YOLO11s-RDS: Response-Guided Feature Enhancement with Spatial Priors for Lightweight Colorectal Polyp Detection

Haoran Wu[1][0009-0006-6949-504X]

[1] School of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
haoran_wu@mail.tust.edu.cn

Abstract. Accurate and efficient colorectal polyp detection is essential for computer-aided colorectal cancer screening. However, endoscopic images often suffer from specular reflections, intestinal folds, motion blur, weak lesion boundaries, and small-scale abnormalities, which introduce severe background interference and increase the risk of missed detections and localization errors. To address these challenges, we propose YOLO11s-RDS, a lightweight colorectal polyp detector built upon YOLO11s. The proposed method redesigns the neck with three complementary modules. First, Response-Guided Calibration Fusion (RGCF) replaces direct feature concatenation with response-aware gated fusion, suppressing shallow pseudo-responses caused by reflections and folds while improving cross-scale semantic consistency. Second, Dual-Pooling Feature Enhancement (DPFE) jointly exploits global average responses and local peak activations to strengthen channel representations associated with small polyps. Third, Spatial Prior-Weighted Convolution (SPWConv) introduces a local center-enhancement prior into downsampling convolutions to mitigate spatial-structure degradation of compact targets. Experiments on a uniformly processed hybrid dataset comprising CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB, and Kvasir-SEG show that YOLO11s-RDS achieves Precision, Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ of 0.9018, 0.8729, 0.8871, 0.9113, and 0.6715, respectively. Compared with YOLO11s, it improves Recall, mAP_{50}, and $mAP_{50:95}$ by 5.51%, 3.77%, and 3.26%, demonstrating stronger robustness in complex endoscopic scenes while maintaining a lightweight design.

Keywords: Colorectal polyp detection, Multi-scale feature fusion, Small-object detection, Spatial prior.

1 Introduction

Colorectal cancer (CRC) is one of the most prevalent gastrointestinal malignancies worldwide and often develops through the adenoma--carcinoma sequence. Early detection and removal of colorectal polyps at the precancerous stage are therefore critical for reducing CRC risk. Colonoscopy enables direct visualization of the intestinal mucosa

and intervention on suspicious lesions, making it a key tool for CRC screening and prevention[1-4].

However, practical colonoscopy screening remains challenging due to image degradation and subtle lesion appearance. Endoscopic images are frequently affected by uneven illumination, mucosal reflections, intestinal folds, occlusion, and motion blur, which obscure lesion boundaries and introduce complex background interference. Moreover, small or flat polyps often resemble the surrounding mucosa in color, texture, and morphology, making their visual cues weak and easily overlooked. Prolonged image inspection may further induce clinician fatigue, increasing the risk of missed or false detections of small lesions[5-7]. Therefore, accurate, robust, and efficient polyp detection in complex endoscopic scenes remains a critical problem for computer-aided CRC screening.

In recent years, deep learning-based computer-aided diagnosis techniques have provided an effective means of improving polyp detection. Existing studies have demonstrated that AI-assisted detection systems can enhance the detection of polyps and adenomas while reducing the risk of missed lesions in manual screening[8-10]. Object detection methods can generally be divided into two-stage and one-stage detectors. Two-stage methods, represented by Faster R-CNN, usually exhibit strong region proposal modeling and localization capabilities. Nevertheless, their proposal generation and multi-stage inference processes introduce considerable computational overhead, making them less suitable for computationally efficient colonoscopy-assisted detection scenarios[11,12]. In contrast, one-stage detectors, particularly the YOLO series, offer end-to-end prediction and relatively low computational cost, making them more appropriate for lightweight endoscopic detection scenarios[13-17]. Among them, YOLO11s achieves a favorable balance between model size and detection performance and can serve as an effective baseline for lightweight colorectal polyp detection[18].

Although YOLO11s provides a favorable balance between detection accuracy and model complexity, directly applying it to complex endoscopic images still presents several limitations. First, YOLO-style detection networks generally rely on multi-scale feature fusion to accommodate objects of different sizes[19-21]. However, shallow features often contain abundant redundant responses caused by specular reflections, intestinal folds, bubbles, and texture variations. Directly concatenating these shallow features with deeper semantic features may lead to cross-level semantic interference, thereby weakening the model's discriminative capability for polyp regions[22]. Second, small polyps occupy only a limited proportion of both the original image and feature maps. Their boundary, texture, and color differences are usually weak, and their local peak responses may gradually decay after multiple convolution and downsampling operations, increasing the risk of missed detection. Finally, while standard downsampling operations compress spatial resolution, they may also weaken the central responses and local topological structures of small targets, resulting in localization deviations under low-contrast or boundary-ambiguous conditions.

To address these issues, this paper proposes an improved lightweight colorectal polyp detection algorithm, termed YOLO11s-RDS. The proposed method is developed based on YOLO11s and introduces targeted architectural improvements to tackle insufficient cross-scale feature fusion, attenuation of weak small-lesion responses, and loss

of spatial information during downsampling in complex endoscopic scenarios. By optimizing the feature fusion strategy, enhancing small-target feature representation, and improving spatial information preservation during downsampling, the proposed model can extract more stable and discriminative polyp features, thereby better adapting to complex backgrounds, small-scale lesions, and weak-boundary targets.The main contributions of this paper are summarized as follows:

1. We propose YOLO11s-RDS, a lightweight colorectal polyp detector designed for complex endoscopic scenes with small lesions and weak boundaries.
2. We design three complementary modules: RGCF for calibrated cross-scale fusion, DPFE for weak-response enhancement, and SPWConv for spatial-structure preservation during downsampling.
3. Extensive experiments on four public polyp datasets demonstrate that the proposed method improves detection accuracy and recall over YOLO11s and other YOLO-based detectors with moderate computational overhead.

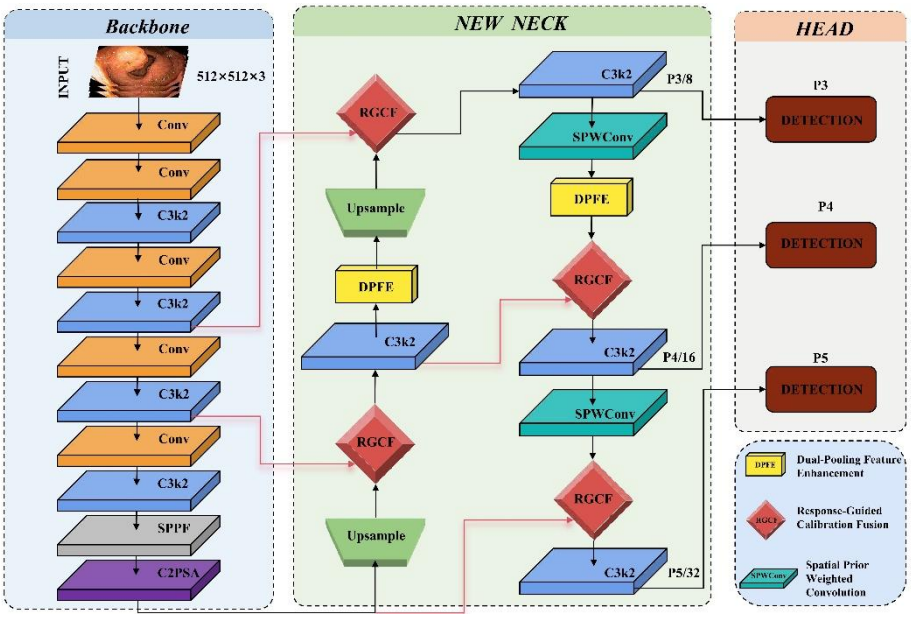

Fig. 1. Overall architecture of the proposed YOLO11s-RDS.

2 Methodology

2.1 Overview

Fig. 1 illustrates the overall architecture of YOLO11s-RDS. Built upon YOLO11s, the proposed model retains the original backbone and detection head while redesigning the neck to improve multi-scale feature interaction. The redesigned neck integrates RGCF, DPFE, and SPWConv for cross-scale fusion calibration, weak small-target response

enhancement, and spatial-structure preservation during downsampling, respectively. With these designs, YOLO11s-RDS improves the representation of small and weak-boundary polyps in complex endoscopic scenes while preserving the lightweight nature of YOLO11s.

2.2 Response-Guided Calibration Fusion Module

Conventional YOLO necks commonly adopt FPN/PAN-style multi-scale feature aggregation with channel concatenation. However, shallow features in endoscopic images may contain pseudo-responses from specular reflections, folds, bubbles, and blur, while deep features may weaken small-polyp responses after repeated downsampling. Direct concatenation therefore introduces unreliable responses into the fused representation. To address this problem, we propose the Response-Guided Calibration Fusion (RGCF) module, as shown in **Fig. 2**.

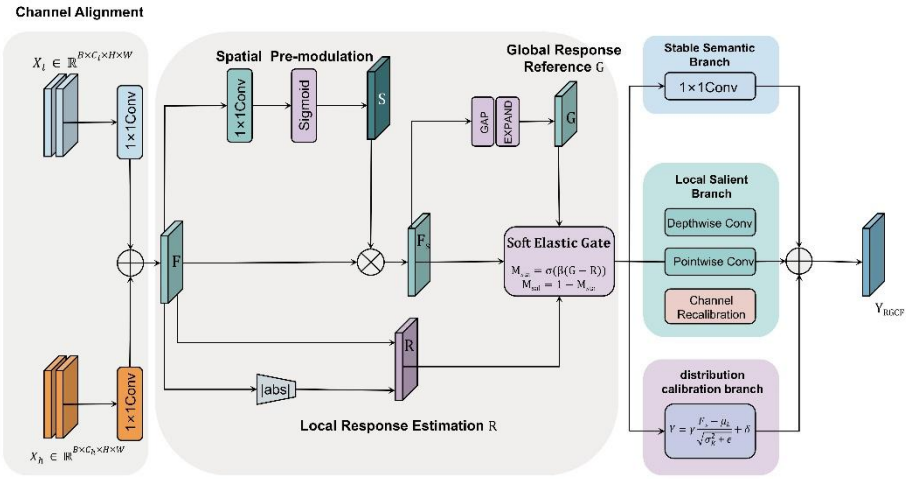

Fig. 2. Structure of the Response-Guided Calibration Fusion (RGCF) module.

Given two input features $X_l \in \mathbb{R}^{B\times C_l\times H\times W}$ and $X_h \in \mathbb{R}^{B\times C_h\times H\times W}$, where B denotes the batch size, C_l and C_h denote the input channel numbers, and H and W denote the spatial height and width, respectively. If their spatial resolutions are different, upsampling or downsampling is first applied to align them. Then, two independent 1×1 convolutions are used for channel alignment:

$$\hat{X}_l = f_l(X_l), \quad \hat{X}_h = f_h(X_h), \tag{1}$$

where $f_l(\cdot)$ and $f_h(\cdot)$ are channel mapping functions, and $\hat{X}_l, \hat{X}_h \in \mathbb{R}^{B\times C\times H\times W}$. The aligned features are fused by element-wise addition:

$$F = \hat{X}_l + \hat{X}_h. \tag{2}$$

where F denotes the initial fused feature. Compared with concatenation, additive fusion avoids channel expansion and forms a unified feature space.

Before response partitioning, RGCF performs spatial pre-modulation on F:

$$S = \sigma\big(\phi_s(F)\big), \tag{3}$$

$$F_s = S \odot F, \tag{4}$$

where $\phi_s(\cdot)$ is a 1×1 convolution for spatial mapping, $\sigma(\cdot)$ denotes the sigmoid function, S is the spatial modulation map, and $\odot$ denotes element-wise multiplication. The modulated feature F_s is used for subsequent response estimation.

RGCF then computes a global response reference and a local response estimate. The global response reference is obtained by:

$$G = E\big(GAP(F_s)\big), \tag{5}$$

where $GAP(\cdot)$ denotes global average pooling, and $E(\cdot)$ broadcasts the pooled response to the same spatial size as F_s. Thus, $G \in \mathbb{R}^{B\times C\times H\times W}$. The local response estimate is defined as:

$$R = \frac{1}{2}\big(|\hat{X}_l| + |\hat{X}_h|\big), \tag{6}$$

where $|\cdot|$ denotes the element-wise absolute value operation. R measures the local activation intensity of the aligned cross-scale features.

Different from generic attention or channel recalibration methods[23-26], RGCF generates soft gates according to the response difference between G and R:

$$M_{sta} = \sigma\big(\beta(G - R)\big), \quad M_{sal} = 1 - M_{sta}, \tag{7}$$

where M_{sta} denotes the stable-response gate, M_{sal} denotes the local-saliency gate, and β is a temperature coefficient controlling the smoothness of the gate distribution. In this work, β is set to 1. Since the gates are generated by the sigmoid function, the gating process is differentiable.

Based on the soft gates, F_s is divided into a stable-response component and a local-saliency component:

$$F_{sta} = M_{sta} \odot F_s, \quad F_{sal} = M_{sal} \odot F_s. \tag{8}$$

For the stable-response component, a lightweight 1×1 convolution is used for semantic integration:

$$Y_{sta} = \phi_{sta}(F_{sta}), \tag{9}$$

where $\phi_{sta}(\cdot)$ denotes the stable semantic branch, and Y_{sta} preserves globally consistent semantic information.

For the local-saliency component, RGCF adopts a local enhancement branch:

$$Y_{sal} = \mathcal{C}\left(\phi_{pw}\big(\phi_{dw}(F_{sal})\big)\right), \tag{10}$$

where $\phi_{dw}(\cdot)$ denotes depthwise convolution, $\phi_{pw}(\cdot)$ denotes pointwise convolution, and $\mathcal{C}(\cdot)$ denotes channel recalibration based on global pooling and channel normalization. The depthwise convolution captures local spatial details, the pointwise convolution enables channel interaction, and the channel recalibration emphasizes more discriminative local responses. It should be noted that locally salient responses are not necessarily equivalent to polyp regions. They may contain small-polyp boundaries and abrupt texture variations, but may also include background pseudo-responses caused by specular reflections or intestinal folds. Therefore, RGCF does not directly enhance all locally strong responses. Instead, it further filters candidate salient regions through local convolution and channel recalibration.

To reduce feature distribution shifts caused by spatial modulation and gated partitioning, RGCF introduces a lightweight distribution calibration branch. The channels of F_s are divided into K groups. For the k-th group, the mean and variance are computed as:

$$\mu_k = \frac{1}{|\Omega_k|}\sum_{i\in\Omega_k} F_{s,i}, \tag{11}$$

$$\sigma_k^2 = \frac{1}{|\Omega_k|}\sum_{i\in\Omega_k}\left(F_{s,i} - \mu_k\right)^2, \tag{12}$$

where Ω_k denotes the set of elements in the k-th channel group, and $|\Omega_k|$ is the number of elements in this group. The calibrated output is formulated as:

$$Y_{cal} = \gamma\frac{F_s - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + \delta, \tag{13}$$

where γ and δ are learnable affine parameters, and ϵ is a small constant for numerical stability. This branch stabilizes the feature distribution after gated reorganization.

Finally, the outputs of the stable semantic branch, local saliency branch, and distribution calibration branch are integrated as:

$$Y_{RGCF} = Y_{sta} + Y_{sal} + Y_{cal}. \tag{14}$$

2.3 Dual-Pooling Feature Enhancement Module

After cross-scale fusion by RGCF, weak responses of small or low-contrast polyps may still be insufficient, especially when sparse lesion cues are suppressed by background mucosal textures, folds, or specular highlights. To further compensate for such weak target-related responses with limited computational cost, we introduce the Dual-Pooling Feature Enhancement (DPFE) module after selected fused features, as shown in **Fig. 3**. Different from complex attention structures, DPFE serves as a lightweight channel-wise response compensation module by jointly exploiting global average and local peak statistics[23,24,26].

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, DPFE first generates two channel descriptors by global average pooling and global max pooling:

$$z_{avg} = GAP(X), \quad z_{max} = GMP(X), \tag{15}$$

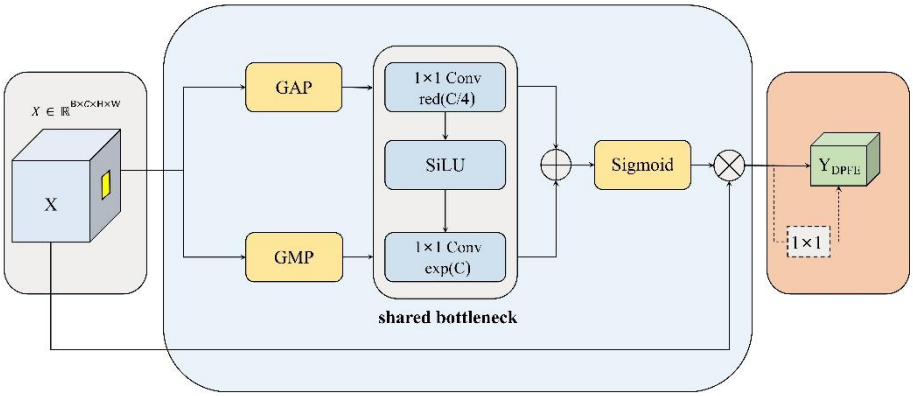

Fig. 3. Structure of the Dual-Pooling Feature Enhancement (DPFE) module.

where $GMP(\cdot)$ denotes global max pooling. Here, z_{avg} captures the overall channel response, while z_{max} emphasizes sparse local peak activations that may correspond to small-polyp cues.

The two descriptors are then processed by a shared bottleneck mapping:

$$f(z) = W_2\delta(W_1 z), \tag{16}$$

where W_1 and W_2 are 1×1 convolutions for channel reduction and expansion, respectively, and $\delta(\cdot)$ denotes the SiLU activation function. The bottleneck dimension is set to $C/4$ to reduce the parameter cost.

The channel enhancement weight is obtained by combining the two pooled responses:

$$A = \sigma\left(f(z_{avg}) + f(z_{max})\right), \tag{17}$$

where A denotes the channel-wise enhancement weight. This weight integrates global semantic responses and local peak responses, enabling the network to emphasize channels related to weak polyp features.

Finally, the input feature is reweighted as:

$$Y = X \odot A, \tag{18}$$

If the input and output channel dimensions are inconsistent, a 1×1 projection is used for channel matching:

$$Y_{DPFE} = P(Y), \tag{19}$$

where $P(\cdot)$ is an identity mapping when the channel dimensions are unchanged, and a 1×1 convolution otherwise.

2.4 Spatial Prior-Weighted Convolution Module

Stride-2 downsampling in feature pyramids reduces spatial resolution and may dilute compact responses of small polyps with surrounding mucosal textures, folds, or highlights[27,28]. To alleviate this problem, we propose the Spatial Prior-Weighted Convolution (SPWConv) module, which introduces a fixed local center-enhancement prior into downsampling convolutions. The prior is imposed within each local convolutional window rather than on the global image center, encouraging the convolution to preserve local center responses during scale transitions.

Let W denote the weight tensor of a standard convolution kernel. Specifically,$W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k}$,where C_{in}, C_{out}, and k denote the input channel number, output channel number, and kernel size, respectively. For a $k \times k$ convolution kernel, SPWConv first constructs a one-dimensional center-decay vector:

$$\alpha = [\rho, \cdots, 1, \cdots, \rho], \tag{20}$$

where the center position is assigned a weight of 1, and ρ denotes the decay coefficient from the center to the boundary. Then, a two-dimensional spatial prior matrix is generated by the outer product:

$$\Phi = \alpha^T \alpha \tag{21}$$

In this work, SPWConv adopts a 3×3 kernel with $\rho = 0.5$. Thus, $\alpha = [0.5,1.0,0.5]$, and the corresponding prior matrix is:

$$\Phi = \begin{bmatrix} 0.25 & 0.50 & 0.25 \\ 0.50 & 1.00 & 0.50 \\ 0.25 & 0.50 & 0.25 \end{bmatrix} \tag{22}$$

This prior assigns the largest weight to the local kernel center, intermediate weights to edge positions, and smaller weights to corner positions.

Based on this spatial prior, the convolution kernel is reweighted as:

$$\hat{W} = W \odot \Phi, \tag{23}$$

where Φ is broadcast along the input-channel and output-channel dimensions. The SPWConv operation is then formulated as:

$$Y = \text{Conv}\big(X, \hat{W}, b\big), \tag{24}$$

where X is the input feature, Y is the convolution output, and b is the bias term. Following the standard convolutional block in YOLO11s, Batch Normalization and SiLU activation are further applied:

$$Y_{\text{SPW}} = \text{SiLU}(\text{BN}(Y)). \tag{25}$$

Since Φ is fixed, SPWConv introduces no additional learnable parameters. During inference, the reweighted kernel $\hat{W} = W \odot \Phi$ can be pre-integrated into the convolution weights.

3 Experiments

3.1 Datasets and Preprocessing

We evaluate the proposed method on four public colorectal polyp datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB[29,32]. These datasets differ in image resolution, imaging conditions, polyp scale, and background complexity, enabling evaluation under diverse endoscopic scenarios. The dataset statistics and splits are summarized in **Table 1**.

Table 1. Details and splits of the experimental datasets.

Dataset	Year	Original resolution	Train	Val	Test
Kvasir-SEG	2020	Varied	800	100	100
CVC-ClinicDB	2015	384 × 288	428	61	123
CVC-ColonDB	2012	574 × 500	304	38	38
ETIS-LaribPolypDB	2014	1225 × 966	150	20	26
Total	---	---	1682	219	287

To reduce potential data leakage from highly similar samples within the same examination, we adopt group-wise splitting whenever video- or patient-level information is available, assigning samples from the same video sequence or patient source to the same subset. For datasets without explicit grouping metadata, a fixed random seed is used, and the split indices are kept identical across all comparative experiments. The training subsets of the four datasets are merged for model training, resulting in 1,682 training images, 219 validation images, and 287 test images.

Since these datasets mainly provide pixel-level segmentation masks, we convert them into detection annotations. For each binary polyp mask, connected-component analysis is performed, and the minimum enclosing rectangle of each valid component is used as the bounding box of the corresponding instance. Images with multiple disconnected polyp regions are assigned multiple bounding boxes. All boxes are then converted into YOLO format, including the class index, normalized center coordinates, and normalized width and height.

For fair comparison, YOLO11s-RDS and all compared models use the same input size and preprocessing pipeline. All images are resized to 512×512, with bounding-box coordinates updated accordingly. During validation and testing, only resizing and normalization are applied. During training, we use random horizontal flipping, random rotation, random scaling, brightness/contrast perturbation, and Gaussian blur, with a flipping probability of 0.5, a rotation range of $\pm 15^{\circ}$, a scaling range of 0.8-1.2, and a brightness/contrast perturbation range of ± 0.2. All geometric transformations are applied synchronously to images and bounding boxes to ensure label consistency.

3.2 Experimental Environment and Implementation Details

All experiments were conducted under the same hardware and software environment. The platform ran Windows 11 and was equipped with an NVIDIA RTX 4070 GPU with 12GB memory and an Intel® Core™ i7-13700K CPU. All models were implemented using Python 3.11.7, PyTorch 2.8.0, and CUDA 11.8.

For fair and reproducible comparison, YOLOv3-tiny, YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, and YOLO11s-RDS were trained and evaluated with identical data splits, input resolution, and training protocol. AdamW was used as the optimizer with an initial learning rate of 1×10^{-4}. The batch size was set to 16, and each model was trained for 300 epochs. After each epoch, validation performance was evaluated, and the best-performing weights were selected for final testing.

3.3 Evaluation Metrics

Model performance is evaluated using Precision, Recall, F1-score, mAP_{50}, $mAP_{50:95}$, GFLOPs, and the number of parameters. Precision, Recall, and F1-score measure detection accuracy, recall capability, and their overall balance, respectively. mAP_{50} denotes the average precision at an IoU threshold of 0.50, while $mAP_{50:95}$ averages the precision over IoU thresholds from 0.50 to 0.95 with a step size of 0.05. GFLOPs and parameter count are used to assess computational complexity and model size. Since the task contains only one category, i.e., colorectal polyps, mAP is equivalent to the AP of this class under the corresponding IoU setting.

4 Results and Analysis

4.1 Comparative Experiments

Table 2 reports the results of different YOLO-based detectors on the mixed test set. YOLO11s-RDS achieves the best Precision, F1-score, mAP_{50}, and $mAP_{50:95}$, reaching 0.9018, 0.8871, 0.9113, and 0.6715, respectively. Its Recall reaches 0.8729, only slightly lower than that of YOLOv9s, indicating a favorable balance between detection accuracy and sensitivity.

Table 2. Performance comparison with different YOLO-based detection models.

Method	Precision	Recall	F1	mAP_{50}	$mAP_{50:95}$	GFLOPs	Params
YOLOv3-tiny [14]	0.8808	0.7117	0.7872	0.7794	0.5553	18.9	12.13M
YOLOv8s [15]	0.8994	0.8528	0.8755	0.8843	0.6617	28.4	11.13M
YOLOv9s [16]	0.8239	0.8804	0.8512	0.9029	0.6589	26.7	7.17M
YOLOv10s [17]	0.8176	0.7577	0.7865	0.8304	0.6146	21.4	7.22M
YOLO11s [18]	0.8969	0.8273	0.8607	0.8782	0.6503	21.3	9.41M
YOLO11s-RDS	0.9018	0.8729	0.8871	0.9113	0.6715	23.1	10.87M

Compared with the YOLO11s baseline, YOLO11s-RDS improves Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ by 5.51%, 3.07%, 3.77%, and 3.26%, respectively, demonstrating improved polyp recall and localization quality in complex endoscopic scenes.

Meanwhile, GFLOPs and parameters increase by only 8.45% and 15.52%, suggesting that the performance gain does not rely on substantial model-scale expansion.

Compared with other YOLO variants, YOLO11s-RDS also shows a competitive accuracy-complexity trade-off. Relative to YOLOv8s, it improves mAP_{50} and $mAP_{50:95}$ by 3.05% and 1.48% with lower computational cost. Relative to YOLOv9s, although its Recall is slightly lower, YOLO11s-RDS achieves higher Precision, F1-score, and $mAP_{50:95}$, indicating better false-positive suppression while maintaining high sensitivity. Overall, YOLO11s-RDS is more suitable for lightweight colorectal polyp detection.

4.2 Ablation Study

To evaluate the contribution of RGCF, DPFE, and SPWConv, we conduct ablation experiments using YOLO11s as the baseline under the same experimental settings, as reported in **Table 3**. RGCF calibrates cross-scale feature fusion, DPFE enhances channel-wise response representation, and SPWConv introduces a fixed spatial prior during downsampling. Accordingly, we evaluate single-module variants, the RGCF+DPFE combination, and the complete YOLO11s-RDS model.

Table 3. Ablation study of the proposed modules.

Method	RGCF	DPFE	SPW-Conv	Precision	Recall	F1	mAP50	mAP50:95	GFLOPs	Params
YOLO11s	×	×	×	0.8969	0.8273	0.8607	0.8782	0.6503	21.3	9.41M
	×	×	✓	0.8912	0.8543	0.8724	0.9042	0.6730	21.3	9.41M
	×	✓	×	0.9197	0.7945	0.8525	0.8875	0.6342	21.4	9.45M
	✓	×	×	0.8912	0.8037	0.8452	0.8868	0.6402	24.4	11.10M
	✓	✓	×	0.9046	0.8144	0.8571	0.8951	0.6639	24.0	10.86M
	✓	✓	✓	0.9018	0.8729	0.8871	0.9113	0.6715	23.1	10.87M

Among the single-module variants, SPWConv provides the most stable performance gain. With only SPWConv, Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ increase by 3.26%, 1.36%, 2.96%, and 3.49%, respectively, while the number of parameters remains unchanged. When DPFE is used alone, Precision improves by 2.54% to 0.9197, but Recall and $mAP_{50:95}$ decrease by 3.96% and 2.48%, respectively, indicating that it does not yield consistent gains across all key metrics when used independently. RGCF alone improves mAP_{50} by 0.98%, but its overall gain remains limited.

Combining RGCF and DPFE improves Precision, mAP_{50}, and $mAP_{50:95}$ by 0.86%, 1.92%, and 2.09% over the baseline, respectively, showing more stable detection gains. After further incorporating SPWConv, the complete YOLO11s-RDS improves Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ by 5.51%, 3.07%, 3.77%, and 3.26%, respectively, achieving the best overall performance. Overall, the ablation results demonstrate the complementary contribution of the three modules and confirm that their joint integration improves the polyp detection performance of YOLO11s with moderate computational overhead.

4.3 Curve Analysis

To evaluate the stability of different models under varying threshold settings, we plot the Precision-Recall and F1-confidence curves in Fig.4. As shown in **Fig. 4**(a), YOLO11s-RDS achieves a generally higher Precision-Recall curve than most

compared models, with an AP of 0.911, which is consistent with the mAP_{50} results in **Table 2**. In particular, it maintains relatively high Precision in the medium-to-high Recall range, indicating that the model improves polyp recall while effectively controlling false positives.

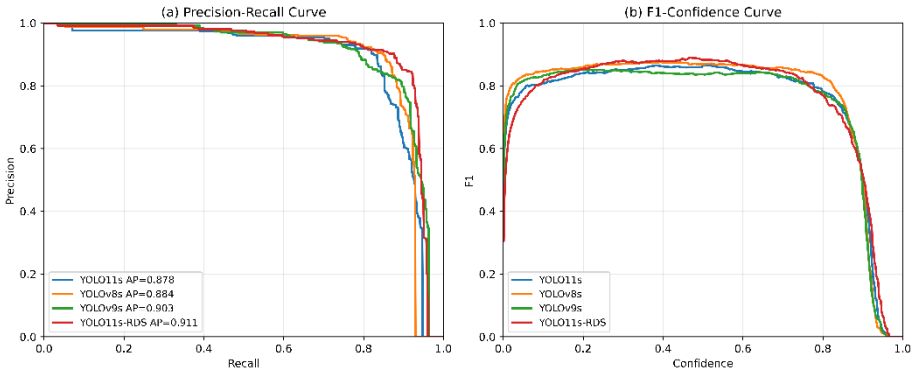

Fig. 4. Performance curves of different models. (a) Precision--Recall curve. (b) F1-confidence curve.

Fig. 4 (b) further shows the F1-score variation under different confidence thresholds. YOLO11s-RDS maintains a high F1-score over a relatively broad threshold range, suggesting that its balance between Precision and Recall is more stable against threshold changes. This property reduces the dependence on careful threshold tuning in practical deployment. Overall, the curve analysis confirms that YOLO11s-RDS provides favorable threshold robustness and detection stability beyond fixed metric evaluation.

4.4 Qualitative Analysis

To visually evaluate detection performance in complex endoscopic scenes, **Fig. 5** compares different models on representative samples involving common polyps, small-scale polyps, and polyps affected by blur or specular reflection. The first column shows the ground-truth annotations, while the remaining columns present the results of YOLO11s, YOLOv8s, YOLOv9s, and YOLO11s-RDS, respectively.

As observed in **Fig. 5**, all models can detect the target in the common polyp case, but YOLO11s-RDS produces bounding boxes that are more consistent with the annotations. For small-scale polyps with low boundary contrast, some compared models exhibit localization deviations, whereas YOLO11s-RDS maintains better box alignment. In samples affected by blur and specular reflection, YOLO11s-RDS also provides relatively stable localization, indicating stronger adaptability to complex background interference.

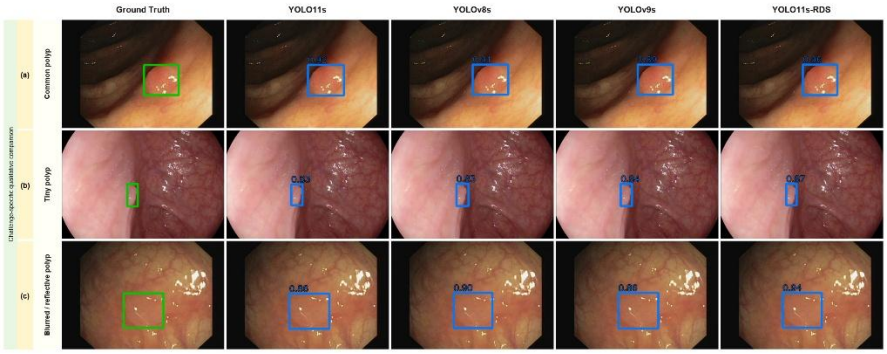

Fig. 5. Qualitative comparison of different models in representative endoscopic scenarios.

5 Conclusion

This paper presents YOLO11s-RDS, a lightweight colorectal polyp detector for complex endoscopic scenes. By integrating RGCF, DPFE, and SPWConv into the YOLO11s neck, the proposed method improves response-calibrated cross-scale fusion, weak small-target response enhancement, and spatial-structure preservation during downsampling, thereby improving the robustness of detecting small and weak-boundary polyps.

On the mixed dataset comprising Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB, YOLO11s-RDS achieves Precision, Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ of 0.9018, 0.8729, 0.8871, 0.9113, and 0.6715, respectively. Compared with YOLO11s, it improves Recall, F1-score, mAP_{50}, and $mAP_{50:95}$ by 5.51%, 3.07%, 3.77%, and 3.26%. Ablation results show that the three modules are complementary and improve detection performance with moderate computational overhead. Future work will further evaluate real-time video-level performance and generalization on multi-center clinical datasets.

References

1. Yin, Z., Yao, C., Zhang, L., Qi, S.: Application of artificial intelligence in diagnosis and treatment of colorectal cancer: A novel prospect. Frontiers in Medicine 10, 1128084 (2023)
2. Chen, Q., Xu, Y.-H., Kang, S., Lin, W., Luo, L., Yang, L., Zhang, Q.-H., Yang, P., Huang, J.-Q., Zhang, X., et al.: The tissue and circulating cell-free DNA-derived genetic landscape of premalignant colorectal lesions and its application for early diagnosis of colorectal cancer. MedComm 5(12), e70011 (2024)
3. Lieberman, D., Dekker, E.: How good is good enough? What should be the target for CRC screening? Digestive Diseases and Sciences 70(5), 1660–1667 (2025)
4. Moon, S.Y., Lee, J.Y., Lee, J.H.: Comparison of adenoma detection rate between high-definition colonoscopes with different fields of view: 170 degrees versus 140 degrees. Medicine 102(2), e32675 (2023)

5. Rodge, G.: Artificial intelligence for colonic polyp and adenoma detection: The way forward. Journal of Digestive Endoscopy 14(01), 064–066 (2023)
6. Wang, S.-Y., Gao, J.-C., Wu, S.-D.: Artificial intelligence for reducing missed detection of adenomas and polyps in colonoscopy: A systematic review and meta-analysis. World Journal of Gastroenterology 31(21), 105753 (2025)
7. Gan, P., Li, P., Xia, H., Zhou, X., Tang, X.: The application of artificial intelligence in improving colonoscopic adenoma detection rate: Where are we and where are we going. Gastroenterología y Hepatología 46(3), 203–213 (2023)
8. Thomas, J., Ravichandran, R., Nag, A., Gupta, L., Singh, M., Panjiyar, B.K.: Advancing colorectal cancer screening: A comprehensive systematic review of artificial intelligence (AI)-assisted versus routine colonoscopy. Cureus 15(9) (2023)
9. Rondonotti, E., Di Paolo, D., Rizzotto, E.R., Alvisi, C., Buscarini, E., Spadaccini, M., Tamanini, G., Paggi, S., Amato, A., Scardino, G., et al.: Efficacy of a computer-aided detection system in a fecal immunochemical test-based organized colorectal cancer screening program: A randomized controlled trial (AIFIT study). Endoscopy 54(12), 1171–1179 (2022)
10. Wang, P., Berzin, T.M., Glissen Brown, J.R., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al.: Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. Gut 68(10), 1813–1819 (2019)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems 28 (2015)
12. J. Li, J. Zhang, D. Chang, and Y. Hu, "Computer-assisted detection of colonic polyps using improved Faster R-CNN," Chinese Journal of Electronics, vol. 28, no. 4, pp. 718–724, 2019.
13. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
14. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
15. Lalinia, M., Sahafi, A.: Colorectal polyp detection in colonoscopy images using YOLO-V8 network. Signal, Image and Video Processing 18(3), 2047–2058 (2024)
16. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: YOLOv9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision, pp. 1–21. Springer (2024)
17. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: YOLOv10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems 37, 107984–108011 (2024)
18. Khanam, R., Hussain, M.: YOLOv11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
19. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
20. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
21. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)

22. Liu, S., Huang, D., Wang, Y.: Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516 (2019)
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
24. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
25. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)
26. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
27. Wei, W., Cheng, Y., He, J., Zhu, X.: A review of small object detection based on deep learning. Neural Computing and Applications 36(12), 6283–6303 (2024)
28. Zhang, R.: Making convolutional networks shift-invariant again. In: International Conference on Machine Learning, pp. 7324–7334. PMLR (2019)
29. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A segmented polyp dataset. In: International Conference on Multimedia Modeling, pp. 451–462. Springer (2019)
30. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Viñarino, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics 43, 99–111 (2015)
31. Bernal, J., Sánchez, J., Viñarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45(9), 3166–3182 (2012)
32. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293 (2014)