HRSNet: Hierarchical Recursive Scaling for Efficient UAV Object Detection

Bowen Yang, Qing Dong, and Gang Wu*

Computer Science and Technology, School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China;
wugang@mail.neu.edu.cn

Abstract. The rapid development of UAV technology has driven changes in fields such as intelligent transportation and disaster rescue, and the rise of low-altitude economy has further accelerated its application. In this context, UAV object detection technology faces challenges such as large object size difference, high density, and complex background. To this end, we propose a lightweight and efficient detector, HRSNet, whose core consists of three self-designed modules: the Feature Pyramid Shared Convolution Module (FPSCM) to achieve efficient fusion of multi-scale features, the Dynamic Aware Modulation Module (DAMM) to reduce background interference, and the Hierarchical Receptive Field Scale Detection Head (HRSDH) to improve classification and localization accuracy. Experimental results show that HRSNet outperforms the baseline by 2.1% and achieves 41.2% mAP50 while reducing parameters by 0.28% on the VisDrone dataset, while maintaining a good balance between accuracy and efficiency.

Keywords: UAV · object detection · high precision · YOLOv10 · real time.

1 Introduction

The rapid advancement of UAV technology has propelled its crucial role in smart transportation, disaster rescue, and infrastructure inspection. With their flexible deployment and wide coverage capabilities, UAVs have become essential perception nodes in smart city systems. However, as shown in Figure 1, UAV-based object detection faces significant challenges: identifying objects in dense environments and tracking dynamic objects in complex urban landscapes, demanding high accuracy, real-time performance, and environmental adaptability. Although current mainstream detection frameobjectworks demonstrate strong performance in general scenarios, they fail to address the unique challenges of UAV object detection. Through comprehensive benchmarking, we identify two critical limitations in existing approaches: (1) Transformer-based detectors (OWRT-DETR) [1] achieve superior context modeling with their global attention mechanisms, but suffer from prohibitive computational complexity, limiting their real-time performance on UAV edge devices. In contrast, HRSNet achieves

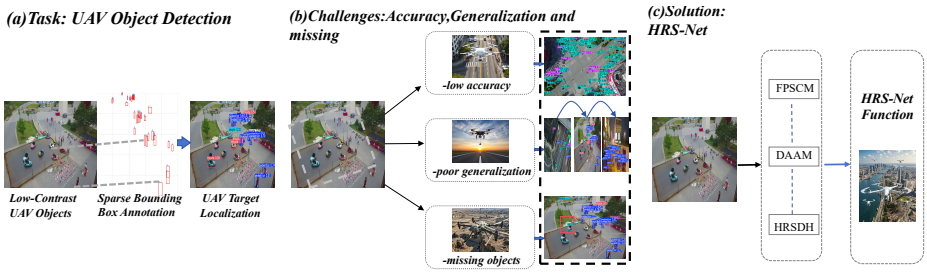


Fig. 1: The contribution of HRSNet.

real-time performance through our innovative shared-parameter backbone, which significantly reduces computational requirements while fully maintaining detection accuracy. This architectural innovation enables efficient deployment on UAV edge devices without compromising model capability. (2) Lightweight models like WSA-YOLO [2] achieve 20-30 FPS through model compression, but at the cost of significantly reduced detection effectiveness, particularly for small objects. HRSNet addresses this through our FPSCM module, which enhances small object recall by 1.8% via multi-scale feature fusion, while maintaining real-time performance. Existing methods fail to address UAV-specific challenges like spatial correlations between objects and dynamic interference suppression. To overcome these limitations, we propose HRSNet, an efficient lightweight detector with three key innovations:

1. We propose HRSNet, a high resolution network for small object detection in aerial imagery that introduces three key innovations to address the challenges of feature degradation, complex backgrounds, and computational efficiency in drone-based vision applications.
2. We propose the FPSCM that enhances feature representation through cross-layer feature reuse and parameter sharing.
3. The DAMM module enhances detection in complex scenarios through gated feature selection and adaptive background suppression, while the HRSDH detection head improves efficiency via task decoupling and dynamic sample allocation.
4. HRSNet achieves pioneering performance, attaining 41.2% mAP50 using merely 7.2M parameters VisDrone and surpassing existing detectors in accuracy and efficiency.

2 Related Work

2.1 Traditional and Deep Learning for UAV Detection

Early UAV detection relied on hand-designed features like Haar [3] and Adaboost [4], but these methods struggled with small objects [5] and complex backgrounds. While CNN-based detectors (Faster R-CNN, YOLO) improved performance, they still face challenges with small UAV objects and background interference [6]. Recent works address these limitations: Ye et al. [7] combined CNNs

and transformers via LEM/ECTB modules, while Liu et al. [8] introduced self-attention in Trans R-CNN. Others like TLSTMF-YOLO [9] and TTSDA-YOLO [10] employed attention mechanisms and multi-scale fusion, yet the accuracy-speed tradeoff remains challenging for real-time aerial scenarios.

2.2 Real Time Object Detection for UAV Applications

Recent advances in real-time object detection leverage optimized deep learning architectures to balance computational efficiency and accuracy. Zhang et al. [11] presented the FFCA-YOLO, integrating three lightweight plug and play modules to optimize accuracy speed balance under resource constraints. Zhao et al. [12] proposed SLGA using prototype learning to enhance detection in complex backgrounds via scene level supervision.

Although current research on high precision real time object detection [13] has made significant breakthroughs in performance metrics, these methods still face a fundamental challenge: how to achieve an intrinsic improvement in computational efficiency while maintaining feature characterization capabilities. Existing schemes often increase network complexity in exchange for accuracy improvement, but fail to establish an adaptive matching mechanism between feature extraction and computational resources, leading to difficulties in achieving the ideal accuracy efficiency balance in resource constrained scenarios. Unlike the previous works, our method proposes the full fusion of multi-scale features [14] through cross layer feature multiplexing and parameter sharing mechanisms, and proposes a new dynamic input aware modulation module to maintain high accuracy of detection.

3 Methods

3.1 HRSNet: An Overview of the Architecture

HRSNet enhances the YOLOv10 architecture with specific optimizations for small object detection while maintaining computational efficiency. Our design addresses three key limitations in existing approaches: (1) inefficient feature extraction for small objects; (2) redundant network parameters; and (3) excessive computation in detection heads. As shown in Figure 2, the framework introduces three core innovations:

a) Parameter-Shared Backbone: The backbone $\mathcal{B}$ employs a novel convolution scheme with cross layer parameter sharing. This design stems from the observation that small objects benefit from consistent low level feature extractors while reducing network complexity.

b) Adaptive Feature Fusion: To dynamically enhance small object features, we propose a gating mechanism:

$$\mathcal{F}_{\text{DAAM}} = \sigma(\mathbf{W}_g \mathbf{x} + \mathbf{b}_g) \odot \mathcal{F}_{\text{in}} \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters, and $\odot$ denotes element wise multiplication. This automatically suppresses irrelevant background while preserving subtle object patterns.

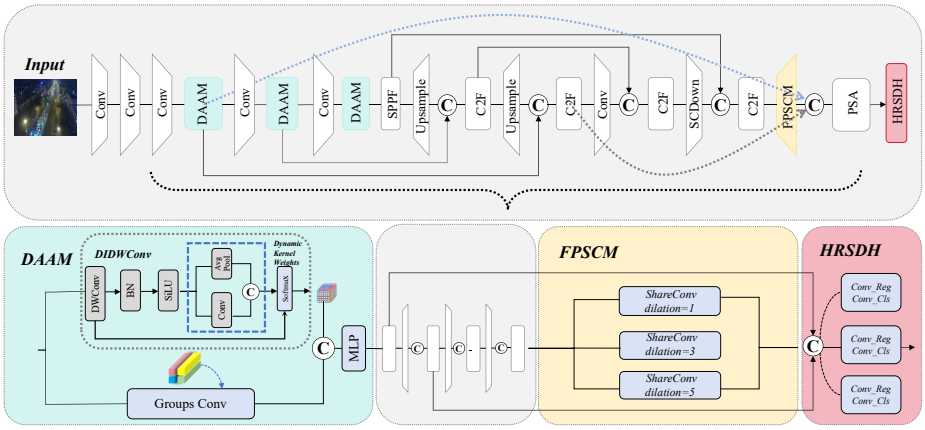


Fig. 2: Architecture of the HRSNet.

where $\mathbf{x}$ is the input feature vector to the gating mechanism, $\mathcal{F}_{\text{in}}$ represents the input feature map to be modulated.

c) Efficient Detection Head: The head uses depthwise separable convolution to maintain spatial sensitivity with reduced computation:

$$\mathcal{H}(x) = \mathcal{DW}_{k\times k}(\mathcal{PW}_{1\times 1}(x)) \tag{2}$$

where $\mathcal{DW}$ and $\mathcal{PW}$ represent depth wise and point-wise convolutions respectively. This design is crucial for processing high-resolution feature maps efficiently.
where $k \times k$ denotes the kernel size of the depthwise convolution operation, x is the input feature map to the detection head.

The complete framework integrates these components to achieve balanced performance in accuracy and speed for small object detection tasks.

3.2 Feature Pyramid Shared Convolution Module (FPSCM)

To address the computational redundancy and feature inconsistency in conventional multi-scale feature extraction, we propose the FPSCM based on two key insights: (1) dilated convolutions should share fundamental feature patterns across scales, and (2) sequential processing is more computationally efficient than parallel branches.

The input feature map $x \in \mathbb{R}^{B\times C_1\times H\times W}$ first undergoes channel reduction for efficiency:

$$y_0 = \text{Conv1}(x), \quad y_0 \in \mathbb{R}^{B\times \frac{C_1}{2}\times H\times W} \tag{3}$$

where x is the input feature tensor with batch size B, C_1 input channels, height H and width W, y_0 is the channel-reduced intermediate feature map, $\text{Conv1}(\cdot)$ denotes a pointwise convolution with output channels $C_1/2$.

Weight shared dilated convolutions are then applied sequentially with increasing dilation rates $\mathcal{D} = \{1, 3, 5\}$:

$$y_i = \text{Conv}_{d_i}(y_{i-1}), \quad d_i \in \mathcal{D} \tag{4}$$

where y_i represents the output feature map at dilation stage i, $\text{Conv}_{d_i}(\cdot)$ is a weight shared 3×3 convolution with dilation rate d_i, $\mathcal{D}$ is the predefined set of dilation rates.

By concatenating intermediate features before final projection, the module maintains multi-scale representation while optimizing computational efficiency.

3.3 Dynamic Aware Modulation Module (DAMM)

To address the limitations of standard convolutions and attention mechanisms, we propose the DAAM. This module enables input adaptive feature modulation through a lightweight gating mechanism that dynamically adjusts feature importance without expensive attention computations.

Given an input feature map $x \in \mathbb{R}^{B \times C \times H \times W}$, DAAM performs the following transformation:

$$x \leftarrow x + \mathcal{P}_1\big(\boldsymbol{\lambda}_1 \circledast \mathcal{M}(\text{BN}_1(x))\big) \tag{5}$$

where $\mathcal{P}_1 : \mathbb{R}^{C \times H \times W} \to \mathbb{R}^{C \times H \times W}$ denotes a 1×1 pointwise convolution that preserves spatial dimensions, $\boldsymbol{\lambda}_1 \in \mathbb{R}^C$ represents a learnable channel-wise scaling vector, $\circledast$ indicates channel wise multiplication, $\text{BN}_1(\cdot)$ stands for batch normalization.

The core dynamic mixer $\mathcal{M}(\cdot)$ integrates multi-branch features through:

$$\mathcal{M}(x) = \sum_{k=1}^{3} \text{DWConv}_k(x) \odot \sigma(\mathbf{W}_k \text{AvgPool}(x) + \mathbf{b}_k) \tag{6}$$

where $\mathbf{W}_k \in \mathbb{R}^{C \times C}$ as the branch-specific transformation matrix, $\mathbf{b}_k \in \mathbb{R}^C$ as the branch specific bias term, $\sigma(\cdot)$ denoting the sigmoid activation function, $\odot$ representing element-wise multiplication.

The design achieves an optimal balance between representation power and computational efficiency, making it suitable for resource constrained applications.

3.4 Hierarchical Receptive Field Scale Detection Head (HRSDH)

To address the challenges of multi-scale feature extraction and cross-scale interaction in small object detection, we propose the HRSDH with two key components. First, the local feature enhancement branch processes high resolution feature maps through parallel GroupNorm convolution layers with adaptive kernel sizes:

$$\mathcal{F}_i = \text{ConvGN}_{k \times k}(P_i), \quad k \in \{1, 3\},\ i \in \{3, ..., 7\} \tag{7}$$

where P_i represents the input feature map at pyramid level i. Second, the global context integration branch employs a novel scale-aware fusion mechanism that dynamically combines features from different pyramid levels using bilinear interpolation for precise feature resizing:

$$\mathcal{F}_j = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \text{Conv}_{3\times3}(\text{Resize}_s(P_j, \text{bilinear})), \quad j \in \{9, 12\} \tag{8}$$

where $\mathcal{S}$ denotes the set of target scales for feature fusion, P_j represents the input feature map at pyramid level j, and $\text{Resize}_s(\cdot, \text{bilinear})$ performs scale adjustment using bilinear interpolation to preserve feature continuity.

The module combines multi-resolution features through channel concatenation and a lightweight projection, maintaining efficiency via depthwise separable convolutions while preserving scale awareness. BatchNorm and ReLU ensure stable training.

4 Experiment

4.1 VisDrone Dataset

VisDrone [15] is an authoritative object detection benchmark for UAV aerial photography scenarios. The dataset contains 6,471 training set images, 548 validation set images, and 3,190 test set images, for a total of 10,209 high-resolution images, with more than 540,000 bounding boxes annotated, covering 10 categories of objects such as pedestrians and vehicles.

4.2 Experiment Set Up

Our experiments ran on Ubuntu 20.04 LTS with 16 vCPU Intel Xeon Gold 6430 and NVIDIA RTX 4090. We implemented HRSNet using Python 3.8 with PyTorch 1.11.0 (CUDA 11.3). Evaluation metrics included mAP50, GFLOPs, and parameter count to assess deployment efficiency in resource limited scenarios.

4.3 Ablation Experiment Analysis

We conducted ablation studies on VisDrone to validate HRSNet's components. The baseline YOLOv10s showed limited performance, particularly in late-stage mAP50 saturation. Adding FPSCM improved detection with faster mid-training convergence, demonstrating effective spatial compression. The FPSCM+DAAM combination achieved 40.1% mAP50 (1% gain over FPSCM alone), where DAAM's depth-invertible structure enhanced occlusion handling while reducing parameters. The full HRSNet system achieved a 2.1% mAP50 improvement over baseline, with 7.20M parameters due to HRSDH's multi-scale processing branch. These results demonstrate HRSNet's optimal balance between accuracy and efficiency through FPSCM's compression, DAAM's lightweight modeling, and enhanced feature integration(see Table 1).

Table 1: Outcome Analysis of Ablation Experiments

Dataset	Model	Para. (M)	GFLOPs	P (%)	R (%)	mAP50 (%)
VisDrone	YOLOv10s	7.22	16.5	49.8	38.3	39.1
VisDrone	+FPSCM	8.69	22.1	51.1	38.4	39.7
VisDrone	+DAAM	6.66	15.5	49.7	39.4	40.1
VisDrone	+HRSDH	7.20	18.2	49.2	37.8	39.4
VisDrone	HRSNet	7.20	18.4	50.6	40.1	41.2

4.4 Relative Performance Evaluation of Enhanced Algorithms

We evaluate HRSNet against a baseline model on the VisDrone dataset, with quantitative results in Table 2. HRSNet demonstrates improved detection perfor-

Table 2: Performance Metrics Comparison of Models on Visdrone

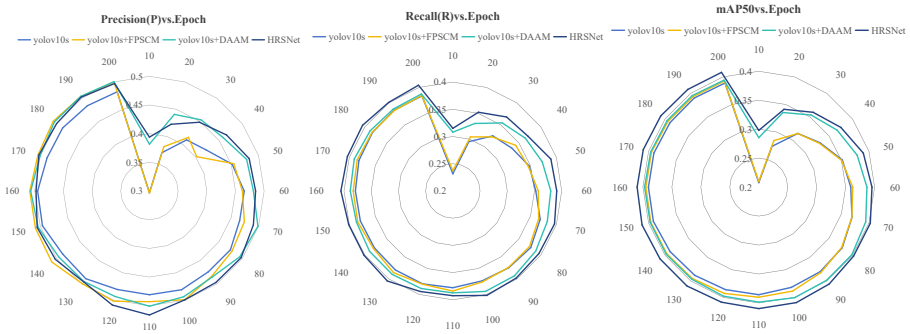
Module	Dataset	Par. (M)	GFLOPs	P (%)	R (%)	mAP50 (%)
YOLOv10s	VisDrone	7.22	16.5	49.8	38.3	39.1
HRSNet	VisDrone	7.20	16.7	50.6	40.1	41.2

mance on the VisDrone dataset with a 2.1 percent increase in mAP50 from 39.1% to 41.2% while maintaining low computational complexity between 16.5 and 16.7 GFLOPs. The model shows enhanced adaptability to complex UAV aerial scenarios and meets real-time processing requirements. Overall, the HRSNet model not only achieved enhanced detection performance but also successfully overcame the computational challenges typical in drone-based aerial imaging tasks. Particularly in demanding UAV visual processing scenarios, the model maintains superior monitoring accuracy while employing a more efficient parameter architecture, marking substantial progress in real-time aerial object detection.

Figure 3 vividly illustrates HRSNet's performance progression on the VisDrone dataset across training epochs, showing three key improvements: 1) a sustained increase in mAP50, indicating enhanced object detection capability; 2) rising precision values, demonstrating better identification accuracy; and 3) improved recall rates, reflecting greater proficiency in recognizing true objects. These combined metrics clearly establish HRSNet's superior and more efficient detection performance.

4.5 Benchmarking Against Leading Algorithms

As shown in Table 3, HRSNet leads the other comparison models across the board in terms of key performance metrics. On the VisDrone dataset, it achieves



(a) Precision (b) Recall (c) mAP50

Fig. 3: Performance comparison on VisDrone dataset showing: (a) Precision curves; (b) Recall curves; and (c) mAP50 trend over training iterations (y-axis to 200).

a precision of 50.6%, a recall of 40.1%, and a mAP50 of 41.2%. These metrics significantly outperform mainstream detection models including YOLOv8, fully demonstrating HRSNet's stable detection capability for small objects in complex aerial scenarios, especially in dealing with high-density objects and multi-scale variations.

Table 3: Performance comparison of SOTA models on VisDrone

Model	Precision (%)	Recall (%)	mAP50 (%)	Params (M)
RMVAD-YOLO [16]	46.3	35.6	33.8	10.61
YOLOv5s	49.5	37.8	38.5	9.12
YOLOv8s	48.3	38.6	38.5	9.83
YOLOv10s	49.8	38.3	39.1	7.22
YOLOv11s	49.4	38.9	39.4	9.42
TPH-YOLO	48.4	38.0	39.3	9.20
FFCA-YOLO [11]	48.5	35.8	37.0	7.32
PS-YOLO [17]	51.4	39.4	40.7	5.53
HRSNet (Ours)	**50.6**	**40.1**	**41.2**	**7.20**

4.6 Experimental Results in Visual Format

Figure 4 demonstrates HRSNet's superior small object detection on UAV aerial challenges. Unlike YOLOv10s, which suffers from missed detections and poor localization, HRSNet achieves precise bounding box regressionespecially for clustered and low-contrast objects.

Fig. 4: Visual Comparison of HRSNet Detection Results.

This improvement stems from enhanced feature representation and multi-scale processing, making HRSNet particularly effective for UAV-based aerial inspection. Its balance of accuracy and efficiency enables robust detection in complex aerial scenarios, even for small and occluded targets critical in UAV surveillance and monitoring applications.

5 Conclusions

HRSNet is an advanced UAV detection framework that tackles small object challenges through three innovations: FPSCM for multi-scale fusion, DAAM for irregular object modeling, and Dynamic Sparse Convolution for efficient feature learning. On VisDrone benchmark, it achieves superior accuracy in dense scenarios while maintaining real-time performance. Future work will explore deformable attention for enhanced ultra-dense detection.

References

1. S. Ma, Y. Zhang, L. Peng, C. Sun, L. Ding, and Y. Zhu, "Owrt-detr: A novel real-time transformer network for small object detection in open water search and rescue from uav aerial imagery," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2025.
2. Y. Hui, J. Wang, and B. Li, "Wsa-yolo: Weak-supervised and adaptive object detection in the low-light environment for yolov7," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.

3. D. Y. Ardhito, D. Susilo, D. Ruswanti, D. Retnoningsih, A. Kristianto, and Setiyowati, "Employee attendance through face recognition using the haar cascade classifier method," in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, 2024, pp. 1–4.
4. X. Kejun, W. Jian, N. Pengyu, and H. Jie, "Automatic nipple detection using cascaded adaboost classifier," in *2012 Fifth International Symposium on Computational Intelligence and Design*, vol. 2, 2012, pp. 427–432.
5. X. Yan, B. Shen, and H. Li, "Small objects detection method for uavs aerial image based on yolov5s," in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, 2023, pp. 61–66.
6. Z. Zhuang, P. Liu, D. Xu, and J. Cheng, "Yolo-ked: A novel framework for rotated object detection in complex environments," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
7. T. Ye, W. Qin, Z. Zhao, X. Gao, X. Deng, and Y. Ouyang, "Real-time object detection network in uav-vision based on cnn and transformer," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.
8. H.-I. Liu, Y.-W. Tseng, K.-C. Chang, P.-J. Wang, H.-H. Shuai, and W.-H. Cheng, "A denoising fpn with transformer r-cnn for tiny object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
9. S. Meng, Z. Shi, S. Pirasteh, S. Liberata Ullo, M. Peng, C. Zhou, W. Nunes Gonçalves, and L. Zhang, "Tlstmf-yolo: Transfer learning and feature fusion network for earthquake-induced landslide detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.
10. M. Zhang, Q. Rong, and H. Jing, "Ttsda-yolo: A two training stage domain adaptation framework for object detection in adverse weather," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.
11. Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "Ffca-yolo for small object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
12. T. Zhao, R. Feng, and L. Wang, "Scene-yolo: A one-stage remote sensing object detection network with scene supervision," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.
13. Y. Tao, Z. Zongyang, Z. Jun, C. Xinghua, and Z. Fuqiang, "Low-altitude small-sized object detection using lightweight feature-enhanced convolutional neural network," *Journal of Systems Engineering and Electronics*, vol. 32, no. 4, pp. 841–853, 2021.
14. H. Lin, B. Liu, G. Zhang, Q. Yin, L. Yang, and P. Lan, "Multi-scale feature fusion network for lip recognition," in *2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA)*, 2024, pp. 541–545.
15. P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.
16. K. Li, X. Zheng, J. Bi, G. Zhang, Y. Cui, and T. Lei, "Rmvad-yolo: A robust multi-view aircraft detection model for imbalanced and similar classes," *Remote Sensing*, vol. 17, no. 6, 2025. [Online]. Available: https://www.mdpi.com/2072-4292/17/6/1001
17. H. Zhong, Y. Zhang, Z. Shi, Y. Zhang, and L. Zhao, "Ps-yolo: A lighter and faster network for uav object detection," *Remote Sensing*, vol. 17, no. 9, 2025. [Online]. Available: https://www.mdpi.com/2072-4292/17/9/1641