

PSF-YOLO11: an enhanced YOLO11 framework for small object detection in UAV imagery

Dan Shan, Xiangdong Meng, Carlos Ferran, Xuan Tong, and Xifeng Guo*

Published 17 June 2026

Abstract

Images captured by unmanned aerial vehicles are characterized by complex backgrounds, significant scale variations, and a large proportion of small and densely distributed objects, which pose substantial challenges to object detection algorithms. To address these issues, this paper proposes PSF YOLO11, a detection framework for complex aerial scenes that aims to improve small object detection performance. First, to enhance high level semantic representation under complex backgrounds, a parallel multi scale convolution module is introduced into the high level feature extraction stages of YOLO11. Second, a small object feature pyramid is proposed to address the insufficient representation of small objects in shallow layers by redesigning the small object detection branch, thereby improving fine grained spatial feature extraction and multi scale information interaction. Third, to alleviate feature inconsistency during multi scale aggregation, a shared weight feature pyramid structure is designed to enhance feature fusion effectiveness. Experimental results on the VisDrone2019 dataset demonstrate that, compared with the baseline YOLO11 model, the proposed method improves $mAP@0.5$ and $mAP@0.5:0.95$ by 4.6% and 3.0%, respectively, while achieving a 3.5% improvement in Recall over the baseline. On the drones in optics recognition dataset, which contains a broader range of aerial object categories and diverse imaging conditions, the proposed method achieves improvements of 2.0% in $mAP@0.5$ and 2.3% in $mAP@0.5:0.95$, and 1.9% in Recall over the baseline. These results indicate that the proposed method exhibits superior performance in detecting small and densely distributed objects in complex aerial environments, with strong generalization capability across diverse remote sensing scenarios.

For the full article, you need to request it directly from the journal. Below is a two-page summary for your convenience.

Summary

Overview and Problem Statement

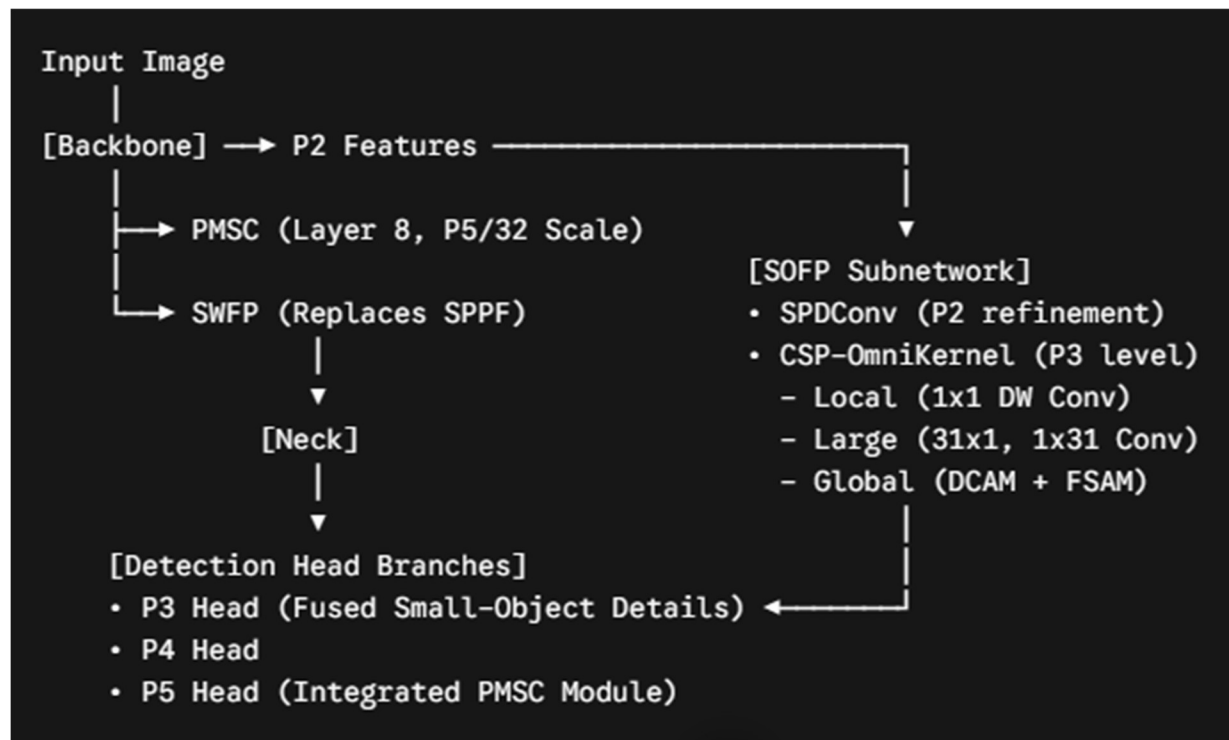
Unmanned aerial vehicle (UAV) imagery presents severe challenges for computer vision models, including extreme scale variation, dense object distributions, sensor resolution limits, oblique viewpoints, and complex visual clutter (e.g., shadows, buildings, and vegetation). From a measurement science perspective, aerial detection functions as a non-contact optical measurement task requiring precise geometric localization and semantic identification. However,

small targets (defined as objects occupying less than 32×32 pixels) suffer extensive pixel-level degradation during progressive downsampling operations.

Existing one-stage detectors such as baseline YOLO11 offer favorable inference speed but exhibit limited sensitivity toward small targets and suffer from scale inconsistency. While adding dedicated low-level prediction heads (such as a P2 head) can improve small object capture, it introduces substantial computational overhead, increased latency, and complex post-processing. To address these trade-offs, Shan et al. (2026) introduce **PSF-YOLO11**, an enhanced detection framework designed to boost small-object representation accuracy while maintaining real-time edge viability.

Architectural Enhancements

PSF-YOLO11 retains the general backbone-neck-head structure of YOLO11 while integrating three targeted architectural modules:



1. Parallel Multi-Scale Convolution (PMSC)

The PMSC module improves high-level semantic representation and contextual modeling. It operates via a channel-split mechanism:

- **Feature Preservation Branch (X1):** Directly propagates half of the input channels without convolution to preserve fine-grained spatial information.

- **Multi-Scale Convolution Branch (X2):** Divides the remaining channels into G groups (optimized at $G=2$) processed in parallel using 3×3 and 5×5 convolutions to capture diverse receptive fields without substantial parameter growth.
- **Integration:** Outputs are stacked, rearranged, concatenated with $X1$, and refined via a 1×1 convolution.

PMSC is inserted at layer 8 of the backbone and layer 24 of the neck/head.

2. Small Object Feature Pyramid (SOFB)

Rather than introducing a dedicated P2 detection head, SOFB refines shallow P2 features and fuses them directly into the P3 detection branch:

- **Spatial-to-Depth Conversion (SPDConv):** Replaces strided convolutions or pooling, which discard crucial spatial data, by folding spatial dimensions into the channel dimension to preserve weak target cues.
- **CSP-OmniKernel:** Positioned at the P3 level to balance multi-scale representation across three parallel paths:
 1. *Local Branch:* Employs 1×1 depthwise convolutions for minute spatial textures.
 2. *Large Branch:* Employs large ($K=31$) and strip-shaped convolutions (1×31 , 31×1) for extended contextual cues.
 3. *Global Branch:* Couples Dual Domain Channel Attention (DCAM) with Frequency Spatial Attention (FSAM) using Fast Fourier Transforms (FFT/IFFT) to suppress background clutter.

3. Shared-Weight Feature Pyramid (SWFP)

Replacing standard Spatial Pyramid Pooling Fast (SPPF) structures, SWFP reduces multi-scale feature inconsistency:

- Compresses channels by half via a 1×1 convolution.
- Employs shared 3×3 convolutional kernel parameters across successive dilation rates of 1, 3, and 5.
- Captures expanding receptive fields without growing the parameter count, aligning features across scales and mitigating scale mismatch errors.

Experimental Evaluation and Results

The framework was evaluated on two demanding aerial benchmarks: the **VisDrone2019** validation set (small and dense UAV targets across 10 classes) and the **DIOR** dataset (optical remote sensing targets across 20 classes). Testing was conducted at 640×640 resolution on an NVIDIA RTX 4060 Ti GPU.

Performance Comparison Across Benchmarks

Model	Parameters (M)	GFLOPs	VisDrone mAP@0.5 (%)	VisDrone Recall (%)	DIOR mAP@0.5 (%)	Inference Speed (FPS)
Faster R-CNN	118.7	282.4	25.6	23.0	72.4	21.6
RT-DETR	42.0	85.1	31.5	32.7	83.5	34.8
YOLOv8	2.0	8.7	30.1	30.9	83.1	354.6
YOLO11 Baseline	2.6	6.5	30.4	31.4	84.1	261.5
YOLO-MS	23.3	79.1	32.3	33.6	86.3	95.2
PSF-YOLO11 (Proposed)	3.0	12.4	35.0	34.9	86.1	213.0

On VisDrone2019, PSF-YOLO11 improved mAP@0.5 by 4.6% and Recall by 3.5% over baseline YOLO11, dramatically cutting missed detections among densely clustered targets. On DIOR, the model achieved an mAP@0.5 of 86.1% (a 2.0% increase over baseline). While YOLO-MS achieved a fractionally higher mAP@0.5 on DIOR (86.3%), it required nearly eight times more parameters (23.3 M vs. 3.0 M) and ran at less than half the frame rate (95.2 FPS vs. 213.0 FPS).

Key Ablation Insights

- **SOFP as Core Contributor:** Introducing SOFP alone contributed the largest gain on VisDrone2019 (+2.6% mAP@0.5), proving that preserving early spatial details is paramount for small objects.
- **SPDConv Superiority:** Replacing SPDConv with standard strided convolutions or max pooling dropped mAP@0.5 by 0.8% and 1.2%, respectively, confirming the necessity of non-lossy spatial-to-channel transformation.
- **Synergistic Integration:** While PMSC or SWFP alone provided minimal or slightly negative shifts when isolated, combining all three modules produced the peak score (35.0% mAP@0.5). This highlights that PMSC contextualizes the low-level details preserved by SOFP, while SWFP stabilizes multi-scale alignment.

Limitations and Future Directions

The authors identified four core limitations in the current framework:

1. **Ambiguous Feature Sensitivity:** Enhanced shallow sensitivity occasionally causes false positives in uniform or textureless regions that visually mimic tiny objects.
2. **Computational Overhead:** Processing high-resolution P2 features nearly doubles the computational requirement from 6.5 to 12.4 GFLOPs.
3. **Absence of On-Board Hardware Testing:** Evaluations were conducted on desktop GPUs rather than embedded, low-power edge units (e.g., NVIDIA Jetson modules) subjected to flight vibrations, thermal throttles, and power constraints.
4. **Data Dependency:** The architecture relies heavily on extensive manual bounding box annotations, limiting rapid deployment to novel aerial scenes.

Proposed future research includes edge-oriented model compression (INT8 quantization, structured channel pruning, and knowledge distillation), integration of evidential deep learning or Monte Carlo dropout for uncertainty estimation, and semi-supervised or few-shot learning paradigms.