

Aerial Small Object Detection with Boundary-Aware and Semantic-Guided Feature Pyramid

Xiaoyu Xiong

¹School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Xiaoyu Xiong (e-mail: U202340049@xs.ustb.edu.cn)

ABSTRACT Aerial small object detection in UAV imagery faces multiple challenges, including highly imbalanced scale distribution, low object resolution, complex background textures, and constrained computational resources. To address these issues, this paper proposes Aerial-YOLO, an aerial object detection algorithm that integrates boundary-aware enhancement and a semantic-guided feature pyramid. First, a boundary-aware and scale-adaptive collaborative enhancement module is constructed in the backbone network. By introducing directional structural reinforcement and adaptive multi-scale convolution mechanisms, the proposed module strengthens object contours and scale-discriminative features, thereby improving localization stability in dense scenes. Second, a lightweight semantic-guided feature pyramid network is designed, whose core component is an Adaptive Cross-scale Injection Gate module. Through unidirectional semantic guidance and sparse fusion strategies, high-level semantic information selectively enhances low-level fine-grained details, improving multi-scale representation capability while reducing computational overhead. In addition, a structural detail reconstruction module is embedded between the backbone and the neck, where directional enhancement and frequency-domain reweighting mechanisms are employed to amplify high-frequency responses of small objects and alleviate interference from complex backgrounds. Experimental results on the VisDrone2019 dataset demonstrate that Aerial-YOLO significantly improves small object detection performance while maintaining real-time inference capability. Further evaluations on the TinyPerson dataset verify the strong generalization ability and robustness of the proposed method.

INDEX TERMS Unmanned aerial vehicle imagery, small object detection, boundary-aware enhancement, semantic-guided feature pyramid, lightweight feature pyramid

I. INTRODUCTION

UAV aerial imagery has been widely used in environmental monitoring, urban management, disaster assessment, intelligent transportation, and other fields because of its high flexibility, broad coverage, and low acquisition cost [1]. Compared with conventional ground-view images, aerial images provide a distinctive perspective, but they also introduce special challenges for small object detection. Objects are usually small, sparsely distributed, and easily affected by background interference and occlusion. Meanwhile, small objects in images have low resolution and large scale variation, which imposes higher requirements on localization accuracy and model robustness [2], [3]. In addition, real-time processing usually requires models to operate under limited computational resources, further increasing the difficulty of model design.

In recent years, deep-learning-based object detection methods have made substantial progress. Convolutional neural networks have become the basis of high-performance

detectors, with representative methods including two-stage detectors such as Faster R-CNN, one-stage detectors such as SSD, and the YOLO series [4], [5]. For UAV aerial small object detection, existing studies have attempted to improve performance through multi-scale feature extraction, attention mechanisms, and lightweight network structures [6], [7]. As UAV platforms are increasingly used in inspection and monitoring, disaster assessment, military reconnaissance, and related scenarios, object detection for UAV scenes has received continuous attention. Compared with general natural scenes, UAV aerial images usually have a top-down view, large scale span, high proportion of small objects, and complex backgrounds. At the same time, onboard computing resources are limited, making the balance between accuracy and complexity especially important.

Domestic and international studies have explored this problem from two major perspectives: lightweight model design and feature modeling optimization. In feature mod-

eling, existing methods mainly focus on multi-scale feature fusion and small object feature enhancement. The former integrates features from different levels through cross-scale information interaction, while the latter improves the discriminability of small objects by enhancing fine-grained features. Although these strategies improve detection performance to some extent, they still have limitations in explicit boundary-direction modeling, selective guidance from high-level semantics to low-level details, and recovery of high-frequency structural information.

In lightweight backbone design, Zhou et al. proposed Drone-YOLO, which introduces Ghost modules to construct CSPGhostNet, reduces redundant computation through cross-stage feature generation, combines a hybrid attention mechanism to improve multi-scale feature representation, and designs a shared small-object detection head and WIoU loss to jointly optimize detection accuracy and model size [8]. Zhang et al. addressed the high computational cost of RT-DETR by constructing a LOB-Vovnet backbone and designing an LRFF feature enhancement module and a cascaded group attention mechanism, improving recognition capability in complex backgrounds while maintaining lightweight properties [9]. Wang et al. designed a lightweight feature extraction module based on YOLOv5s, embedded SENet, and combined it with an adaptive multi-scale feature fusion network named NASFF, effectively reducing parameters and FLOPs while improving deployment suitability in resource-constrained environments [10].

For improvements to the YOLO series, Liao et al. proposed YOLOv10n-CIG based on YOLOv10n. They improved the C2f module and SPPF structure, removed part of the downsampling operations to preserve fine-grained spatial information of small objects, and designed a group-convolution-based lightweight detection head for embedded real-time performance [11]. Chang et al. proposed CSP-YOLO, introducing a space-to-depth convolution module and an improved attention mechanism, reconstructing the neck structure, adding a micro detection head for dense small object detection, and using WIoU loss to address sample imbalance [12]. Jing et al. built MDPE and DAFI modules based on YOLOv11, strengthening feature allocation and fusion through dynamic channel splitting and dual dynamic weighting, and designed ARSIOU loss to improve small object localization accuracy, achieving performance gains on VisDrone2019 [13].

Although many studies have been conducted on UAV aerial small object detection and progress has been made in lightweight modeling, multi-scale fusion, and loss-function optimization, existing methods still show shortcomings in key feature modeling when considering the practical challenges of low object resolution, highly imbalanced scale distribution, and complex background texture. On the one hand, because small objects occupy very few pixels, their boundary and structural information are inherently weak. Repeated downsampling and convolution operations further weaken fine-grained contours and directional structural in-

formation in shallow features. Most existing methods focus on channel recalibration or global context modeling, while explicit modeling of boundary-sensitive features remains insufficient. As a result, contour information may decay early during feature extraction, affecting subsequent localization accuracy. On the other hand, for large scale variation and strong background interference, existing studies usually improve performance through multi-scale feature fusion or attention mechanisms, but most focus on individual modules or local path optimization. They lack systematic modeling of the cross-level feature evolution process of shallow-structure preservation, high-level semantic guidance, and detail information recovery. Therefore, under lightweight constraints, building a collaborative feature modeling framework that considers boundary information preservation, semantic-guided enhancement, and structural detail recovery is important for improving UAV aerial small object detection.

To address the progressive degradation of fine-grained structural information during cross-level feature propagation, this paper proposes a unified feature enhancement framework named Aerial-YOLO from the perspective of structural information preservation and cross-scale semantic collaborative modeling. Around the feature evolution process of boundary information preservation, semantic-guided enhancement, and structural detail recovery, the main contributions are as follows:

1. A boundary-aware and scale-adaptive collaborative enhancement module is proposed. To address the attenuation of small-object boundary information in shallow features during downsampling, directional structural reinforcement and adaptive multi-scale convolution are introduced to enhance fine-grained contour response and scale-discriminative capability, suppressing structural information loss from the source.
2. A lightweight semantic-guided feature pyramid is designed. To address the problem that high-level semantics may overwhelm low-level details during cross-level feature fusion, an adaptive cross-scale injection gate is constructed to selectively guide structural details with semantic information, improving the consistency and effectiveness of multi-scale feature representation.
3. A structural detail reconstruction module is proposed. To address the attenuation of high-frequency structural information in deep features, directional enhancement and frequency-domain reweighting are used to recover fine-grained structural responses of small objects, improving discriminability and localization stability in complex backgrounds.

Experimental results show that, while maintaining real-time inference performance, Aerial-YOLO significantly alleviates the early-stage attenuation of small-object contour features, improves detection accuracy in UAV aerial scenes, and demonstrates good generalization ability and robustness across datasets. The proposed method provides a structurally

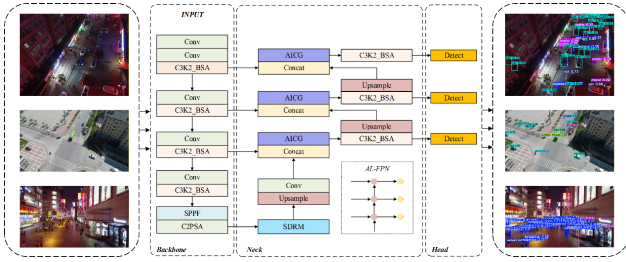


Fig. 1. Aerial-YOLO architecture diagram.

collaborative optimization approach for small object detection under resource-constrained conditions and has strong engineering application potential.

II. AERIAL-YOLO MODEL

This paper proposes a lightweight enhanced model for UAV aerial small object detection based on the YOLOv11 detection framework. The model is named Aerial-YOLO, and its architecture is shown in Fig. 1.

The model systematically reconstructs the backbone and neck of YOLOv11 to address low small-object resolution, drastic scale variation, strong background interference, and dense occlusion in aerial scenes. Although YOLOv11 achieves a good balance between detection speed and accuracy, its standard convolutional stack lacks explicit modeling of high-frequency structural information and directional boundaries. Its default feature pyramid may also cause degradation of small-object details during deep semantic propagation.

Therefore, Aerial-YOLO is designed around structural detail enhancement, scale-adaptive modeling, semantic-guided fusion, and lightweight constraints. First, an improved C3K2_BSA module is introduced into the backbone. By replacing the original Bottleneck unit with boundary-aware enhancement and adaptive multi-scale convolution, directional convolution and scale-weight dynamic allocation are used to strengthen small-object contours and scale-discriminative features from the feature-generation stage. Second, a lightweight AL-FPN is constructed in the feature fusion stage. It removes the redundant high-level detection head and bottom-up path, establishes a unidirectional semantic guidance structure, and uses a cross-scale injection gate to selectively enhance shallow details with high-level semantics, thereby avoiding semantic conflict and detail suppression caused by repeated feature mixing. Meanwhile, a structural detail reconstruction module is embedded between the backbone and neck to further enhance small-object response in complex backgrounds through high-frequency enhancement and scale-adaptive reconstruction. Overall, Aerial-YOLO collaboratively optimizes the backbone representation capability and neck information-flow path of YOLOv11, improving small-object discriminability and localization stability from both feature generation and cross-scale fusion while maintaining a balance between detection accuracy and inference efficiency.

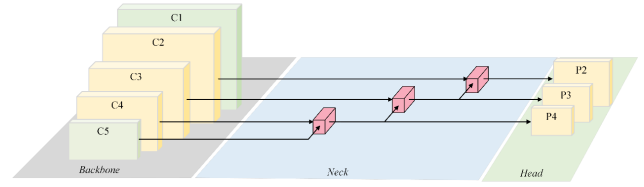


Fig. 2. AL-FPN architecture diagram.

A. ACIG-BASED AL-FPN

During multi-scale feature fusion, small object detection faces a key problem: there is a significant semantic asymmetry between high-level semantic information and low-level detail features. High-level features have stronger semantic discrimination but lower spatial resolution, while low-level features preserve rich structural details but lack semantic expression. Traditional feature pyramid networks usually use layer-by-layer addition or concatenation for fusion. This undifferentiated information aggregation can cause high-level semantics to excessively cover low-level details, weakening the structural response of small objects.

From the perspective of information flow, this problem is essentially non-selective interference of semantic information on structural information. In other words, the cross-level propagation process lacks an effective mechanism for regulating information flow.

To solve this problem, this paper proposes a lightweight semantic-guided feature pyramid named ACIG-Lite FPN (AL-FPN). By constructing an Adaptive Cross-scale Injection Gate (ACIG), as shown in Fig. 2, the model realizes selective guidance and dynamic regulation from high-level semantics to low-level details, thereby suppressing semantic interference while improving the consistency of multi-scale feature representation.

Let the multi-scale features output by the backbone be represented by the corresponding feature maps at different downsampling scales of the input image. Considering that small objects in aerial images are prone to degradation in deep features due to continuous downsampling, and that the proportion of very large objects is relatively low, this paper removes the traditional high-level detection layer and adds a high-resolution detection layer, thereby constructing a three-layer detection structure. In terms of structure, AL-FPN retains only the top-down semantic propagation path and realizes cross-scale information transfer through lightweight channel compression and upsampling.

$$F_4 = \phi(C_4) \quad (1)$$

$$F_3 = \phi(C_3) + U(F_4) \quad (2)$$

$$F_2 = \phi(C_2) + U(F_3) \quad (3)$$

Depthwise separable convolution is used in this formulation to reduce the number of parameters and computational complexity.

To avoid excessive semantic mixing and detail suppression caused by layer-by-layer addition, this paper further introduces the Adaptive Cross-scale Injection Gate (ACIG), which uses high-level semantics to guide the enhancement of low-level details.

For any high-level feature, a channel semantic description vector is first obtained through depthwise separable convolution and global average pooling:

$$z = G(\psi_{dw}(F_h)) \quad (4)$$

Then a semantic weight vector is obtained through a Sigmoid activation function:

$$w = \sigma(z) \quad (5)$$

For the corresponding low-level feature, channel-wise recalibration is performed:

$$\tilde{F}_l = F_l \odot w \quad (6)$$

The operation denotes the Hadamard product. To implement cross-scale detail compensation, a lightweight injection term is constructed:

$$\Delta F = \alpha \cdot U(F_h \odot w) \quad (7)$$

where the scaling factor is learnable. The final fusion result is expressed as:

$$F_l^{out} = \tilde{F}_l + \Delta F \quad (8)$$

This design avoids dimensional expansion caused by channel concatenation and strengthens object-related responses through semantic guidance. Furthermore, to reduce redundant computation and enhance key feature expression, a sparse channel selection strategy is introduced in the fusion stage. The channel importance scoring function is defined as:

$$s_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \left| \tilde{F}_l^{(c)}(i, j) \right| \quad (9)$$

According to this score, the top-ranked channels are selected to construct a sparse mapping operator, and the sparse cross-scale fusion process is expressed as:

$$F_l^{out} = S_k(\tilde{F}_l) + \alpha \cdot U(S_k(F_h \odot w)) \quad (10)$$

By removing the redundant high-level detection head, eliminating the bottom-up path, replacing standard convolution with depthwise separable convolution, and introducing gated sparse cross-scale injection, AL-FPN significantly reduces parameters and FLOPs while alleviating small-object feature degradation and complex background interference in aerial scenes. Structurally, it improves the discriminability and localization stability of small objects and achieves a balance between detection performance and model lightweighting.

B. STRUCTURAL DETAIL RECONSTRUCTION MODULE (SDRM)

In deep feature representations, the fine-grained structural information of small objects often appears as high-frequency components. However, convolutional neural networks tend to preserve low-frequency semantic information during layer-by-layer feature extraction, causing high-frequency structural responses to gradually decay. This phenomenon is especially evident in aerial scenes. Low-frequency textures in complex backgrounds can interfere with small objects, further submerging object boundaries and detail features. From the frequency-domain perspective, high-frequency information is more clearly separable in the spectrum, and therefore structural details can be explicitly enhanced through frequency-domain reweighting. However, relying only on spatial-domain convolution makes it difficult to finely regulate different frequency components.

Based on this analysis, this paper proposes the Structural Detail Reconstruction Module (SDRM). By combining directional structure enhancement, frequency-domain reweighting, and adaptive scale reconstruction, SDRM explicitly restores high-frequency structural information of small objects, improving discriminability and localization stability in complex backgrounds. Given an input feature map, channel compression is first performed through convolution:

$$X_0 = \phi(X) \quad (11)$$

The operation denotes a lightweight channel transformation operator. Directional structure enhancement is then performed. Horizontal and vertical depthwise separable convolution operators are constructed, and learnable directional weights are introduced to obtain the structure-enhanced feature:

$$X_s = X_0 + \beta_h \cdot \psi_h(X_0) + \beta_v \cdot \psi_v(X_0) \quad (12)$$

This design strengthens linear edge and contour responses, effectively improving the separability of small-object boundaries. To suppress redundant low-frequency background information and highlight high-frequency details, a frequency-domain reweighting mechanism is constructed. Let the Fourier transform and inverse Fourier transform be represented by the corresponding operators. The frequency-domain representation is:

$$\hat{X}_s = \mathcal{F}(X_s) \quad (13)$$

The frequency-domain weighting function is defined as:

$$W_f(u, v) = 1 + \gamma \cdot \frac{\sqrt{u^2 + v^2}}{\max(\sqrt{u^2 + v^2})} \quad (14)$$

where the parameter controlling high-frequency enhancement is learnable. The frequency-enhanced result is:

$$\hat{X}_e(u, v) = \hat{X}_s(u, v) \cdot W_f(u, v) \quad (15)$$

The spatial-domain enhanced feature is then obtained through inverse transformation:

$$X_e = \mathcal{F}^{-1}(\hat{X}_e) \quad (16)$$

To improve cross-scale adaptability, SDRM uses parallel multi-scale depthwise convolution branches for scale reconstruction:

$$O_k = \psi_k(X_e) \quad (17)$$

Adaptive fusion weights are then constructed:

$$\omega_k = \frac{\exp(\delta_k)}{\sum_i \exp(\delta_i)} \quad (18)$$

where the scale coefficients are learnable. The final fused output is:

$$O = \sum_k \omega_k O_k \quad (19)$$

Considering that depthwise convolution weakens inter-channel information interaction, a channel shuffle operator is introduced after fusion to promote cross-channel information flow:

$$O' = T(O) \quad (20)$$

Finally, convolution is used to restore the original dimension:

$$Y = \phi^{-1}(O') \quad (21)$$

SDRM strengthens edge responses through directional structure enhancement, improves small-object high-frequency feature representation through frequency-domain reweighting, and realizes scale-dynamic matching through adaptive multi-scale reconstruction. While significantly improving small-object discriminability, all convolutions use depthwise separable forms, giving much lower parameter complexity than standard convolution. Thus, small object detection accuracy is improved while maintaining lightweight model properties.

C. BOUNDARY-AWARE AND SCALE-ADAPTIVE COLLABORATIVE ENHANCEMENT MODULE (C3K2_BSA)

In UAV aerial scenes, small objects usually occupy very few pixels, and their boundary and structural information mainly appears as high-frequency and direction-sensitive features. However, standard convolution operators use isotropic modeling and cannot explicitly model edges with clear directional structures. Repeated downsampling further weakens high-frequency responses in shallow features, causing small-object contour information to degrade early in the feature extraction process.

In addition, small objects differ significantly in representation across scales. Fixed-receptive-field convolution

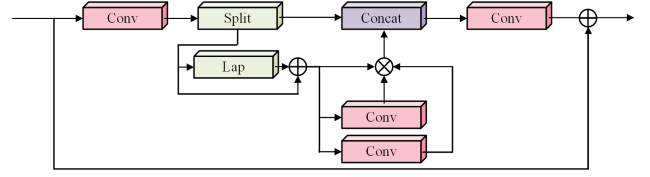


Fig. 3. C3K2_BSA architecture diagram.

structures struggle to adapt to scale variation, which can lead to insufficient feature expression or response aliasing.

Based on this analysis, this paper proposes the Boundary-aware and Scale-adaptive Attention module, named C3K2_BSA. Its structure is shown in Fig. 3. By introducing directional structural modeling and multi-scale adaptive fusion, the module enhances small-object boundary response and scale discrimination while maintaining lightweight characteristics, thereby alleviating structural information degradation from the feature-generation stage.

The original C3K2 module is based on the Cross Stage Partial (CSP) idea and improves feature representation through branched residual stacking. However, its internal Bottleneck structure still uses standard convolution and simple residual connections, lacking explicit modeling of boundary structure and scale variation. As a result, small-object edge information can be submerged by background textures in UAV aerial scenes. Therefore, C3K2_BSA enhances the core Bottleneck unit while preserving the branch-splitting structure of the original C3K2 framework.

Given an input feature, channel splitting is first performed according to the original C3K2 structure:

$$X_1, X_2 = \text{Split}(X) \quad (22)$$

One branch enters the enhanced bottleneck branch, while the identity mapping branch is preserved to maintain gradient stability. In the enhanced bottleneck branch, a boundary-aware operator is first introduced to strengthen small-object contour response. Specifically, depthwise separable Laplacian convolution is used for structural response extraction, and a learnable scaling factor is introduced to obtain boundary-enhanced features:

$$X_b = X_1 + \lambda \cdot \psi_{lap}(X_1) \quad (23)$$

This design explicitly strengthens high-frequency edge components and improves contrast between small objects and the background. Next, to improve scale adaptability, parallel multi-scale depthwise convolution branches are constructed:

$$S_k = \psi_k(X_b) \quad (24)$$

Adaptive weights are then used for dynamic fusion. The weights are generated from a global descriptor:

$$z = \text{GAP}(X_b), \quad a_k = \frac{\exp(\theta_k z)}{\sum_i \exp(\theta_i z)} \quad (25)$$

The multi-scale fusion result is:

$$X_s = \sum_k a_k S_k \quad (26)$$

After obtaining scale-enhanced features, a lightweight spatial-channel joint attention mechanism is introduced to suppress redundant responses. Directional encoding pooling is first performed:

$$f_h(i) = \frac{1}{W} \sum_{j=1}^W X_s(i, j), \quad f_w(j) = \frac{1}{H} \sum_{i=1}^H X_s(i, j) \quad (27)$$

The fused feature is then obtained through shared convolutional mapping:

$$f = \delta(\text{Conv}_{1 \times 1}([f_h, f_w])) \quad (28)$$

The operation denotes a nonlinear activation function. Directional attention weights are then generated:

$$A_h = \sigma(\text{Conv}_{1 \times 1}(f)), \quad A_w = \sigma(\text{Conv}_{1 \times 1}(f)) \quad (29)$$

The final enhanced output is:

$$X_a = X_s \odot A_h \odot A_w \quad (30)$$

The enhanced branch output and identity mapping branch are concatenated:

$$Y = \text{Conv}_{1 \times 1}(\text{Concat}(X_a, X_2)) \quad (31)$$

The final output is formed through a residual connection:

$$Y_{out} = Y + X \quad (32)$$

C3K2_BSA preserves the parameter-efficient nature of C3K2 while explicitly strengthening small-object contours and scale-discriminative information through boundary-aware enhancement, small-scale adaptive fusion, and directional attention. It effectively alleviates localization ambiguity and feature aliasing in dense aerial scenes. Meanwhile, all spatial convolutions use depthwise separable forms, reducing parameter complexity compared with traditional convolution while maintaining overall lightweight properties.

III. EXPERIMENTS AND RESULT ANALYSIS

A. EXPERIMENTAL DATASET SETTINGS

To comprehensively verify the effectiveness and generalization capability of Aerial-YOLO for UAV aerial small object detection, this paper uses two representative public datasets: VisDrone2019 and TinyPerson. These datasets focus respectively on complex urban scenes and extremely small object detection, enabling model performance to be evaluated from different perspectives.

The VisDrone2019 dataset is collected by UAV platforms in real urban scenes and contains multiple categories such

as pedestrians, vehicles, and bicycles. It features large view-point variation, imbalanced object scales, a high proportion of small objects, and complex backgrounds. The dataset provides training, validation, and testing splits. Object sizes vary widely, and many objects occupy only a small number of pixels, making it a typical small-object detection scenario. This paper follows the official split for training and evaluation and uses consistent input resolution and evaluation metrics to ensure fair comparison.

The TinyPerson dataset focuses on extremely small pedestrian detection. Its targets are usually smaller than 20×20 pixels and are densely distributed in complex backgrounds, placing higher requirements on fine-grained feature representation and interference resistance. Compared with VisDrone2019, TinyPerson is more challenging in terms of target size and distribution and is suitable for evaluating detection performance and cross-dataset generalization under extreme small-object conditions. This paper uses its standard training and testing split and conducts experiments without introducing additional data.

B. EXPERIMENTAL ENVIRONMENT SETTINGS

To ensure the stability and reproducibility of the experimental results, model training and testing are performed in a unified hardware and software environment. The hardware platform uses a high-performance computing device equipped with an NVIDIA GeForce RTX 3090 GPU with 32 GB of memory, satisfying the memory and computation requirements of large-scale aerial data training. A multi-core CPU architecture and sufficient system memory are used to improve data loading and preprocessing efficiency. In terms of software, the training process is implemented in Python on Linux. PyTorch is used as the deep learning framework, and the official YOLO implementation is modified for development. CUDA and cuDNN acceleration libraries are configured to improve training and inference efficiency.

For training, this paper follows common hyperparameter settings used by the YOLO series and makes appropriate adjustments according to the characteristics of UAV aerial small object detection. The input image resolution is uniformly set to 640×640 , balancing detection accuracy and inference speed. SGD is used as the optimizer, with an initial learning rate of 0.01 and cosine annealing for dynamic adjustment to improve convergence stability. Momentum is set to 0.937, and weight decay is set to $5e-4$ to alleviate overfitting. Batch size is set to 16 according to GPU memory. Automatic mixed precision (AMP) is enabled during training to improve efficiency and reduce memory consumption.

For data augmentation, Mosaic, random flipping, color perturbation, and other common augmentation strategies are used to improve adaptability to complex backgrounds and scale variation. The number of training epochs is set to 300 to ensure sufficient convergence. During inference, data augmentation is disabled, and the same confidence threshold

and non-maximum suppression (NMS) parameters are used to ensure fair comparison across models.

C. EVALUATION METRICS

To comprehensively evaluate the performance of the proposed method for UAV aerial small object detection, this paper uses common accuracy and efficiency metrics in object detection, including precision, recall, average precision, parameter count, and floating-point operations.

Precision measures the proportion of predicted positive samples that are true positives:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (33)$$

where TP denotes correctly detected targets and FP denotes false positive detections.

Recall measures the proportion of all real targets that are correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (34)$$

where FN denotes real targets that are not detected.

Average Precision (AP) is obtained by calculating the area under the precision-recall curve at different confidence thresholds and is used to comprehensively measure detection performance. Mean Average Precision (mAP) is then obtained by averaging AP over all categories:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (35)$$

where the variable represents the number of categories and each term represents the AP of a class. This paper uses mAP@0.5 and mAP@0.5:0.95 as the main evaluation metrics. mAP@0.5 denotes detection accuracy at an IoU threshold of 0.5, while mAP@0.5:0.95 denotes average detection performance over IoU thresholds from 0.5 to 0.95 with a step size of 0.05.

In addition, parameter count and FLOPs are introduced to evaluate model lightweightness and computational efficiency. Parameter count measures model size, while FLOPs measure computational cost during forward propagation, thereby jointly evaluating model applicability in resource-constrained scenarios.

D. ABLATION EXPERIMENTS

1) BACKBONE NETWORK IMPROVEMENT ANALYSIS

To verify the performance improvement brought by introducing the boundary-aware and scale-adaptive collaborative enhancement module into the backbone network, the proposed model is compared with the YOLOv11n baseline and models using different improved C3K2 modules on the VisDrone2019 dataset. The experimental results are shown in Table I.

As shown in Table I, after introducing the C3K2_BSA module, the model outperforms other improved C3K2 variants in precision, recall, mAP@50, and mAP@50-95. The

improvement in mAP@50 is especially evident, increasing by 1.9 percentage points compared with the original C3K2 model, while mAP@50-95 improves by 0.9 percentage points. This indicates that BSA can effectively strengthen the capture of small-object boundary and scale features, reduce feature loss during downsampling and convolution, and significantly improve small-object detection accuracy in complex backgrounds.

2) FEATURE PYRAMID NETWORK IMPROVEMENT ANALYSIS

For the problems of high small-object proportion, complex background texture, and highly imbalanced scale distribution in aerial scenes, this paper reconstructs the PAFPN structure commonly used in YOLO models and proposes a lightweight semantic-guided feature pyramid network named AL-FPN. To verify the effectiveness of ACIG in feature aggregation, AL-FPN is compared with YOLOv11n and different feature fusion strategies. The lightweight unidirectional fusion structure without the gating mechanism is denoted as L-FPN, while the model with ACIG is denoted as AL-FPN. The experimental results are shown in Table II.

As shown in Table II, compared with the PAFPN baseline, A2-FPN achieves the highest performance in mAP@50 and mAP@50-95, indicating that more complex attention mechanisms can improve feature fusion to some extent. However, the proposed AL-FPN still outperforms the baseline in multiple metrics while maintaining a lightweight structure. Its mAP@50 increases by 0.6 percentage points and its mAP@50-95 increases by 0.7 percentage points, while its precision reaches 67.4%, higher than the other compared methods.

Further comparison between L-FPN and AL-FPN shows that all metrics improve after introducing ACIG. This indicates that the gating mechanism effectively enhances cross-scale feature interaction, realizes selective guidance from high-level semantics to low-level details, and improves small-object feature representation.

It should be noted that although AL-FPN does not achieve the best result in every metric, this is because the proposed structure prioritizes model lightweightness and computational efficiency, constraining the fusion path and module complexity. Compared with complex structures that introduce large amounts of attention computation, AL-FPN achieves stable performance improvement with lower computational overhead, demonstrating a favorable accuracy-efficiency tradeoff.

3) MULTI-MODULE ABLATION ANALYSIS

To systematically evaluate the influence of each proposed module on model performance, YOLOv11n is used as the baseline, and C3K2_BSA, AL-FPN, and SDRM are introduced individually and in combination. The experimental results are shown in Table III, where the check mark indicates that the corresponding module is introduced and the first row is the baseline.

TABLE I. Improved ablation experiments of the C3K2 module

Method	P	R	mAP@50	mAP@50-95
C3K2	66.9	42.4	46.8	26.9
C3K2-EMA [14]	66.5	42.6	46.5	27.2
C3K2-Faster-EMA [15]	67.1	42.8	46.9	27.1
CMA-C3K2 [16]	67.3	42.6	47.2	27.3
HS-C3K2 [17]	67.2	43.1	47.6	27.4
C3K2_BSA	67.5	43.1	48.7	27.8

TABLE II. Improved ablation experiments of the FPN module

Method	P	R	mAP@50	mAP@50-95
PAFPN	66.9	42.4	46.8	26.9
MuRDE-FPN [18]	66.8	42.9	46.7	27.3
A2-FPN [19]	67.2	42.8	47.1	27.4
L-FPN	66.8	42.6	46.5	26.8
AL-FPN	67.4	42.7	47.4	27.6

TABLE III. Aerial-YOLO ablation experiment

BSA	AL-FPN	SDRM	FLOPs	Parameters	mAP@50	mAP@50-95
			6.2	2.59	46.8	26.9
Yes			7.1	2.61	48.7	27.2
	Yes		9.3	2.28	47.4	27.8
		Yes	7.2	2.71	49.2	28.9
Yes	Yes	Yes	11.2	2.43	50.8	30.6

As shown in Table III, each improved module enhances the baseline model to varying degrees for UAV aerial small object detection. First, after C3K2_BSA is introduced into the backbone, mAP@50 increases from 46.8% to 48.7%, and mAP@50-95 increases to 27.2%. This shows that the module improves representation of small-object contours and structural information through boundary awareness and scale adaptability. However, the multi-scale convolution and structural enhancement operations slightly increase computation.

Second, when AL-FPN is used alone, the parameter count decreases to 2.28M, while mAP@50 increases to 47.4% and mAP@50-95 reaches 27.8%, showing strong lightweight properties and performance improvement. This result indicates that ACIG can reduce redundant feature transfer while enabling high-level semantics to effectively guide low-level details, improving multi-scale feature fusion efficiency.

Third, introducing SDRM between the backbone and the neck brings a more significant performance improvement. mAP@50 rises to 49.2%, and mAP@50-95 rises to 28.9%. This shows that directional enhancement and frequency-domain reweighting can effectively restore high-frequency details of small objects and strengthen discrimination in complex backgrounds. At the same time, parameter count increases, indicating that this module brings a certain computational cost while improving performance.

When all three modules are integrated, the complete Aerial-YOLO model achieves mAP@50 and mAP@50-95 of 50.8% and 30.6%, respectively, improving over the base-

line by 4.0 and 3.7 percentage points. Although total computation increases, the parameter count remains controlled at 2.43M, maintaining good lightweight characteristics.

The experimental results demonstrate good synergy among the proposed modules. C3K2_BSA provides boundary and scale perception, AL-FPN enables efficient semantic-guided fusion, and SDRM further supplements detail recovery. Together, they improve detection performance for aerial small objects. In summary, the proposed multi-module collaborative optimization strategy enhances feature expression and fusion while maintaining lightweight characteristics, achieving better accuracy and robustness in complex UAV scenes.

E. COMPARATIVE ANALYSIS

To comprehensively verify the detection performance and lightweight advantages of Aerial-YOLO, this paper compares it with representative two-stage detection models, mainstream YOLO-series models, and recent related improved methods. The results are shown in Table IV.

Overall, traditional two-stage detection models such as Faster R-CNN and RetinaNet perform significantly worse than YOLO-based one-stage detection models in mAP@50 and mAP@50-95. Their parameter counts and computational costs are also much larger, making it difficult for them to satisfy real-time and lightweight requirements on UAV platforms.

Among YOLO-series models, YOLOv5n, YOLOv6n, YOLOv8n, YOLOv10n, and YOLOv11n all achieve relatively stable detection performance with low model

complexity. YOLOv11n reaches 46.8% mAP@50 and 26.9% mAP@50-95, showing balanced performance among lightweight models. In comparison, Aerial-YOLO further improves mAP@50 to 50.8% and mAP@50-95 to 30.6%, demonstrating that the proposed multi-module collaborative optimization strategy effectively enhances small-object detection capability.

Compared with recent improved methods, MSAF-YOLO and STF-YOLOv8n show relatively strong accuracy, and STF-YOLOv8n reaches 52.5% mAP@50, but its computation increases significantly. Transformer-based IF-DETR and CSFF-DETR have certain advantages in mAP@50-95, but their parameter counts and computational complexity are high and unfavorable for practical deployment. In contrast, Aerial-YOLO achieves competitive mAP@50 and mAP@50-95 while maintaining a low parameter count and moderate computation.

In summary, Aerial-YOLO achieves a good balance between detection accuracy and model complexity. Although it is slightly lower than some high-complexity models in certain metrics, it still delivers substantial performance gains under lightweight constraints and is more suitable for resource-constrained UAV aerial scenarios, verifying the effectiveness and feasibility of the proposed method in practical applications.

F. VISUALIZATION ANALYSIS

To intuitively verify the detection performance of the proposed method in complex aerial scenes, representative scenes are selected for visualization analysis, including insufficient illumination, long-distance small objects, and complex background interference. The visualization results show that Aerial-YOLO performs better in these challenging scenarios.

As shown in Fig. 4, in low-light scenes, YOLOv11 is prone to missed detections because small objects have low contrast and weak boundary information. By introducing the structural detail reconstruction module, the proposed method strengthens high-frequency responses through frequency-domain reweighting and directional enhancement. This improves the boundary clarity of small objects in dark regions and enables the model to capture weak-texture targets more accurately, thereby significantly reducing missed detections.

As shown in Fig. 5, in long-distance small object detection scenes, target sizes are extremely small and feature information is limited. Conventional models easily lose key structural information after repeated downsampling. Through the boundary-aware and scale-adaptive feature extraction module, the proposed method enhances shallow feature representation of small-object contours and scale variation, allowing long-distance targets to be more fully preserved in feature space. This effectively improves detection recall and reduces missed detections.

As shown in Fig. 6, in complex background scenes, ground textures, building structures, and dense object distributions can introduce interference, causing false detections

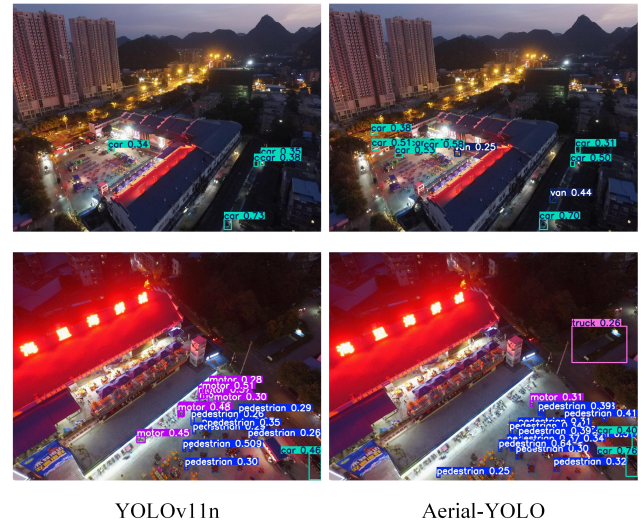


Fig. 4. Detection results of YOLOv11 and Aerial-YOLO under insufficient light conditions.

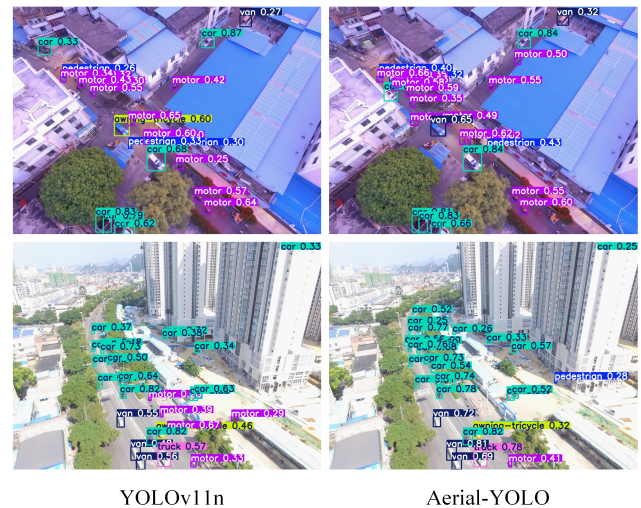


Fig. 5. Detection results of YOLOv11 and Aerial-YOLO for small targets at long distances.

and misdetections. To address this issue, this paper designs a lightweight semantic-guided feature pyramid in the neck network. Through the adaptive cross-scale injection gate, high-level semantics selectively enhance low-level details, effectively suppressing background noise interference, improving distinction between objects and the background, reducing false detections, and improving detection accuracy.

Overall, the proposed method achieves clear improvements in representative difficult scenarios such as insufficient illumination, long-distance small objects, and complex backgrounds. This verifies the effectiveness and robustness of the proposed multi-module collaborative optimization strategy for improving UAV aerial small object detection.

TABLE IV. Aerial-YOLO comparison experiment

Category	Method	mAP@50	mAP@50-95	Parameters	FLOPs
Two-stage / Retina-style detectors	Faster-RCNN	33.1	16.9	108.5	171.5
Two-stage / Retina-style detectors	RetinaNet	36.3	15.8	37.8	91.8
YOLO series	YOLOv5n	46.3	26.5	2.5	7.1
YOLO series	YOLOv6n	45.7	26.2	4.3	11.8
YOLO series	YOLOv8n	45.6	26.6	3.1	8.2
YOLO series	YOLOv10n	46.1	26.8	2.3	6.7
YOLO series	YOLOv11n	46.8	26.9	2.6	6.2
Recent methods	MSAF-YOLO [20]	47.3	29.4	8.7	36.8
Recent methods	STF-YOLOv8n [21]	52.5	30.5	2.9	14.2
Recent methods	IF-DETR [22]	50.6	31.4	17.2	64.5
Recent methods	CSFF-DETR [23]	51.7	32.5	11.8	59.8
Proposed	Aerial-YOLO	50.8	30.6	2.4	11.2



Fig. 6. Detection results of YOLOv11 and Aerial-YOLO under complex background conditions.

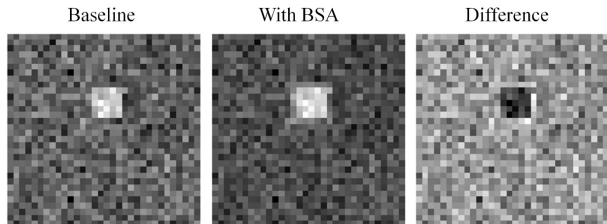


Fig. 7. Local feature response comparison and difference analysis between the baseline and the proposed method.

G. FEATURE RESPONSE AND STRUCTURAL INFORMATION ANALYSIS

To further analyze the model's feature response capability in small-object regions, this paper compares the baseline model and improved model from the perspective of local features. Because global feature maps contain a large amount of background information and cannot clearly show response differences in small-object regions, representative target regions are cropped for analysis.

As shown in Fig. 7, compared with the baseline model, after introducing the boundary-aware module, the response in the target region becomes more concentrated and the boundary structure becomes clearer. This indicates that the module effectively enhances shallow feature representation of small-object contour information.

The difference map further shows that response enhancement is mainly concentrated around target edges, while

changes in background regions are small. This demonstrates that the proposed method improves structural information representation while effectively suppressing background interference.

Overall, the proposed method significantly improves local feature representation quality for small objects, verifying the reason for performance improvement at the feature level.

H. MODEL GENERALIZATION ANALYSIS

To further verify the generalization capability of the proposed method under different data distributions, this paper evaluates the model on the TinyPerson dataset. This dataset focuses on extremely small pedestrians. Targets are generally small and densely distributed, with complex background interference, which places higher requirements on fine-grained feature extraction and cross-scene adaptability. As shown in Table V, compared with the YOLOv11n baseline, Aerial-YOLO maintains higher detection accuracy on TinyPerson and achieves improvements in both mAP@0.5 and mAP@0.5:0.95, indicating strong cross-scene adaptability.

From the perspective of model structure, the boundary-aware and scale-adaptive module enhances robust representation of small-object contours and scale variation, allowing the model to maintain stable detection performance under different resolution distributions. The semantic-guided feature pyramid improves the consistency of multi-scale feature representation through selective cross-scale information injection. The structural detail reconstruction module strengthens high-frequency response and effectively alleviates feature suppression of extremely small objects in complex backgrounds. Therefore, the synergy of these modules enables the model to maintain strong generalization capability when facing the distribution differences of the TinyPerson dataset.

TABLE V. Generalization results of Aerial-YOLO on the TinyPerson dataset

Method	P	R	mAP@50	mAP@50–95	Parameters	FLOPs
YOLOv11n	73.3	57.12	64.5	26.9	2.59	6.2
Aerial-YOLO	75.2	59.63	65.1	27.1	2.43	11.2

IV. CONCLUSION

This paper proposes an improved lightweight object detection model named Aerial-YOLO for UAV aerial scenes with a high proportion of small objects, large scale variation, and complex backgrounds. By introducing a boundary-aware and scale-adaptive feature extraction module into the backbone, designing a semantic-guided feature pyramid in the neck, and combining it with a structural detail reconstruction mechanism, the proposed method effectively improves feature representation and multi-scale fusion for small objects. Experimental results on VisDrone2019 and TinyPerson show that the proposed method significantly improves detection accuracy while maintaining low parameter count and computational cost, and it demonstrates strong robustness and generalization capability in complex scenes. In summary, the proposed method achieves a good balance between accuracy and efficiency and provides an efficient and feasible solution for UAV aerial small object detection.

REFERENCES

- [1] R. Patel, "Application and future prospects of geographic information system (GIS) in intelligent transportation," *Transactions on Computational and Scientific Methods*, vol. 5, no. 3, 2025.
- [2] W. Hua and Q. Chen, "A survey of small object detection based on deep learning in aerial images," *Artificial Intelligence Review*, vol. 58, no. 6, p. 162, 2025.
- [3] S. Lu, Y. Guo, J. Long, et al., "Dense small target detection algorithm for UAV aerial imagery," *Image and Vision Computing*, vol. 156, p. 105485, 2025.
- [4] E. Edozie, A. N. Shuaibu, U. K. John, et al., "Comprehensive review of recent developments in visual object detection based on deep learning," *Artificial Intelligence Review*, vol. 58, no. 9, p. 277, 2025.
- [5] M. Trigka and E. Dritsas, "A comprehensive survey of machine learning techniques and models for object detection," *Sensors*, vol. 25, no. 1, p. 214, 2025.
- [6] D. Nguyen and V.-D. Hoang, "A Lightweight Multi-Scale Attention Model for Small Object Detection in UAV Imagery," *IEEE Access*, 2026.
- [7] Y. Wang, S. Li, Y. Zhang, et al., "Lightweight concrete crack segmentation network for drone image with complex backgrounds using multi-scale feature fusion and optimized architecture," *Construction and Building Materials*, vol. 495, p. 143667, 2025.
- [8] S. Zhou, C. Rong, and H. Wang, "A lightweight YOLOv8-based small object detection algorithm for UAV scenes," *Signal Processing*, vol. 42, no. 1, pp. 72–82, 2026.
- [9] S. Zhang, F. Xiao, and C. Ge, "Lightweight aerial image detection algorithm based on improved RT-DETR," *Electronic Measurement Technology*, vol. 48, no. 22, pp. 187–197, 2025.
- [10] G. Wang, X. Li, and Y. Liu, "UAV target recognition algorithm based on residual separable convolution and multi-scale fusion," *Electronics Optics & Control*, vol. 32, no. 12, pp. 43–49, 112, 2025.
- [11] X. Liao and X. Tang, "A lightweight UAV aerial target detection model based on YOLOv10n," *Infrared Technology*, vol. 47, no. 11, pp. 1427–1434, 2025.
- [12] G. Chang, S. Zhao, S. Hou, et al., "Improved YOLOv8n-based UAV aerial small object detection algorithm," *Science Technology and Engineering*, vol. 25, no. 31, pp. 13500–13508, 2025.
- [13] T. Jing, Y. Cao, X. Chen, et al., "Improved YOLOv11 for UAV aerial small object detection," *Computer Engineering and Applications*, vol. 62, no. 2, pp. 138–148, 2026.
- [14] H. Tao, L. Zhang, Z. Sun, et al., "Multi-strategy improvement of coal gangue recognition method of YOLOv11," *Sensors*, vol. 25, no. 7, p. 1983, 2025.
- [15] K. Zhou and S. Jiang, "Forest fire detection algorithm based on improved YOLOv11n," *Sensors*, vol. 25, no. 10, p. 2989, 2025.
- [16] J. Zhang and X. L. Peng, "YOLOv11-LC: accurate detection of benign and malignant endometrial lesions using LCA-enhanced YOLOv11 with CMA-C3k2," *Signal, Image and Video Processing*, vol. 20, no. 1, p. 15, 2026.
- [17] L. Shang, Y. Qiao, H. Lei, et al., "WSADet: A wavelet scale-aware UAV object detector for complex conditions," *Pattern Recognition*, p. 112965, 2025.
- [18] M. Kisieliute and I. Daugela, "MuRDE-FPN: Precise UAV Localization Using Enhanced Feature Pyramid Network," *Drones*, vol. 10, no. 3, p. 162, 2026.
- [19] M. Hu, Y. Li, L. Fang, et al., "A2-FPN: Attention aggregation based feature pyramid network for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [20] J. Chen and J. Liu, "MSAF-YOLO: A high-performance small object detection network for aerial images," *Infrared Technology*, pp. 1–9.
- [21] D. Zhu, Y. Wei, F. Wang, et al., "UAV ground small object detection algorithm based on multiple attention mechanisms," *Laser & Optoelectronics Progress*, vol. 63, no. 4, pp. 452–466, 2026.
- [22] Y. Yang, W. Wei, S. Zheng, et al., "IF-DETR: An improved RT-DETR aerial image detection algorithm," *Electronic Measurement Technology*, pp. 1–12.
- [23] M. Li and C. She, "UAV aerial small object detection algorithm based on cross-scale feature fusion," *Electronic Measurement Technology*, pp. 1–10.