Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080341

ARTICLE

MFCI-YOLO: Lightweight UAV Aerial Photography Small Object Detection Method Based on Multi-Scale Feature Fusion and Contextual Information

Weiguang Wang[1,2], Jincai Li[1], Mengqi Liu[1], Mengke Liu[1], Yuan Zhang[1], Jingyan Wu[1,*], Yang Liu[3,*], Junbin Lou[4] and Yixin He[5]

[1]School of Information Engineering, Henan University of Science and Technology, Luoyang, China
[2]Industry Research Institute of Intelligent Systems, Longmen Laboratory, Luoyang, China
[3]School of Electrical and Information Engineering, Guangdong Baiyun University, Guangzhou, China
[4]College of Mechanical Engineering, Jiaxing University, Jiaxing, China
[5]College of Information Science and Engineering, Jiaxing University, Jiaxing, China

*Corresponding Authors: Jingyan Wu. Email: wujingyan@jxust.edu.cn; Yang Liu. Email: liuyang@jxust.edu.cn

Received: 07 February 2026; Accepted: 08 May 2026; Published: 15 June 2026

ABSTRACT: To improve the accuracy of small object feature detection in complex backgrounds for Unmanned Aerial Vehicle (UAV) aerial photography and reduce computational complexity, we propose the lightweight UAV aerial photography small object detection method based on multi-scale feature fusion and contextual information. Firstly, by introducing the grouped content-aware reassembly (GCA) operator and designing lightweight pinwheel context convolution (LPConv), we extend the feature fusion path to the P2 layer, constructing a lightweight multi-scale feature fusion network (SG-PANet). Through the decoupling of fine-grained small object features and background interference features by the GCA operator, combined with the anisotropic receptive field constructed by LPConv, our proposed method can effectively preserve the geometric details of small objects. Furthermore, we introduce the cross-stage dense feature refinement (CSPStage) module as the pre-refining unit of the detection head, and use the full history state awareness mechanism to strengthen feature reuse and gradient propagation to solve the problem of feature degradation across layers. We utilize the Wise-IoU v3 loss function to dynamically optimize the gradient gains of high-quality and low-quality samples, thereby enhancing the detection accuracy and convergence speed of the proposed method in complex scenarios. Finally, we verified the superiority and generalization of the proposed method on the VisDrone2019 dataset and DOTAv1.5 dataset. The results show that compared with YOLOv11n, MFCI-YOLO's detection mAP50-95 increased by 11.1%, small object mAP50 increased by 16.1%, and mAP50 reached 80.3%. It provides a practical solution for detecting small objects in dense scenes.

KEYWORDS: Small object detection; UAV aerial photography; YOLOv11n; grouped structure; context-aware; feature fusion; lightweight

1 Introduction

With UAVs widely deployed in the global low-altitude economy, establishing robust communication and edge-computing infrastructures—such as Internet of Things (IoT)-enhanced emergency communications [1], Multiple-Input Multiple-Output (MIMO) cooperative networks [2], Non-Orthogonal Multiple Access with Mobile Edge Computing (NOMA-MEC) offloading [3], and Reconfigurable Intelligent Surface (RIS)-based Integrated Sensing and Communication (ISAC) frameworks [4]—has become a foundational prerequisite. However, the ultimate efficacy of these advanced networks relies entirely on accurate visual

perception. In such high-altitude missions, critical targets inherently appear as extremely small objects. Consequently, robust small-object detection has become the perceptual bottleneck and crucial enabler for UAV intelligence within these complex networks. However, existing UAV aerial detection suffers from severe missed or false detections due to extreme perspective changes and complex scenes.

Selecting appropriate algorithms is key for accurate scene analysis. Mainstream Convolutional Neural Network (CNN)-based object detection algorithms perform poorly in UAV aerial small-object detection, as UAV aerial photography has three features: (1) small object pixel ratio; (2) limited hardware computing power and storage; (3) complex scenes with strong external interference.

YOLO series models are widely used in UAV aerial detection for their speed and accuracy. Li et al. [5] proposed SOD-YOLO (enhanced YOLOv8) to solve blurred small-object features, but its high-resolution head limits edge inference speed. Wan et al. [6] proposed a dynamic attention-based ultra-lightweight method to reduce information loss, yet complex attention impairs real-time performance. Yu and Mo [7] proposed YOLO-GCOF to alleviate computational bottlenecks, but lightweight pruning reduces robustness. In summary, UAV object detection research focuses on: (1) improving accuracy while maintaining efficiency via lightweight design; (2) enhancing small-object feature representation in complex backgrounds.

Based on YOLOv11n, we propose a lightweight UAV aerial small-object detection method using multi-scale feature fusion and contextual information, addressing feature drown-out, misalignment and limited localization robustness in extremely small-object cross-scale fusion. Specifically, to meet the strict deployment standards of resource-constrained UAV edge devices, we explicitly set our lightweight design objects as: total parameters <5M and computational complexity <50 Giga Floating-point Operations Per Second (GFLOPs), while ensuring real-time inference capability. Unlike recently proposed YOLO-based UAV detectors cited in Section 2—such as SOD-YOLO [5], which relies on computationally heavy high-resolution heads, or DAU-YOLO [6], which utilizes complex attention mechanisms that remain susceptible to feature drowning in dense backgrounds—our method fundamentally resolves the efficiency-accuracy conflict. We advance beyond these works by introducing extremely lightweight orthogonal convolutions and decoupled reconstruction mechanisms rather than stacking redundant parameters.

To provide a sharper distinction between our architectural improvements and novel modules, the core innovations and contributions of this paper are summarized as follows:

(1) Novel Modular Operators (LPConv & GCA): We propose the Lightweight Pinwheel Context Convolution (LPConv) to capture anisotropic contextual features of small objects via a dual-stream orthogonal alignment mechanism, drastically reducing parameter overhead. Concurrently, we introduce the Grouped Content-Aware (GCA) operator, a novel decoupled upsampling module that isolates fine-grained object features from complex backgrounds to fundamentally prevent feature drowning.

(2) Architectural Improvement (SG-PANet): Leveraging the aforementioned modules, we design a structural paradigm shift named the Scale-Gradient and Context-Aware Feature Fusion Network (SG-PANet). By abandoning the deep P5 layer and extending the fusion hierarchy directly to the high-resolution P2 layer, this architecture explicitly preserves the geometric details of extremely small objects that are typically lost in standard YOLO frameworks.

(3) Feature Refinement and Dynamic Optimization: We integrate the CSPStage module as a pre-refining unit before the detection head to leverage full-history state awareness, bridging semantic gaps and mitigating deep feature decay. Additionally, the Wise-IoU v3 loss function is employed to dynamically allocate gradient gains, significantly boosting localization robustness for low-quality small object instances.

(4) MFCI-YOLO balances accuracy and efficiency, achieving 48.2% mAP50 at 97 Frames Per Second (FPS) on VisDrone2019. An 80.3% mAP50 on DOTAv1.5 further validates its generalization for practical small-object detection in dense scenes.

2 Related Work

To resolve the core conflict between high-precision detection and computational constraints in UAV aerial photography, two main technical approaches are adopted: feature enhancement to strengthen effective features in small-object and occlusion scenarios, and lightweight network design to reduce overhead, balancing performance and deployability.

For small-object feature loss and severe background interference in UAV aerial photography, studies focus on feature fusion optimization. Jian et al. [8] introduced Bidirectional Feature Pyramid Network (BiFPN) into YOLOv5 to solve unidirectional fusion information loss. Ma and Wang [9] propose a dynamically weighted multi-scale fusion module to improve small-object recall. Lai et al. [10] design an efficient framework with SSFF and MSFE modules to alleviate dense-scene feature blurring. Qiao et al. [11] use recursive feature pyramids, and Sun et al. [12] adopt heterogeneous architectures to optimize feature representation.

To address edge deployment's computational constraints and low accuracy, researches focus on lightweight network reconstruction. Ye and Li [13] use Ghost Bottleneck to reduce complexity and improve performance. Ju et al. [14] integrate LSKA and DCNv4 into YOLO for better geometric perception. Zhao et al. [15] propose a lightweight scheme for UAV search-and-rescue to detect minute objects on low-power devices. Ji et al. [16] combine lightweight convolutions with attention, and Zhu et al. [17] designe a dynamic sparse attention mechanism to balance performance and efficiency.

Despite recent advancements, existing methods exhibit critical bottlenecks. BiFPN and Recursive FPN lack noise suppression, lightweight convolutions sacrifice contextual awareness, and traditional heads fail in dense occlusions. Recent advancements utilizing Spatial Pyramid Multi-scale Common Convolution (SPMCC) [18], Excitation and Modulation Attention (EMA) [19], and content-aware reassembly [20] improve saliency, yet their channel-shared operations often cause small-target 'feature drowning'. Meanwhile, architectures utilizing composite multi-scale fusion [21], multi-stage path aggregation [22], and spatially enhanced polarity sensing [23] achieve high precision but incur prohibitive computational overhead for resource-constrained UAV deployment. To resolve these conflicts, we propose MFCI-YOLO. By synergizing GCA for explicit noise decoupling, LPConv for anisotropic context capture, and CSPStage for occlusion refinement, our framework significantly improves localization accuracy and false-alarm suppression while maintaining a strictly lightweight profile.

3 An Improved Lightweight Multi-Scale Feature Fusion Method for Small Object Detection

3.1 MFCI-YOLO Network Architecture

To mitigate small-object feature loss and background interference in UAV imagery, we propose a lightweight YOLOv11n-based detector. As shown in Fig. 1, our method integrates LPConv for orthogonal feature alignment and geometric detail capture via a dynamic channel strategy. SG-PANet constructs a high-resolution P2 interaction space, while the GCA operator suppresses cross-scale fusion noise, ensuring semantic fidelity for faint objects. It is important to note that MFCI-YOLO refers to our overall end-to-end detection framework, whereas the SG-PANet specifically designates the novel neck architecture (feature fusion network) embedded within it.

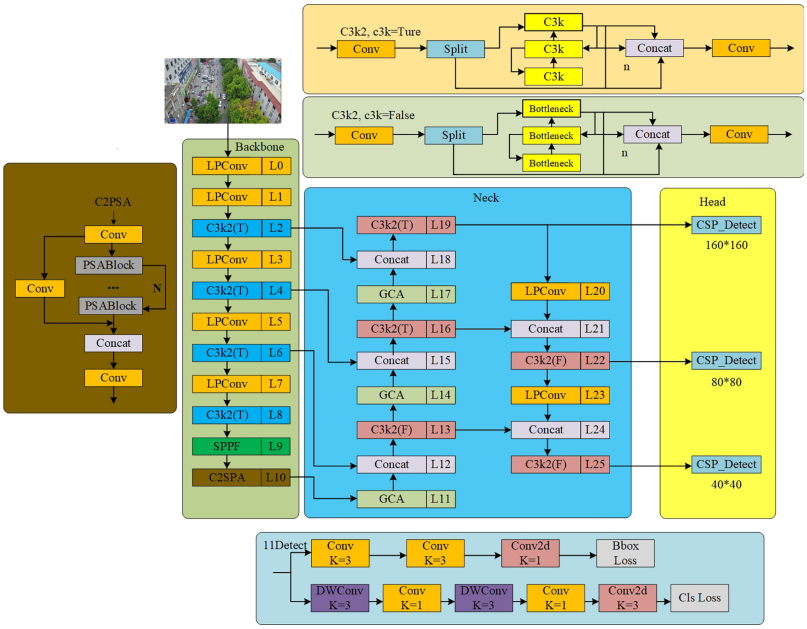

Figure 1: MFCI-YOLO network architecture.

In the refinement and detection stage, CSPStage is deployed before the head to denoise and reconstruct features via full-history state awareness, bridging semantic gaps and enhancing robustness in occlusions. Finally, the Wise-IoU v3 loss function prioritizes high-quality small object instances, significantly boosting performance while maintaining real-time efficiency.

3.2 Lightweight Pinwheel Context Convolution

UAV-detected objects show distinct anisotropic elongated features. Traditional 3 × 3 convolution with square receptive fields introduces excessive background noise when processing such features, so we propose a lightweight pinwheel context convolution (LPConv) based on orthogonal center alignment and dual-stream complementarity.

As shown in Fig. 2, the input feature tensor is defined as: $X \in \mathbb{R}^{C\times H\times W}$. Where X is the input tensor, C is channel count, and H, W are its height and width.

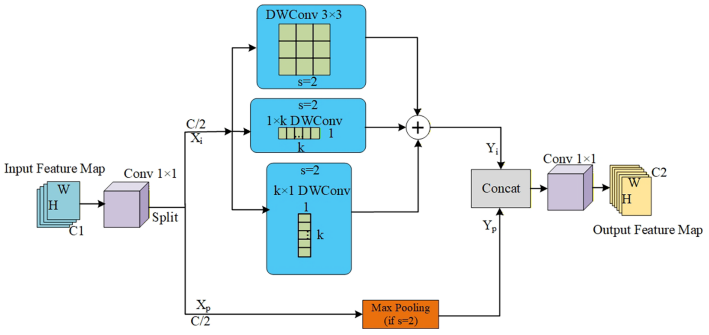

Figure 2: Schematic diagram of LPConv.

To improve detection and avoid channel dimension explosion, we split input features into a processing stream (X_p) for feature enhancement and an identity stream (X_i) for original feature preservation: $X_p = X_{[1:\frac{C}{2},:,:]}, \quad X_i = X_{[\frac{C}{2}+1:C,:,:]}$. Subscripts denote channel slicing, with $X_p, X_i \in \mathbb{R}_2^{C\times H\times W}$.

For X_p, we build a Local Windmill Extraction Unit with three anisotropic convolution branches: 3×3 depthwise separable convolution (core, high-frequency texture), $1 \times k$ (horizontal context), and $k \times 1$ (vertical context). Specifically, the $1 \times k$ and $k \times 1$ orthogonal branches are selected to construct a cross-shaped receptive field that tightly aligns with the anisotropic, slender morphologies of typical aerial objects (e.g., vehicles, pedestrians) from a top-down view, effectively avoiding the diagonal background noise introduced by standard $k \times k$ square filters. Residual additive aggregation unifies these features:

$$Y_p = \mathcal{F}_{\text{core}}(X_p) \oplus \mathcal{F}_{\text{hor}}(X_p) \oplus \mathcal{F}_{\text{ver}}(X_p) \tag{1}$$

where Y_p is the processing stream output, $\mathcal{F}_{\text{core}}/\mathcal{F}_{\text{hor}}/\mathcal{F}_{\text{ver}}$ are the three branches, and $\oplus$ is element-wise addition.

When LPConv is specifically employed as a downsampling module to replace the standard Conv layer, a coordinated spatial reduction strategy is applied to both streams. The initial 1×1 splitting convolution maintains a stride of 1. Subsequently, the parallel depthwise convolutions (the 3×3, $1 \times k$, and $k \times 1$ branches) in the processing stream operate with a stride of 2 ($s = 2$) to halve the spatial dimensions while extracting multi-scale contextual features. Synchronously, a 2×2 Max Pooling operation is applied to the identity stream (X_i). This pooling step is essential to strictly align the spatial dimensions of Y_p and Y_i (both reduced to $H/2 \times W/2$) prior to the final channel concatenation. For standard feature extraction without downsampling, the stride remains $s = 1$ and the Max Pooling operation is bypassed.

Finally, we remap and unify these complementary subspaces by concatenating Y_p and Y_i along the channel dimension. A 1×1 pointwise convolution then aggregates features and adjusts channel depth. The final LPConv output, Y_{out}, is expressed as: $Y_{\text{out}} = \mathcal{F}_{\text{fuse}}(\text{Concat}(Y_p, Y_i))$. Where $\text{Concat}(\cdot)$ is channel-wise splicing and $\mathcal{F}_{\text{fuse}}$ is 1×1 fusion convolution.

In simpler terms, equations mathematically describe a split-and-conquer strategy. By dividing the feature maps, LPConv ensures that one half actively captures the orthogonal shapes of small objects, while the other half preserves the original uncorrupted details, achieving a highly efficient balance between contextual learning and parameter reduction.

Unlike conventional strip convolution methods that simply alternate $1 \times k$ and $k \times 1$ kernels to expand the receptive field at the cost of losing central high-frequency details, LPConv is driven by a dual-stream orthogonal alignment motivation. By explicitly decoupling the feature space into a processing stream for anisotropic contextual capture and an identity stream for fine-grained texture preservation, LPConv fundamentally avoids the detail degradation typical in extremely small objects during deep feature extraction.

For lightweight evaluation, assume input/output channels C, core kernel $k_{core} = 3$, and strip kernel $k_{strip} = k$. The standard 3×3 convolution parameter count is $P_{std} = 9C^2$. Because LPConv performs spatial convolution on only half the channels, its total parameter count is $P_{LPC} = C^2 + \frac{C}{2}(9 + 2k)$. The parameter compression ratio is thus defined as $\eta_{param} = P_{LPC}/P_{std}$. Setting $C = 128$ and $k = 7$, we obtain $\eta_{param} \approx 12.1\%$. This theoretically demonstrates that LPConv expands the effective receptive field to 7×7 while retaining robust feature extraction capabilities using only 12.1% of the parameters of a standard convolution.

3.3 Optimized Upsampling Operator

The CARAFE operator proposed by Wang et al. [24] employs a dynamic kernel generation mechanism, yet its channel sharing approach exhibits limitations in UAV aerial photography scenarios. As background

information dominates the scene, the shared kernel often functions as a smoothing filter, leading to the misidentification of small object edges as noise and their subsequent suppression. This results in the phenomenon of feature drowning.

To address this issue, we propose a group-content-aware (GCA) reconstruction method based on low-level decoupling. As illustrated in Fig. 3, we partition the channel space C into G independent semantic groups ($C_g = C/G$ for each group) and process them through two parallel branches. In our implementation, the number of groups G is empirically set to 4. The choice of G significantly impacts both detection performance and computational efficiency. A smaller G (e.g., $G = 1$) regresses to a channel-shared mechanism, failing to decouple the background from the object and leading to feature drowning. Conversely, an excessively large G forces the network to generate too many independent kernels, which not only drastically increases the memory access cost (MAC) and computational overhead but also disrupts the synergistic feature representation within semantic channel groups. Setting $G = 4$ effectively strikes an optimal balance, ensuring sufficiently fine-grained semantic decoupling for small objects while adhering to strict lightweight deployment constraints.

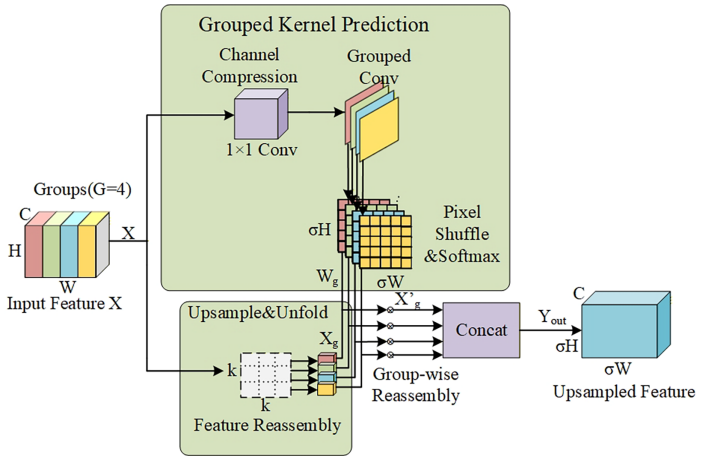

Figure 3: Schematic diagram of GCA.

In the group kernel prediction branch, input features undergo 1×1 convolution compression before being processed through group convolutions (content encoder), which concurrently output G sets of independent parameters. Following Pixel Shuffle and Softmax normalisation, this generates G distinct re-organised kernel tensors $\{W_1, W_2, \ldots, W_G\}$.

Simultaneously, within the grouping and reorganisation branch, the input X is decomposed into G sub-streams $\{X_1, \ldots, X_G\}$. Each sub-stream undergoes grouped dot product reorganisation solely with its corresponding kernel W_g:

$$X'_{l',c,g} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W^{g}_{l'(n,m)} \cdot X^{g}_{(l'+n,j+m),c,g} \tag{2}$$

where $X'_{l',c,g}$ denotes the pixel at position l', channel c, and group g. W_g is the reorganization kernel with radius r, while (n, m) represents the local coordinate offset.

To restore channel capacity and spatial alignment, high-frequency details are reaggregated by concatenating all G reconstructed subflows $\{X'_1, \ldots, X'_G\}$ along the channel dimension. The final output Y_{out} is

expressed as: $Y_{\text{out}} = \text{Concat}(X'_1, X'_2, \ldots, X'_G)$. Where $\text{Concat}(\cdot)$ denotes the concatenation operation along the channel dimension.

The recombined sub-streams are reshaped from $B \times G \times (C/G) \times oH \times oW$ back to the standard $B \times C \times H \times W$ format. Unlike CARAFE's shared smoothing kernels, GCA assigns dedicated high-frequency kernels to independent semantic groups. This prevents small-object textures from being suppressed by dominant backgrounds, effectively alleviating feature drowning and enhancing discriminative details for detection.

Intuitively, a globally shared kernel acts as a single "broad brush" that inevitably blurs fine object details into the dominant background. In contrast, our grouped GCA mechanism assigns dedicated "fine-tipped brushes" exclusively for the high-frequency textures of small objects, explicitly decoupling them from noise and completely avoiding the feature drowning effect.

3.4 CSPStage Module

To address the high-frequency feature decay of small objects in deep convolutional networks, we draw inspiration from the designs of GiraffeDet [25] and CSPNet [26] to propose the Cross-Stage Dense Feature Refinement module (CSPStage). This module enhances the model's representational capability for low-pixel-count objects through full-spectrum feature reuse and structural reparameterization.

As shown in Fig. 4, CSPStage employs a dual-stream feature evolution system. The input tensor X is projected onto the baseline subspace Y_1 and the evolutionary subspace $Y_2^{(0)}$.

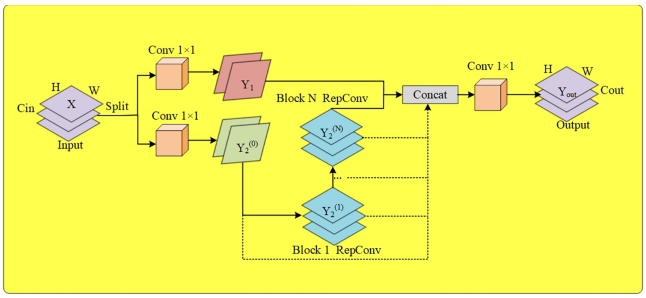

Figure 4: Schematic diagram of CSPStage architecture.

In the baseline branch, 1×1 convolutions construct truncated gradient highways, enabling raw high-resolution spatial features to bypass deep nonlinear transformations and reach the fusion layer directly, which ensures precise pixel-level localization of small objects.

Within the evolutionary branch, the dense aggregation mechanism integrates the outputs $Y_2^{(i)}$ from n cascaded units. This unit employs a RepConv operator, which leverages a multi-branch topology during training to capture multi-scale features. During inference, it fuses equivalently into a single kernel, enhancing extraction robustness without increasing.

By fusing shallow spatial and deep semantic features, Y_{out} mitigates small-object scarcity. Positioning CSPStage before the detection head aligns task features and bridges semantic gaps. Under dense occlusions, the shallow branch bypasses deep non-linearities to preserve high-resolution textures of unoccluded fragments, while the deep branch provides robust semantic context. Aggregating these streams enables MFCI-YOLO to infer partially visible objects from surviving local textures and global priors, significantly reducing missed detections.

3.5 *Wise-IoU Loss Function*

To improve localization robustness against low-quality annotations, we employ Wise-IoU v3. It first constructs a geometric distance penalty $R_{WIoU} = \exp\left(\frac{\rho^2}{(W_g^2+H_g^2)^*}\right)$, where ρ is the Euclidean distance between predicted and ground-truth centers, and W_g, H_g are the minimum bounding box dimensions illustrated in Fig. 5. The asterisk (*) denotes detachment from the computation graph, ensuring it acts purely as an attention weight without hindering small-object convergence, yielding the base loss $L_{WIoU\_v1} = R_{WIoU} \cdot L_{IoU}$.

To optimize gradient allocation across inconsistent annotation qualities, we introduce an outlier degree $\beta = L_{WIoU\_v1}/\overline{L_{IoU}}$ (with $\overline{L_{IoU}}$ as the batch's moving average loss) and a dynamic focusing gain $r = \frac{\beta}{\delta\alpha^{\beta-\delta}}$, producing the final loss $L_{WIoU\_v3} = r \cdot L_{WIoU\_v1}$. As depicted by the non-monotonic curve in Fig. 6, this mechanism acts as an intelligent dynamic filter. Calculus analysis indicates r peaks at $\beta = 1/\ln\alpha$, effectively down-weighting both "easy" high-quality samples ($\beta \ll 1$) to prevent overfitting, and "hard" extreme outliers ($\beta \gg 1$) to avoid harmful gradient updates from noisy labels. Consequently, the network optimally prioritizes ordinary-quality, challenging small objects.

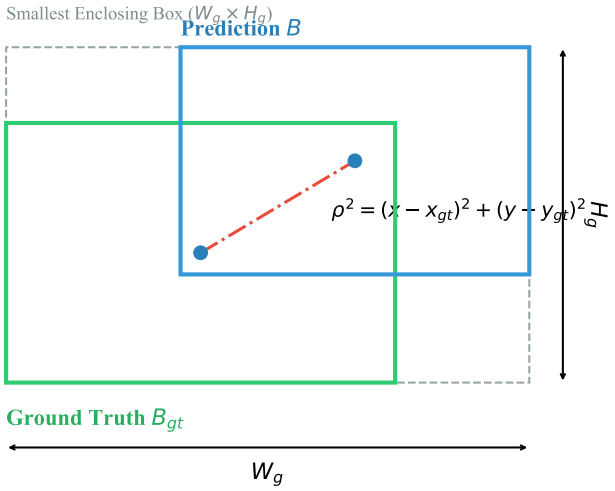

Figure 5: Illustration of geometric definition.

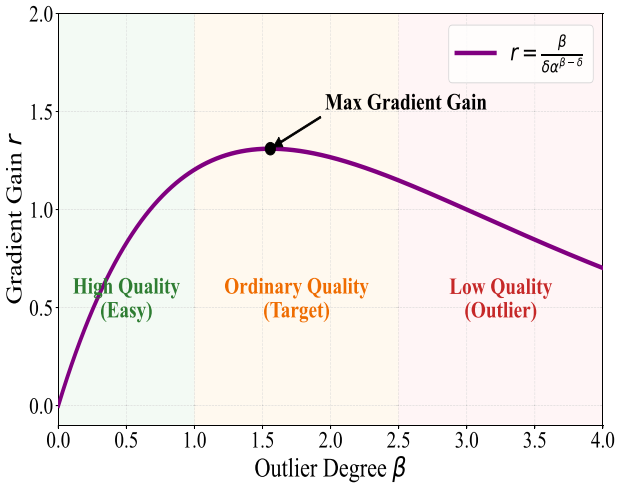

Figure 6: Illustration of dynamic mechanism dataset.

3.6 *Multi-Scale Convergence Network*

Although YOLOv11n [27] native path aggregation network enhances feature flow, its P3–P5 architecture causes scale mismatch and fusion loss in UAV aerial images: 32x downsampling loses small object information, and simple interpolation or convolution fails to distinguish objects from backgrounds, drowning small object features and amplifying background noise. To solve this, we propose Scale-Gradient and Context-Aware Feature Fusion Network (SG-PANet). As shown in Fig. 7, SG-PANet adopts a scale-down strategy: discarding P5 and extending fusion paths to shallow layers, constructing a P2-P3-P4 high-resolution fusion hierarchy. This preserves geometric textures and resolves small object information diffusion.

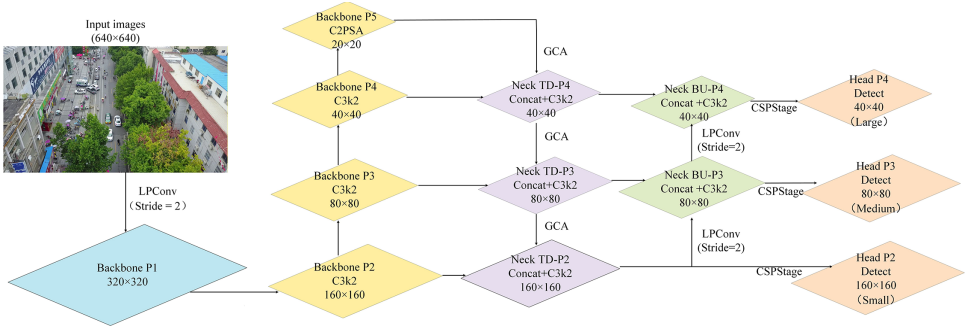

Figure 7: Schematic diagram of SG_PANet architecture.

We introduce GCA as the fusion operator to address feature drowning. Unlike traditional interpolation (blending background and object features), GCA uses group decoupling to partition channels, dynamically enhancing object-related channels, thus injecting deep semantics into shallow features while preserving object details. Meanwhile, we replace standard convolutions with LPConv in the bottom-up localization enhancement path to reduce computational redundancy and texture noise in high-resolution features. LPConv's orthogonal alignment uses strip convolution branches to capture small-object morphology and additive fusion to preserve high-frequency details, reducing computation and resolving localization deviations. In summary, SG-PANet achieves performance improvement via physical-scale reconstruction and bidirectional path optimization.

4 Experimental Process and Analysis

4.1 Datasets

VisDrone2019 [28] (Fig. 8) by Tianjin University's AISKYEYE team focuses on UAV-based computer vision tasks, comprising 10,209 static images (6471 training, 548 validation, 1610 testing) across 14 Chinese cities' urban/rural scenes, with 10 core categories.

To validate MFCI-YOLO's robustness for aerial small object detection, we adopted DOTA v1.5 [29] (Fig. 9)—an upgraded DOTA v1.0 retaining 2806 images, with finer annotations and abundant <10-pixel small instances.

Figs. 8 and 9 visually confirm that most UAV instances occupy sub-10-pixel ratios with severe class imbalance. This extreme distribution empirically justifies SG-PANet for preserving high-resolution details, and Wise-IoU v3 for dynamically balancing the gradients of small, low-quality samples.

The ablation and comparison experiments utilize the VisDrone2019 dataset (Fig. 8) to validate the model's detection performance, while the model's generalization capability is evaluated using the DOTA v1.5 dataset (Fig. 9).

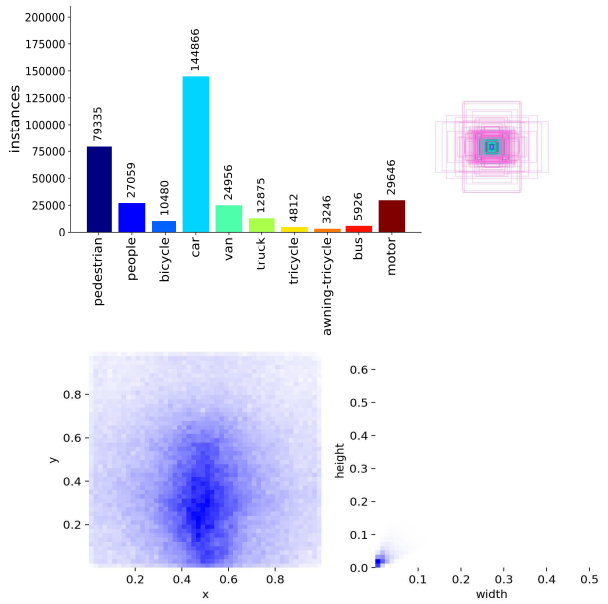

Figure 8: Label analysis diagram of the VisDrone2019 dataset.

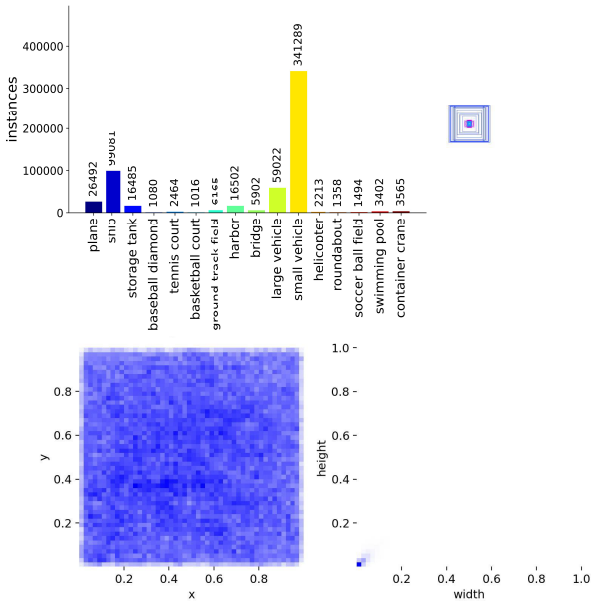

Figure 9: Label analysis diagram of the DOTAv1.5 dataset.

4.2 Experimental Environment and Evaluation Indicators

The experiments were conducted using an NVIDIA RTX 5090 GPU (32 GB VRAM) with CUDA 12.8 and PyTorch 2.8.0. The model was trained for up to 200 epochs (with early stopping upon convergence) using the SGD optimizer (momentum: 0.937, weight decay: 0.0005) and a Cosine Annealing scheduler with an initial learning rate of 0.01. Furthermore, to ensure statistical stability and reproducibility, all reported evaluation metrics represent the average results from three independent training runs under identical hyperparameter settings.

We comprehensively evaluate the model's detection performance using standard accuracy metrics: Precision (P), Recall (R), mAP50, and mAP50-95. Furthermore, to strictly assess the model's deployability on resource-constrained UAV edge devices, we measure computational efficiency using total parameters (M), computational complexity (GFLOPs), and real-time inference speed (FPS).

4.3 Ablation Experiments

To systematically validate the effectiveness of each component within the MFCI-YOLO framework and their respective contributions to small object detection performance in UAV aerial photography, we designed a layer-by-layer ablation study on the VisDrone2019 dataset. Using YOLOv11n as the baseline model, we progressively integrated the Wise-IoU v3, SG-PANet, LPConv, GCA, and CSPStage modules. We quantitatively analyzed the trade-offs among accuracy (mAP), number of parameters, and FPS for each module. The quantitative results are shown in Table 1.

Table 1: Ablation experiment.

YOLOv11n	WIoU v3	SG-PANet	LPConv	GCA	CSPstage	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	Params (M)	GFLOPS	FPS
✓						32.1	18.7	42.3	32.6	2.58	6.3	194
✓	✓					32.4	18.8	42.6	33.4	2.58	6.3	194
✓	✓	✓				47.3	29.1	56.8	45.9	4.79	42.4	168
✓	✓	✓	✓			47.0	28.9	56.7	44.2	3.38	38.5	136
✓	✓	✓	✓	✓		47.7	29.3	57.3	45.4	5.52	45.9	114
✓	✓	✓	✓	✓	✓	**48.2**	**29.8**	**55.7**	**45.0**	**4.50**	**45.0**	97

Note: Bold formatting indicates the optimal performance values.

Table 1 demonstrates that the baseline YOLOv11n struggles with deep downsampling-induced feature diffusion, yielding only 32.1% mAP50. Integrating SG-PANet drastically boosts mAP50 to 47.3% by preserving high-resolution details. Replacing standard convolutions with LPConv effectively cuts parameters to 3.38M and GFLOPs to 38.5 with negligible accuracy drop, as its anisotropic receptive field perfectly matches elongated aerial objects. Furthermore, introducing GCA suppresses background noise, and CSPStage mitigates deep feature decay via historical state awareness. Ultimately, MFCI-YOLO achieves an optimal 48.2% mAP50 and 29.8% mAP50-95 at a real-time 97 FPS, significantly outperforming the baseline across all extremely small object categories (e.g., Pedestrian and Bicycle AP improved by 21.3% and 14.1%, respectively).

The training process visualized in Fig. 10 shows that MFCI-YOLO achieves a higher growth rate and superior final steady-state values for both mAP50 and mAP50-95 compared to YOLOv11n. Regarding loss functions, MFCI-YOLO demonstrates faster convergence and more stable final values in classification (Cls Loss) and distribution focal loss (DFL Loss). This confirms that the synergy of SG-PANet and Wise-IoU v3 effectively optimizes gradient propagation and enhances learning efficiency for small objects.

Fig. 11's visualization offers intuitive proof. The baseline has many blue missed detections in dense crowds and distant vehicles, while MFCI-YOLO corrects these to green true detections. In summary, MFCI-YOLO optimally balances accuracy and robustness for UAV aerial scenarios under limited computation.

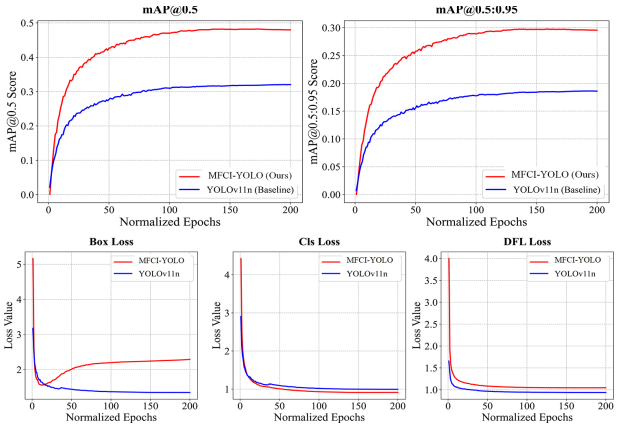

Figure 10: Visualization of the performance comparison between the basic model and the improved model.

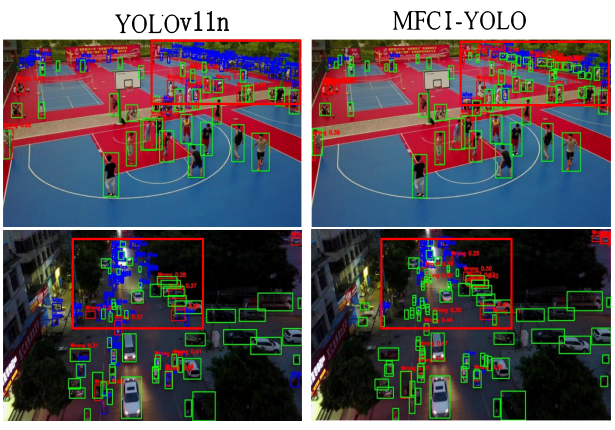

Figure 11: Comparison of the detection diagram of the basic model and the improved model. The left base model, the right improvement model. The green box is correctly detected, the blue box is not detected, and the red box is falsely detected.

4.4 Comparison Experiments

4.4.1 Convolutional Comparison Experiment

Table 2 shows LPConv has notable lightweight advantages. Compared to baseline standard convolution (Conv), LPConv reduces model parameters from 4.79 million to 3.38 million and GFLOPs from 42.4 to 38.5, with only a 0.3% mAP50 drop and nearly unchanged mAP50-95. We specifically selected pConv (Partial Convolution) as a baseline because it represents a state-of-the-art generic lightweight operator focused on spatial redundancy reduction. This comparison critically highlights that for UAV imagery, the domain-specific anisotropic alignment of our LPConv is superior to generic redundancy-reduction approaches. Compared with PConv, LPConv cuts parameters by an extra 18% while maintaining higher accuracy, validating its orthogonal center alignment mechanism.

Table 2: Convolutional comparison table (WIou+SG_PANet).

Category	Conv	PConv	LPConv
P	0.568	0.557	0.567
R	0.459	0.453	0.442
mAP50	0.473	0.473	0.470
mAP50-95	0.291	0.291	0.289
Params (M)	4.79	4.14	3.38
GFLOPS	42.4	45.3	38.5

Table 3 presents ablation results for core hyperparameter k. Accuracy rises steadily as k increases from 3 to 7 (expanded receptive field supplements contextual semantics), but declines when k reaches 9 or 11 (excessively large kernels introduce background noise, diluting features in sparse aerial objects). Thus, $k = 7$ is selected as optimal, balancing lightweight design and feature representation capability.

Table 3: Comparison table of convolution lengths of different bars of LPConv (WIoU).

k	mAP50	mAP50-95	P	R	$P/10^6$	GFLOPs
3	0.319	0.183	0.431	0.322	2.113	6.0
5	0.319	0.183	0.420	0.328	2.115	6.0
7	**0.322**	**0.184**	**0.424**	**0.328**	**2.116**	**6.0**
9	0.319	0.184	0.434	0.318	2.118	6.0
11	0.316	0.181	0.424	0.316	2.119	6.0

Note: Bold formatting indicates the optimal performance values.

4.4.2 Comparison Experiments with Different Models

To evaluate MFCI-YOLO's competitiveness in UAV aerial scenarios, we conducted comparative analyses on VisDrone2019, comparing it with classic algorithms (RetinaNet, Faster R-CNN), general benchmarks (YOLOv5-YOLOv11 series), and aerial-optimized advanced algorithms (HPRS-YOLO, DI-YOLO); detailed results are in Table 4.

Experimental results confirm MFCI-YOLO's architectural superiority, abandoning performance enhancement via pure parameter stacking. vs. baseline YOLOv11n, MFCI-YOLO boosts mAP50 to 48.2% while remaining lightweight via SG-PANet/GCA-based feature alignment. Notably, vs. higher-parameter

YOLOv11s, MFCI-YOLO achieves a 10.4% accuracy gain with <50% parameters, validating that small-object feature refinement outperforms blind network depth scaling.

Table 4: Comparison table of models under VisDrone2019 validation set.

Model	mAP50%	mAP50-95%	P/%	Parameters/M	Gflops	FPS
RetinaNet	22.1	16.6	41.3	19.8	93.7	41.3
Faster R-CNN	33.5	19.3	45.7	41.2	206.6	24
YOLOv5s	37.9	22.7	49.1	9.11	23.7	133
YOLOv7-Tiny	35.4	18.9	45.9	6.33	13.3	121
YOLOv8n	31.9	18.4	43.2	3.00	8.1	232
YOLOv8s	39.1	23.3	50.2	11.2	28.8	118
YOLOv10n	31.7	18.4	43.3	2.71	8.4	241
YOLOv11n	32.1	18.7	42.3	2.58	6.3	250
YOLOv11s	37.8	22.5	48.0	9.4	21.3	–
HPRS-YOLO [18]	38.4	22.7	49.9	3.30	9.3	212
SBE-YOLOv8s [19]	42.1	24.3	51.7	6.2	105.4	28
DI-YOLO [20]	47.1	29.0	–	26.14	96.3	113.4
CM-YOLOv8s [21]	45.9	27.8	55.4	3.49	31.2	72
MPAM-YOLO [22]	43.1	26.9	45.5	41.5	46.5	52.4
SE-MSPA-DETR [23]	40.1	23.2	–	16.03	65.4	78.5
Ours	**48.2**	**29.8**	**55.7**	**4.50**	**45.0**	**97**

Note: Bold formatting indicates the optimal performance values.

MFCI-YOLO demonstrates superior performance over state-of-the-art algorithms. Compared to DI-YOLO, it achieves 48.2% mAP50 with only 1/6 the parameters and 1/2 the FLOPs. It significantly outperforms MPAM-YOLO despite having 9.2 times fewer parameters. Furthermore, unlike the resource-heavy Transformer-based hybrid framework SE-MSPA-DETR, MFCI-YOLO maintains a high-speed 97 FPS while achieving an 8.1% higher mAP50. Overall, MFCI-YOLO provides an optimal balance of accuracy and efficiency for resource-constrained UAV deployment.

To intuitively verify the effectiveness of the proposed modules in enhancing contextual information and suppressing background noise, we generated feature activation heatmaps for both the baseline model and MFCI-YOLO, as shown in Fig. 12. The visualization captures a highly challenging night-time plaza scenario with dense, extremely small pedestrian objects.

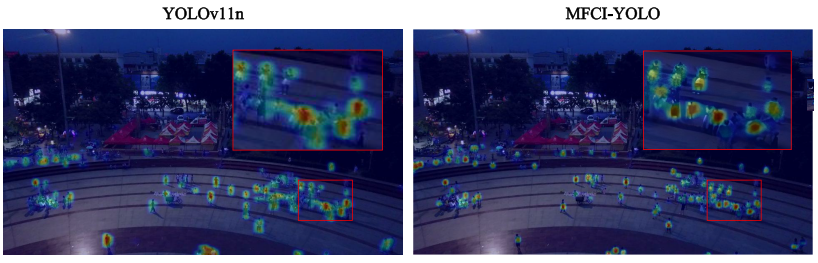

Figure 12: Heatmap visualization comparison in a dense and low-light scenario.

As detailed in the red zoom-in boxes, the baseline heatmap (left) exhibits diffuse activation. Its feature responses for distant small objects blur into the dark background, causing severe "feature drowning". Conversely, the MFCI-YOLO heatmap (right) displays highly focused red spots strictly centered on the objects. The magnified view confirms that our model accurately activates individual tiny targets without falsely highlighting the surrounding pavement. This visual comparison verifies that LPConv's anisotropic receptive field effectively captures object-specific contexts, while the GCA operator cleanly decouples fine-grained features from complex background noise, successfully preserving the signals of extremely small objects.

4.5 Generalization Experiments

To validate robustness, we evaluated the model on the challenging DOTA v1.5 dataset. As shown in Table 5, our method achieves an outstanding 80.3% mAP50.

Table 5: Generalization experiments of different models on the DOTAv1.5 dataset.

Model	mAP50 (%)	Para (M)	GFLOPs
YOLOv5n	62.5	1.8	4.2
YOLOv8n	65.6	3.0	8.1
YOLOv11n	75.3	2.6	6.3
BRSTD [30]	65.4	1.8	64.3
AG-YOLO [31]	72.8	16.6	–
AESOD [32]	48.9	2.7	27.4
ours	80.3	4.5	45.0

DOTA v1.5 presents extreme challenges: arbitrary orientations, massive scale variations, and dense sub-10-pixel clusters. MFCI-YOLO directly addresses these: LPConv's orthogonal branches capture slender geometries, while GCA's independent high-frequency kernels suppress severe background interference, effectively preventing dense objects from drowning.

Fig. 13 corroborates this analysis. Even under extreme lighting and occlusion, MFCI-YOLO achieves 0.894 precision for swimming pools and minimizes missed detections in dense vehicle areas. This confirms that CSPStage effectively bridges the semantic gap via historical state awareness. Overall, MFCI-YOLO demonstrates exceptional robustness and practical value for complex aerial deployment.

Figure 13: Visualization of DOTAv1.5 detection.

To highlight our network's advantages, Table 5 compares MFCI-YOLO against mainstream baselines and recent state-of-the-art detectors on the DOTAv1.5 dataset. Earlier models like YOLOv5n and YOLOv8n

achieve only 62.5% and 65.6% mAP50. Recent advanced networks like BRSTD, AG-YOLO, and AESOD also struggle with dense, arbitrary-oriented targets, scoring 65.4%, 72.8%, and 48.9%, respectively. While the baseline YOLOv11n reaches 75.3%, it remains limited by feature drowning in complex scenes. In contrast, MFCI-YOLO achieves an exceptional 80.3% mAP50. Notably, it outperforms AG-YOLO by 7.5% using nearly one-fourth of its parameters. Although this precision involves a parameter increase (from 2.6M to 4.5M), our model strictly satisfies the 5M lightweight constraint. This trade-off confirms its superior accuracy, robust generalization, and essential lightweight profile for complex aerial imagery.

Although MFCI-YOLO exhibits exceptional generalization in dense scenarios, it still occasionally encounters missed detections and false positives under extreme physical conditions. Specifically, during severe motion blur caused by high-speed UAV maneuvering, the structural contours of extremely small objects become severely distorted, causing the anisotropic receptive field of LPConv to fail in feature alignment. Additionally, under heavy nighttime glare or extreme low-light conditions, background noise and object textures become virtually indistinguishable, occasionally bypassing the GCA operator's decoupling mechanism. Addressing these extreme multi-degradation scenarios via multi-modal fusion or semi-supervised learning remains an important avenue for our future research.

5 Conclusion

We proposed MFCI-YOLO, a lightweight UAV small-object detector, to address feature drowning and missed detections in complex backgrounds. Functionally, LPConv captures slender object morphologies with minimal computational redundancy, while the GCA-equipped SG-PANet decouples object signals from noise and preserves high-resolution geometric details. Integrated with CSPStage and Wise-IoU v3, the model successfully bridges semantic gaps and optimizes dynamic gradient allocation. Experimental results confirm its superiority: achieving 48.2% mAP50 at 97 FPS (only 4.50M parameters) on VisDrone2019, and 80.3% mAP50 on DOTAv1.5. Despite its exceptional generalization, future research will focus on integrating multi-modal data (e.g., infrared imaging) and semi-supervised learning to mitigate occasional missed detections under extreme motion blur or severe low-light conditions.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the Natural Science Foundation of Henan Province under Grant 252300423317, the Science and Technology Research Project of Henan Province under Grant 262102211081, the Key Scientific Research Projects of Colleges and Universities in Henan Province under Grant 25B510012 and "Pioneer" and "Leading Goose" R&D Program of Zhejiang under grant 2026LDC01003(JT).

Author Contributions: The authors confirm their respective contributions to the paper as follows: Weiguang Wang and Jincai Li: Conceptualization, Methodology, Software, Writing—Original Draft. Mengqi Liu: Validation, Data Curation, Investigation. Mengke Liu and Yuan Zhang: Formal Analysis, Visualization. Jingyan Wu and Yang Liu: Supervision, Project Administration, Funding Acquisition, Writing—Review & Editing. Junbin Lou and Yixin He: Resources, Validation. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The VisDrone2019 and DOTAv1.5 datasets analyzed in this study are publicly available. The code and models that support the findings of this study are available from the corresponding authors upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He Y, Huang F, Wang D, Chen B, Zhang R. Emergency communications in post-disaster scenarios: IoT-enhanced airship and buffer support. IEEE Internet Things J. 2025;12(9):11457–68.
2. He Y, Huang F, Wang D, Zhou X, Zhang R. Uplink outage probability analysis of AAV and intelligent connected vehicle cooperative communication using full-duplex MIMO. IEEE Commun Lett. 2025;29(9):2068–72. doi:10.1109/lcomm.2025.3585337.
3. He Y, Huang F, Wang D, Yang L, Zhang R. Delay minimization for NOMA-MEC offloading in ABS-aided maritime communication networks. IEEE Trans Veh Technol. 2025;74(6):9577–90. doi:10.1109/tvt.2025.3539335.
4. Wang D, Wang Z, Yang W, Zhao H, He Y, Li L, et al. Enhanced ISAC framework for moving target assisted by beyond-diagonal RIS: accurate localization and efficient communication. IEEE Trans Netw Sci Eng. 2025;12(5):4299–315. doi:10.1109/tnse.2025.3571278.
5. Li Y, Li Q, Pan J, Zhou Y, Zhu H, Wei H, et al. SOD-YOLO: small-object-detection algorithm based on improved YOLOv8 for UAV images. Remote Sens. 2024;16(16):3057. doi:10.3390/rs16163057.
6. Wan Z, Lan Y, Xu Z, Shang K, Zhang F. DAU-YOLO: a lightweight and effective method for small object detection in UAV images. Remote Sens. 2025;17(10):1768. doi:10.3390/rs17101768.
7. Yu W, Mo K. YOLO-GCOF: a lightweight low-altitude drone detection model. IEEE Access. 2025;13(10):53053–64. doi:10.1109/access.2025.3553477.
8. Jian J, Liu L, Zhang Y, Xu K, Yang J. Optical remote sensing ship recognition and classification based on improved YOLOv5. Remote Sens. 2023;15(17):4319. doi:10.20944/preprints202307.0150.v1.
9. Ma Y, Wang H. Improved multiscale feature fusion and small object detection layer optimization for UAV aerial small object detection based on YOLOv8. In: Proceedings of the International Conference on Advances in Computer Vision Research and Applications (ACVRA 2025); 2025 Feb 28–Mar 2; Nanjing, China. p. 213–7.
10. Lai D, Kang K, Xu K, Ma X, Zhang Y, Huang F, et al. Enhancing UAV object detection with an efficient multi-scale feature fusion framework. PLoS One. 2025;20(10):e0332408. doi:10.1371/journal.pone.0332408.
11. Qiao S, Chen LC, Yuille A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 10208–19.
12. Sun H, Wang R, Li Y, Yang L, Lin S, Cao X, et al. SET: spectral enhancement for tiny object detection. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 4713–23.
13. Ye D, Li G. A Small object detection model based on improved YOLO. J Agric Big Data. 2025;7(2):173–82. (In Chinese). doi:10.19788/j.issn.2096-6369.000073.
14. Ju Z, Shui J, Huang J. GLDS-YOLO: an improved lightweight model for small object detection in UAV aerial imagery. Electronics. 2025;14(19):3831. doi:10.3390/electronics14193831.
15. Zhao B, Zhao J, Song R, Yu L, Zhang X, Liu J. Enhanced YOLO11 for lightweight and accurate drone-based maritime search and rescue object detection. PLoS One. 2025;20(7):e0321920. doi:10.1371/journal.pone.0321920.
16. Ji CL, Yu T, Gao P, Wang F, Yuan RY. Yolo-tla: an efficient and lightweight small object detection model based on YOLOv5. J Real Time Image Process. 2024;21(4):141. doi:10.1007/s11554-024-01519-4.
17. Zhu L, Wang X, Ke Z, Zhang W, Lau R. BiFormer: vision transformer with bi-level routing attention. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 10323–33.
18. Yang Y, Jiang W, Gao Z. Real-time target detection algorithm for low altitude UAVs. Acta Aeronautica et Astronautica Sinica. 2025;46(16):210–23. (In Chinese). doi:10.7527/S1000-6893.2025.31619.
19. Feng Y, Guo X, Yan J. Small UVA target detection algorithm based on multi-scale attention mechanism. Acta Armamentarii. 2025;46(1):12–21. (In Chinese). doi:10.12382/bgxb.2023.1124.
20. Ding H, He W, Wan J, Shen YH, Cui XH. DI-YOLO: an efficient small object detection framework for UAV aerial imagery. Control Decision. 2025;40(10):3106–16. (In Chinese). doi:10.13195/j.kzyjc.2025.0425.

21. Liao NS, Cao TX, Liu KY, Xu M, Zhu M, Gu YX, et al. Small target detection algorithm for UAV based on composite feature and multi-scale fusion. Comput Eng Appl. 2025;61(3):111–20. (In Chinese). doi:10.3778/j.issn.1002-8331.2407-0520.
22. Fan W, Xu X, Jiang Z, Zhu Z. A multi-stage path aggregation module for small object detection on drone-captured scenarios. Digit Signal Process. 2026;173:105901. doi:10.1016/j.dsp.2026.105901.
23. Li H, Liu H. A spatially enhanced multiscale polarity sensing framework for UAV small target detection. Appl Soft Comput. 2026;186(20):114248. doi:10.1016/j.asoc.2025.114248.
24. Wang J, Chen K, Xu R, Liu Z, Loy CC, Lin D. CARAFE: content-aware reassembly of features. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 3007–16.
25. Jiang Y, Tan Z, Wang J, Sun X, Lin M, Li H. GiraffeDet: a heavy-neck paradigm for object detection. arXiv:2202.04256. 2022.
26. Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW, Yeh IH. CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 1571–80.
27. Zhang X, Ding Q, Dou Y, Ren S. Accurately detecting dense small objects from aerial images by fusing local enhanced features and balanced category loss. In: Proceedings of the 2024 36th Chinese Control and Decision Conference (CCDC); 2024 May 25–27; Xi'an, China. p. 5548–53.
28. Zhu P, Wen L, Bian X, Ling H, Hu Q. Vision meets drones: a challenge. arXiv:1804.07437. 2018.
29. Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3974–83.
30. Huang S, Lin C, Jiang X, Qu Z. BRSTD: bio-inspired remote sensing tiny object detection. IEEE Trans Geosci Remote Sens. 2024;62:1–15.
31. Wang X, Han C, Huang L, Nie T, Liu X, Liu H, et al. AG-yolo: attention-guided yolo for efficient remote sensing oriented object detection. Remote Sens. 2025;17(6):1027.
32. Peng J, Lv K, Lin D, Yuan L. AESOD: towards accurate and efficient general-purpose small object detection. Digit Signal Process. 2026;176(4):106037. doi:10.1016/j.dsp.2026.106037.