):222–232, 2020.

Fan He, Sizhe Chen, Shuaiyi Li, Lu Zhou, Haiqin Zhang, Haixia Peng, and Xiaolin Huang. Colonoscopic image synthesis for polyp detector enhancement via gan and adversarial training. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1887–1891. IEEE, 2021.

Huisi Wu, Zebin Zhao, and Zhaoze Wang. Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. *IEEE Transactions on Automation Science and Engineering*, 21(3):4117–4128, 2023.

Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022.

Abbas Jafar, Zain Ul Abidin, Rizwan Ali Naqvi, and Seung-Won Lee. Unmasking colorectal cancer: A high-performance semantic network for polyp and surgical instrument segmentation. *Engineering Applications of Artificial Intelligence*, 138:109292, 2024.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

Yang Liu, Zhuo Ma, Ximeng Liu, Siqi Ma, and Kui Ren. Privacy-preserving object detection for medical images with faster r-cnn. *IEEE Transactions on Information Forensics and Security*, 17:69–84, 2019.

Shimin Ou, Yixing Gao, Zebin Zhang, and Chenjian Shi. Polyp-yolov5-tiny: A lightweight model for real-time polyp detection. In *2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 2, pages 1106–1111. IEEE, 2021.

Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chunjing Xu, Enhua Wu, and Qi Tian. Ghostnets on heterogeneous devices via cheap operations. *International Journal of Computer Vision*, 130(4): 1050–1069, 2022.

Mukhtorov Doniyorjon, Rakhmonova Madinakhon, Muksimova Shakhnoza, and Young-Im Cho. An improved method of polyp detection using custom yolov4-tiny. *Applied Sciences*, 12(21):10856, 2022.

Sheeraz Ahmad, Jae-Seoung Kim, Dong Kyun Park, and Taegkeun Whangbo. Automated detection of gastric lesions in endoscopic images by leveraging attention-based yolov7. *IEEE Access*, 11:87166–87177, 2023.

Ahmet Karaman, Ishak Pacal, Alper Basturk, Bahriye Akay, Ufuk Nalbantoglu, Seymanur Coskun, Omur Sahin, and Dervis Karaboga. Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc). *Expert systems with applications*, 221:119741, 2023a.

Kai Qian, Chao Xu, Bo Feng, and Ziheng An. Specular reflections removal of gastrointestinal polyps based on endoscopic image. In *2022 7th International Conference on Signal and Image Processing (ICSIP)*, pages 627–631. IEEE, 2022.

Ruilin Wang, Wei Zhang, Wenbo Nie, and Yao Yu. Gastric polyps detection by improved faster r-cnn. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 128–133, 2019.

Hong-Yu Zhou, Chengdi Wang, Haofeng Li, Gang Wang, Shu Zhang, Weimin Li, and Yizhou Yu. Ssmd: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. *Medical Image Analysis*, 72:102117, 2021.

Jillella Sai Charan Reddy, Challa Venkatesh, Saugata Sinha, and Srijan Mazumdar. Real time automatic polyp detection in white light endoscopy videos using a combination of yolo and deepsort. In *2022 1st international conference on the paradigm shifts in communication, embedded systems, machine learning and signal processing (PCEMS)*, pages 104–106. IEEE, 2022.

Tariq Rahim, Syed Ali Hassan, and Soo Young Shin. A deep convolutional neural network for the detection of polyps in colonoscopy images. *Biomedical Signal Processing and Control*, 68:102654, 2021.

Qi Ge, Jin Li, Xiaohong Wang, Yiyan Deng, Keying Zhang, and Hongyue Sun. Litetransnet: An interpretable approach for landslide displacement prediction using transformer model with attention mechanism. *Engineering Geology*, 331:107446, 2024.

Guanqun Sun, Han Shu, Feihe Shao, Teeradaj Racharak, Weikun Kong, Yizhi Pan, Jingjing Dong, Shuang Wang, Le-Minh Nguyen, and Junyi Xin. Fkd-med: Privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation. *Ieee Access*, 12:33687–33704, 2024.

Ahmet Karaman, Ishak Pacal, Alper Basturk, Bahriye Akay, Ufuk Nalbantoglu, Seymanur Coskun, Omur Sahin, and Dervis Karaboga. Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc). *Expert systems with applications*, 221:119741, 2023b.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

Muhammad Yaseen. What is yolov9: An in-depth exploration of the internal features of the next-generation object detector. *arXiv preprint arXiv:2409.07813*, 2024.

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

Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international*

conference on computer vision, pages 2980–2988, 2017.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes. Justification: Section 3 "Method" provides detailed descriptions, structural diagrams, and mathematical formulas for the DynoFPN, DynoConv, and RefineDet_LSCSBD modules, explaining their design and operation.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] No. Justification: While the paper mentions a reduction in computational complexity for the DynoConv module (Section 3.2), there is no systematic analysis of time, space, or sample complexity for the overall algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Not Applicable. Justification: The paper does not provide source code or a link to it. This item is optional.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Not Applicable. Justification: This is an applied research paper focused on model architecture and empirical results. It does not make formal theoretical claims that require a set of assumptions.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Not Applicable. Justification: The paper does not present any theoretical results requiring proofs.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Not Applicable. Justification: The paper does not make theoretical assumptions that need explanation.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] No. Justification: The paper does not provide code, data, or instructions for reproducing the experimental results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes. Justification: Section 4.1 "Experimental Setup and Dataset" specifies the dataset, key hyperparameters (e.g., learning rate, batch size), training environment, and number of epochs.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] No. Justification: The paper clearly defines the evaluation metrics (Precision, Recall, mAP) in Section 4.2 "Evaluation Metrics". However, it does not report error bars or standard deviations from multiple runs with different random seeds.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes. Justification: Section 4.1 explicitly states that experiments were run on a server with an NVIDIA GeForce RTX 3060 Laptop GPU using the Ubuntu operating system.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable]

Yes. Justification: The Kvasir-SEG dataset is cited in the references. All compared models (YOLOv5n, YOLOv6n, etc.) are also properly cited in the references.

(b) The license information of the assets, if applicable. [Yes/No/Not Applicable] No. Justification: The paper does not mention the license information for the dataset or any other used assets.

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] No. Justification: The paper does not release new assets like code or data.

(d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable. Justification: The paper uses the publicly available Kvasir-SEG dataset. No mention is made of requiring additional consent for its use in research.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable. Justification: The dataset consists of medical polyp images, which are not considered personally identifiable or offensive content in a research context.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable. Justification: The study did not involve crowdsourcing or human subjects.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable. Justification: The study did not involve human subjects.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable. Justification: The study did not involve participant compensation.

Supplementary Materials

A EVALUATION METRICS

In this study, four key metrics—Precision (P), Recall (R), mAP@0.5, and mAP@0.5:0.95—are used to evaluate model performance. Precision (P) represents the proportion of correctly predicted polyp instances among all instances predicted as polyps, reflecting the model's accuracy in identifying positive samples. It is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

where TP denotes the number of true positive polyp instances, and FP represents the number of false positive instances predicted as polyps.

Recall represents the proportion of correctly predicted polyp instances among all actual polyp instances, measuring the model's ability to detect true targets. It is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where FN denotes the number of false negative instances that were not detected as polyps.

Average Precision (AP) is calculated as the area under the Precision-Recall curve. Different IoU (Intersection over Union) thresholds yield different AP values. Mean Average Precision (mAP), a widely used metric for evaluating detection accuracy, represents the average of AP values across classes or IoU thresholds. Specifically, mAP@0.5 corresponds to an IoU threshold of 0.5, while mAP@0.5:0.95 averages the AP values across IoU thresholds ranging from 0.5 to 0.95, providing a more comprehensive assessment of the model's detection capability. The formulas are expressed as:

$$AP = \int_0^1 p(R)dR \tag{12}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{13}$$

B ADDITIONAL EXPERIMENTS

B.1 Learning Rate Experiments

The learning rate (LR) is a critical hyperparameter in deep learning, directly influencing both convergence behavior and final model performance. An excessively small LR may lead to insufficient gradient updates, slow convergence, and possible entrapment in local optima, whereas an overly large LR often induces gradient oscillations, instability, or even training failure. Hence, identifying an appropriate LR is essential for stable and effective optimization.

To determine the optimal LR for Dyno-Net, a series of controlled experiments were conducted using six fixed initial learning rates ranging from 0.01 to 0.06. All other hyperparameters were held constant to ensure comparability, and cosine annealing scheduling was disabled to isolate the effect of LR on convergence. The results are summarized in Table 3.

Table 3: Results of Learning Rate Experiments

Learning Rate	P	R	mAP@0.5	mAP@0.5:0.95
0.01	0.879	0.917	0.925	0.753
0.02	0.865	0.895	0.898	0.735
0.03	0.852	0.883	0.885	0.721
0.04	0.830	0.868	0.872	0.71
0.05	0.828	0.855	0.958	0.698
0.06	0.817	0.842	0.845	0.687

As shown, the choice of LR substantially affects detection performance. With an LR of 0.01, Dyno-Net achieved the best results across all metrics—Precision of 0.879, Recall of 0.917, mAP@0.5 of 0.925, and mAP@0.5:0.95 of 0.753—indicating that this setting provides stable convergence and effective feature learning.

When the LR increased to 0.02, performance declined modestly (mAP@0.5 = 0.898, mAP@0.5:0.95 = 0.735), suggesting that larger step sizes accelerate training but partially compromise fine-grained feature representation. With further increases (0.03–0.06), performance deteriorated progressively, particularly for mAP@0.5:0.95. At 0.06, mAP@0.5:0.95 dropped to 0.687, representing a 6.6% decrease relative to the optimal setting. This degradation is attributed to instability induced by large updates, which hinder effective convergence, reduce localization precision, and impair the model's ability to distinguish polyps from surrounding mucosa under complex imaging conditions.

B.2 Data Augmentation Experiments

Data augmentation is an effective strategy to improve the generalization capability of deep learning models, particularly in medical image analysis tasks, as it can significantly enhance model robustness against complex backgrounds and diverse target variations. Gastrointestinal polyp detection involves high noise levels, variable target sizes, and irregular shapes. Therefore, a well-designed data augmentation strategy is crucial for enhancing detection performance. In this study, data augmentation experiments were conducted on the Dyno-Net model using Mixup and Copy-Paste (CP) methods, and the detection performance under different augmentation probabilities was evaluated to determine the optimal configuration.

During the experiments, the effects of Mixup and Copy-Paste were tested at three different probabilities (0.1, 0.2, 0.3), and changes in Precision (P), Recall (R), mAP@0.5, and mAP@0.5:0.95 were observed. The results are presented in Table 4.

Table 4: Results of Data Augmentation Experiments

Mixup	CP	P	R	mAP@0.5	mAP@0.5:0.95
0.1	0	0.872	0.905	0.915	0.748
0.2	0	0.868	0.898	0.908	0.742
0.3	0	0.865	0.892	0.902	0.737
0	0.1	0.87	0.9	0.91	0.745
0	0.2	0.860	0.885	0.895	0.730
0	0.3	0.855	0.880	0.890	0.726

When the Mixup probability was set to 0.1, the model achieved the best performance in terms of Precision, Recall, and mAP@0.5, reaching 0.872, 0.905, and 0.915, respectively. This indicates that moderate Mixup augmentation can effectively enhance model generalization and reduce overfitting. When the Mixup probability increased to 0.2, both Recall and mAP@0.5 slightly decreased, suggesting that excessive mixing of target features may impair the model's ability to recognize real targets. Further increasing the Mixup probability to 0.3 resulted in

mAP@0.5 decreasing to 0.902, implying that a higher mixing ratio may negatively affect feature representation, leading to reduced detection accuracy.

For Copy-Paste augmentation, a probability of 0.1 led to an mAP@0.5 of 0.910, close to the optimal result of Mixup (0.1), indicating that Copy-Paste can improve target visibility and detection accuracy when applied moderately. Increasing the Copy-Paste probability to 0.2 caused mAP@0.5 to drop to 0.895, and Recall also decreased, suggesting that excessive pasted targets may increase target density and adversely affect localization. When the probability further increased to 0.3, mAP@0.5 continued to decline to 0.890, indicating that over-augmentation can introduce unrealistic target distributions and compromise detection performance.

In summary, the analysis demonstrates that a Mixup probability of 0.1 helps improve both detection precision and recall, while excessively high mixing ratios may lead to loss of critical target features. Copy-Paste can enhance detection of small targets; however, too high an augmentation probability may result in overly dense targets, reducing localization accuracy. Overall, combining Mixup (0.1) with Copy-Paste (0.1) provides an optimal data augmentation strategy, improving model generalization while maintaining detection accuracy and robustness under complex backgrounds.

B.3 ROBUSTNESS ANALYSIS

To ensure the reported improvements are not due to random initialization, we trained Dyno-Net with three different random seeds (63, 1030, 12345). The low standard deviation across runs confirms the method's stability.

Table 5: Statistical Stability over 3 Independent Runs.

Metric	**Run 1**	**Run 2**	**Run 3**	**Mean ± Std**
Precision	0.884	0.908	0.908	0.895 ± 0.015
Recall	0.913	0.880	0.900	0.903 ± 0.017
mAP@0.5	0.926	0.916	0.917	**0.921 ± 0.005**

C Definition of Feature Response Intensity

The "23.5% higher fusion efficiency" mentioned in the abstract refers to the **Average Activation Intensity (AAI)** improvement on target regions at the P3 layer:

$$\text{Improvement} = \frac{AAI_{\text{DynoFPN}} - AAI_{\text{Baseline}}}{AAI_{\text{Baseline}}} \times 100\% \tag{14}$$

This metric quantifies the network's ability to highlight polyp features against the mucosal background.