An Improved Remote Sensing Object Detection Algorithm Based on YOLOv11

ZiJian Lin

Abstract— **To address the multiple challenges in existing remote sensing images detection methods, including insufficient localization accuracy, imprecise category recognition, and high false positive and false negative rates, this paper proposes RS-YOLOv11 (Remote Sensing-YOLOv11), an improved object detection algorithm specifically designed for remote sensing applications based on the YOLOv11 framework.This study introduces the fine-grained SPD-Conv module to optimize backbone network downsampling, effectively preserving feature information and enhancing small object detection. The detection head employs Dynamic Head architecture with integrated multi-dimensional attention mechanisms, significantly improving model performance.To reduce network complexity, Faster_Block replaces the Bottleneck design, decreasing C3K2 module computational cost and addressing YOLOv11 deployment challenges. This improvement achieves lightweight design while maintaining performance and balancing Dynamic Head overhead. Additionally, WIoU loss function replaces CIoU to suppress gradient issues from low-quality images.Experiments on the VisDrone2021 dataset demonstrate that our improved model achieves a 3.9% increase in mAP50 compared to the YOLOv11n baseline, while maintaining comparable computational complexity and parameter efficiency.**

Index Terms—**Deep learning, remote sensing image, defect detection · yolov11**

I. Introduction

Remote sensing images play a crucial role in precision agriculture, geological disaster monitoring, and military defense[1]. However, the dense distribution of objects, scale variations, and complex environmental factors in these images[2] pose significant challenges to object detection, making the reduction of false positive and false negative rates a critical issue in this field[3].

Deep learning models for object detection are categorized into two types: two-stage models represented by SPPNet[4] and Fast R-CNN[5] offer high accuracy but slow speed, while single-stage models like YOLO[6-9] provide fast inference but limited small object detection capability. As a classic single-stage algorithm, YOLO is renowned for its real-time performance and efficiency, yet parameter complexity and computational cost remain key constraints for its application. Despite YOLOv11's superior performance, its complex structure hinders deployment on edge devices.

Based on YOLOv11n, this study proposes the RS-YOLO algorithm with the following key innovations:

(1)To achieve model lightweight, the Bottleneck structure of C3K2 is improved by adopting a more lightweight structure while maintaining detection accuracy with minimal degradation.

(2)A dynamic detection head (Dyhead) with multiple attention mechanisms is introduced, enabling the model to focus more on dense small target regions and extract enhanced small target features. Model performance is further improved by integrating three attention functions: scale, spatial, and task-oriented attention.

(3)The SPD module is incorporated into the feature extraction network, allowing the network to obtain feature maps without information loss during downsampling, expanding the receptive field and enhancing model detection performance. This approach better preserves feature map details and contextual relationships while reducing small target feature loss in low-light scenarios.

(4)To improve network model convergence speed, the boundary fitting loss function is replaced from CIoU to dynamic non-monotonic focusing Wise IoU, reducing the impact of annotation quality on loss convergence and suppressing background interference.

II. YOLOv11 Model

The architecture of YOLOv11 incorporates several innovative design elements. At the input level, enhanced Mosaic data augmentation implements random image manipulation techniques, strengthening the model's adaptability to diverse real-world scenarios and complex backgrounds.

The backbone network comprises four key modules: Conv, C3K2, C2PSA, and SPPF. The Conv module optimizes image resolution and channel dimensions for feature extraction. C3K2 integrates global semantic context with local target information, enhancing the detector's focus on critical regions. The C2PSA module implements cross-stage partial spatial attention, excelling at processing small and occluded objects through position-sensitive attention mechanisms. SPPF conducts multi-scale feature pooling, enabling flexible processing of varied input dimensions while expanding the receptive field.

In the neck network, multi-level feature map fusion enhances the model's capability to handle diverse scale scenarios. The detection head adopts a decoupled structure[10], separating target localization and classification into independent branches. This design, combined with YOLOX's Anchor-Free mechanism[11-15], optimizes small target edge prediction while reducing hyperparameters and computational complexity. The architecture employs depth-wise separable convolution to minimize computational

Manuscript received March 01, 2025
ZiJian Lin, School of Software, Tiangong university, Tianjin ,China.

redundancy, achieving enhanced operational efficiency while maintaining detection accuracy.The structure of the baseline model for YOLOv11 is depicted in Figure 1.

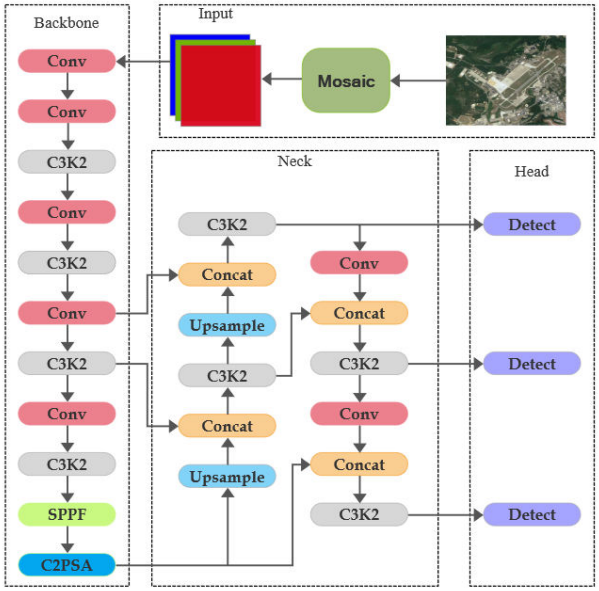

Fig. 1 YOLOv11 network structure

III. IMPROVEMENT METHODS

Building upon YOLOv11, we propose comprehensive architectural enhancements targeting feature fusion, context processing, and loss computation. Key improvements include PConv for C3K2 module optimization, SPD Conv for downsampling, Dyhead for detection, and WIOU loss for boundary regression. The enhanced architecture is depicted in Figure 2.

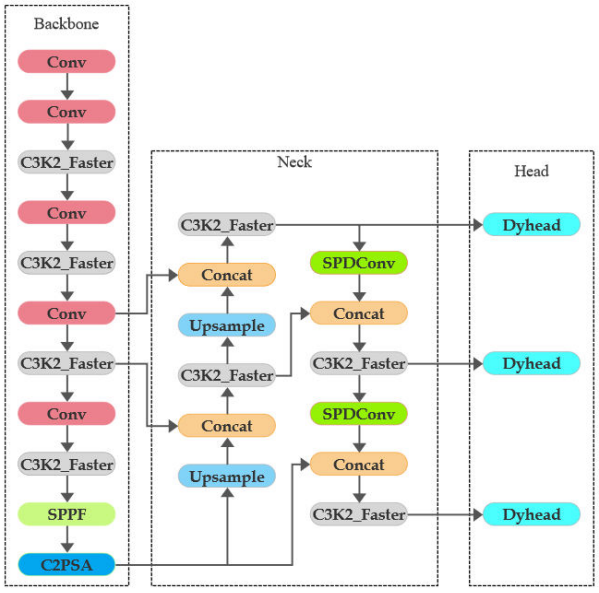

Fig. 2 Improved YOLOv11 model

A. C3K2_Faster

Despite YOLOv11's improved accuracy over its predecessors, its complex architecture and substantial parameter count, particularly in the C3K2 module's bottleneck structures, lead to redundant channel information during feature extraction. While lightweight networks like MobileNet[16], ShuffleNet[17], and GhostNet[18] employ deep or group convolution for spatial feature extraction, these approaches, though reducing FLOPs, often increase memory access costs and computational fragmentation, necessitating additional compensatory structures.

CHEN et al.[20] introduced PConv (Partial Convolution) in their Faster Neural Networks framework, offering an efficient solution to these challenges. PConv selectively convolves only specific input channels while preserving others, effectively reducing computational redundancy and memory access while maintaining network performance. The preserved channels undergo subsequent 1×1 convolutions, ensuring comprehensive feature utilization.

Drawing inspiration from the PConv concept, this paper innovatively designs the Faster Block structure to replace the original Bottleneck components in YOLOv11's backbone network, constructing a novel C3K2_Faster module. The innovation of Faster_Block lies in its selective convolution operation on only 1/4 of the input channels, coupled with lightweight 1×1 convolutions, reducing the computational overhead of each module to approximately 1/16 of the original Bottleneck. Figure 3 illustrates the C3K2_Faster module.

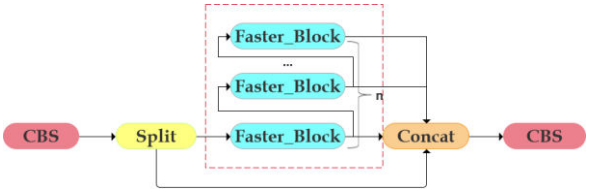

Fig. 3 C3K2_Faster Module

B. Dyhead

Despite YOLOv11's detection head's strong performance, its linear feature transmission design and simplified aggregation mechanism limit multi-scale information capture. While depth-wise separable convolutions reduce computational overhead, the current structure lacks adaptive feature fusion capabilities, necessitating architectural optimization for enhanced detection performance.

Inspired by [21], this study introduces Dynamic Head, a multi-scale detection architecture that unifies scale-aware, spatial-aware, and task-aware attention mechanisms through dimensional-specific feature tensor integration.

To address the computational complexity challenges associated with direct self-attention implementation, Dynamic Head employs an innovative sequential design. The attention mechanism is decomposed into three independent dimensions: scale, spatial, and task-oriented, processed sequentially to enable focused feature enhancement. The scale attention module learns semantic level importance for target feature enhancement, while spatial attention captures object transformation information, improving model adaptability to rotation and scaling. Task-oriented attention guides individual feature channels in executing specific detection tasks.Figure 4 illustrates the internal structure of three attention modules in Dyhead.

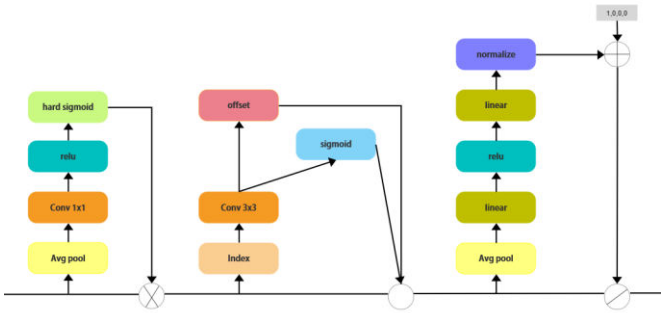

Fig. 4 Dyhead Internal Structure

C. SPD Conv

Current models employ strided convolution layers for feature map downsampling, which, while expanding the receptive field and reducing computational costs, inevitably results in fine-grained information loss. This information degradation is particularly evident when processing low-light remote sensing images: object contours, already blurred due to insufficient illumination, become increasingly indistinct through multiple convolution and pooling layers, with detailed features progressively weakening during layer-wise transmission, leading to inefficient feature learning.

To address the fine-grained feature loss in the original model's downsampling module, this paper proposes integrating the SPD-Conv[22] module into YOLOv11's neck network. Comprising spatial-depth layers and non-strided convolution layers, this module effectively preserves dimensional information during downsampling operations. This design not only enhances the network's feature fusion capabilities but also enables more granular feature learning. Compared to traditional downsampling methods, SPD-Conv better preserves small object features, significantly improving the model's detail capture capability.

SPD-Conv operates through a spatial-depth separation sampling strategy for lossless feature downsampling. Given a feature map T of size cwh, it samples one pixel from each row and column, generating four sub-feature maps of size c. These sub-feature maps are subsequently combined to form a 2x downsampled feature map of size 4c that retains all information. This design integrates width and height feature information into the channel dimension, expanding the channel count fourfold. Figure 5 illustrates the SPD-Conv layer with a scaling factor n of 2.

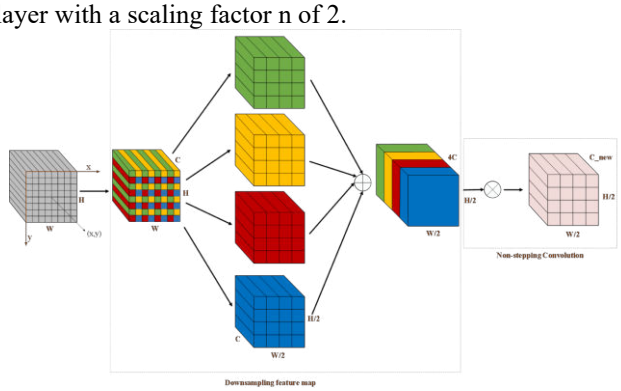

Fig. 5 SPD Conv Working Mechanism

D. WIOU

The conventional CIoU loss function implemented in YOLOv11, despite incorporating multiple geometric factors such as centroid distance, overlap ratio, and aspect ratio, exhibits limitations in handling training sample imbalances. The geometric metrics' inherent bias towards penalizing low-quality samples results in loss oscillation and compromised convergence efficiency. This study proposes adopting the WIoUv3 loss function to optimize annotation quality robustness and small aircraft detection performance. The mathematical formulation is expressed as:

$$L_{WIoUv3} = rL_{WIouv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \quad (1)$$

$$\beta = \frac{L_{IOU}}{\overline{L_{IOU}}} \in [0, +\infty) \quad (2)$$

Where α and δ are hyperparameters, $\overline{L_{IOU}}$ is a dynamic variable, and the criteria for dividing the quality of the anchor frames are also dynamic. This allows WIoUv3 to adopt the gradient gain assignment strategy that best matches the current situation at any given moment.Figure 6 illustrates the principle of the Wise-IoU v3 loss function. The blue section represents the actual box, while the green section represents the predicted box.

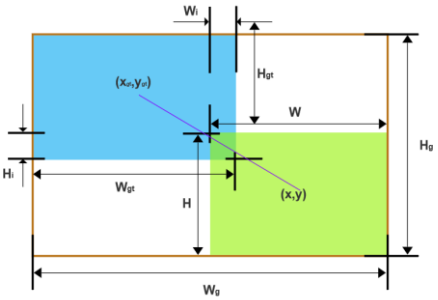

Fig 6 Schematic of Wise-IoU v3

IV. .EXPERIMENT

All experiments were performed on Ubuntu 22.04 LTS with an NVIDIA GeForce RTX 4090 (24GB VRAM), implementing PyTorch 2.0 and CUDA 11.8.

A. Dataset

The VisDrone2021 dataset,developed at Tianjin University, contains 8,629 drone-view images (6,471 training, 548 validation, 1,610 test) across ten object categories including various vehicles and pedestrian types

B. Experimental Environment Configuration

All models shared identical hyperparameters (Table 1) to ensure fair comparison. Critical parameters encompass input resolution, training epochs, and convergence controls (learning rate, momentum, weight decay). Mosaic augmentation (close_mosaic=10) was implemented to enhance training data diversity through controlled image fusion.

Table 1.
Model training hyperparameter settings

Hyperparameter Options	Setting
Input Resolution	640x640
Initial Learning Rate 0	0.01
(lr0)	0.01
Learning Rate Float	0.878
(lrf)	0.0005
Momentum	8
Weight_decay	300
Batch-size	10

C. Ablation experiments

The effectiveness of proposed improvements was evaluated through systematic ablation studies (Table 2). Performance enhancement was analyzed by progressively incorporating different modules, where "√" and "×" denote the presence and absence of methods respectively. D, P, F, and W represent Dyhead, SPDconv, C3K2_Faster, and WIoU loss function implementations.

Table 2
Results of ablation experiment.

Opt	D	P	F	W	mAP@0.5(%)	Params/M	FLOPs/G
1					33.0	2.62	6.6
2	√				36.1	3.14	7.8
3	√	√			36.5	3.25	8.1
4	√		√		32.6	2.39	5.4
7	√	√	√		36.7	3.0	7.9
8	√	√	√	√	36.9	3.0	7.9

Ablation results in Table 2 demonstrate the effectiveness of proposed improvements. From the baseline's 33.0% mAP@0.5, sequential integration of modules culminates in 36.9% mAP@0.5 with the full configuration,while maintaining reasonable computational efficiency .

D. Comparative experiments

To further validate the effectiveness of our proposed method, comparative experiments were conducted against existing classical approaches on VisDrone2021 datasets under identical training parameters. The results are presented in Table 3.

Table 3
Comparative experimental results

Algorithm Model	FLOPs (G)	Parameters （M）	mAP@0.5(%)
YOLOv3-tiny	14.3	9.52	23.3
YOLOv5n	7.1	2.5	33.1
YOLOv8n	8.7	3.2	34.5
YOLOv10n	6.5	2.3	34.0
YOLOv11n	6.5	2.62	33.0
RS-Yolov11	7.9	3.0	36.9

Comparative experiments with mainstream YOLO-series models demonstrate that our proposed model exhibits superior detection performance and robust generalization capability. While achieving performance improvements, the model maintains reasonable computational complexity, indicating that our proposed enhancements achieve an excellent balance among model efficiency, detection accuracy, and generalization ability.

V. CONCLUSION

This study addresses the challenges in remote sensing image detection, particularly focusing on densely distributed targets with significant scale variations under complex backgrounds, while reducing model parameters for improved deployability. Our methodology incorporates several key improvements to YOLOv11:

we replace the original bottleneck structure with Faster_Block to reduce the computational complexity of the C3K2 module. The conventional detection head is superseded by a Dynamic Head that unifies scale, spatial, and task attention mechanisms, enhancing the network's capability to detect small targets in complex scenarios.we introduce SPD convolution for downsampling, facilitating feature fusion across different scales and improving detection accuracy. The WIOU loss function replaces the original CIOU loss to address the penalty term failure during prediction-ground truth box overlap, thereby enhancing localization precision.Experimental validation on the VisDrone2021 dataset demonstrates a 3.9% improvement in mean Average Precision compared to the original YOLOv11n, validating both the effectiveness and applicability of our proposed enhancements.

REFERENCES

[1] Liu L, Pan Z, Lei B, Learning a rotation invariant detector with rotatable bounding box[J]. arXiv preprint arXiv:1711.09405, 2017.

[2] Wan D, Lu R, Wang S, et al. YOLO-HR: Improved YOLOv5 for Object Detection in HighResolution Optical Remote Sensing Images[J]. Remote Sensing, 2023, 15(3): 614-631.[3]

[3] Liu X, Gong W, Shang L, et al. Remote Sensing Image Target Detection and Recognition Based on YOLOv5[J]. Remote Sensing, 2023, 15(18): 4459-4483.

[4] K. He, X. Zhang, S. Ren, et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," in Computer Vision – ECCV 2014, 346−361, Springer International Publishing (2014). [doi:10.1007/978-3-319-10578-9 23].

[5] S. Ren, K. He, R. Girshick, et al., "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis & Machine Intelli-gence 39(06), 1137−1149 (2017). [doi:10.1109/TPAMI.2016.2577031].

[6] J. Redmon, S. Divvala, R. Girshick, et al., "Y ou only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 779−788 (2016).

[7] A. Farhadi, "Yolo9000: Better, faster, stronger," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7263−7271 (2017).

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprintarXiv:1804.02767 (2018). [doi:10.48550/arXiv.1804.02767].

[9] A. Bochkovskiy, C.-Y . Wang, and H.-Y . M. Liao, "Y olov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934 (2020). [doi:10.48550/arXiv.2004.10934].

[10] Song G , Liu Y , Wang X .Revisiting the Sibling Head in Object Detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2020.DOI:10.1109/CVPR42600.2020.01158.

[11] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 840-849.

[12] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[13] Kong T, Sun F, Liu H, et al. Foveabox: Beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[14] Liu W, Liao S, Ren W, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5187-5196.

[15] Wang J, Chen K, Yang S, et al. Region proposal by guided anchoring[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2965-2974
[16] Howard A , Sandler M , Chen B ,et al.Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV).IEEE, 2020.DOI:10.1109/ICCV.2019.00140.
[17] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.
[18] Han K , Wang Y , Tian Q ,et al.GhostNet: More Features From Cheap Operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2020.DOI:10.1109/CVPR42600.2020.00165.
[19] Baffour A , Guo J , Kusi G .Depth-wise Separable Convolution for Real-time Facial Expression Recognition[J]. 2018.
[20] CHEN J, KAO S H, HE H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE Press,2023: 12021-12031.
[21] Dai X , Chen Y , Xiao B ,et al.Dynamic Head: Unifying Object Detection Heads with Attentions[J]. 2021.DOI:10.48550/arXiv.2106.08322.
[22] SUNKARA R, LUO T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Data- bases. Cham: Springer Nature Switzerland, 2022: 443-459.