Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

D3F-Det: Progressive Fine-Grained Feature Refinement and Reuse for Robust Small Object Detection in Aerial Imagery

LingLing Li
Zhengzhou University of Aeronautics
JiaQing Liu
liujiaqing@zua.edu.cn

Zhengzhou University of Aeronautics
XueZhuan Zhao
Zhengzhou University of Aeronautics
XiaoYan Shao
Zhengzhou University of Aeronautics
MengMeng Tang
Zhengzhou University of Aeronautics
YaXuan Xing
Zhengzhou University of Aeronautics

Research Article

Keywords: Small Object Detection, UAV Imagery, Fine-Grained Feature Refinement, Progressive Feature Reuse, Robust Detection

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9197165/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Noname manuscript No.
(will be inserted by the editor)

D3F-Det: Progressive Fine-Grained Feature Refinement and Reuse for Robust Small Object Detection in Aerial Imagery

LingLing Li[1,2,3] · **JiaQing Liu**[1] · **XueZhuan Zhao**[1] · **XiaoYan Shao**[1] · **MengMeng Tang**[1] · **YaXuan Xing**[1]

Received: date / Accepted: date

Abstract The detection of small objects in unmanned aerial vehicle (UAV) imagery remains a formidable challenge due to the severe degradation and underutilization of fine-grained information across the detection pipeline. Conventional methods often suffer from irreversible loss of spatial details during downsampling, a lack of effective same-scale spatial-semantic alignment, and the progressive dilution of shallow features in multi-scale fusion. To address these interconnected issues, we propose a unified real-time detection framework termed D3F-Det, which is designed to preserve, enhance, and progressively reuse fine-grained information for robust small object detection in aerial imagery. Our architecture comprises three novel components: a dual-branch detail-preserving downsampling module (DBDown) that reduces resolution while maintaining critical shallow details; a dual-feature multi-scale collaborative module (DFMS) embedded within same-scale layers to align and enhance spatial and semantic representations; and a three-stage reuse neck (3S-RN) that iteratively reintegrates high-resolution shallow features throughout the feature pyramid. We demonstrate that our method significantly improves performance, achieving a mean Average Precision (mAP@0.5:0.95) of 30.3% on the challenging VisDrone2019 dataset, a 6.2 percentage point gain over the YOLOv12-s baseline, while concurrently reducing the parameter count from 9.3M to 3.7M. Extensive experiments on the AI-TOD and TinyPerson datasets further validate its robustness in extreme small-object and densely populated scenarios. By systematically preserving and exploiting fine-grained features, our D3F-Det framework offers a more effective and generalisable solution for robust object detection in complex aerial environments.The source code is available at `https://github.com/LJQA1/D3F-Det` (DOI: `10.5281/zenodo.19162854`).

Keywords Small Object Detection · UAV Imagery · Fine-Grained Feature Refinement · Progressive Feature Reuse · Robust Detection

1 Introduction

Unmanned aerial vehicle (UAV)–based object detection plays a critical role in a wide range of applications, including traffic surveillance, disaster response, and aerial reconnaissance [1, 2, 3]. Although deep learning–based detectors have achieved remarkable success on natural scene benchmarks, their performance degrades significantly when applied to high-resolution UAV imagery, particularly for small objects [4] .

According to the commonly adopted definition in the MS COCO dataset, objects with spatial resolutions smaller than 32×32 pixels are categorized as small objects [5]. In aerial scenes, such objects usually occupy only a tiny fraction of the image and are easily affected by scale variation, complex backgrounds, and occlusion [6, 7]. More importantly, high-resolution UAV images inevitably undergo multiple stages of downsampling during convolutional feature extraction. As illustrated in Figure 1, successive downsampling operations

Corresponding author. Email: jiaqingliu@zua.edu.cn

[1] School of Computer Science, Zhengzhou University of Aeronautics, No.15 Wenyuan West Road, Zhengzhou 450046, Henan, China
[2] Henan Province Multi-Modal Information Perception and Computing Engineering Research Center, Zhengzhou 450046, Henan, China
[3] Henan Provincial University-Enterprise R&D Center for Artificial Intelligence Technology, Zhengzhou 450046, Henan, China

cause severe degradation of fine-grained spatial structures and texture details at early stages [8], which is especially detrimental to small objects. Once such fine-grained information is lost [9] in shallow layers, it is difficult to recover through subsequent upsampling or feature fusion, fundamentally limiting detection performance.

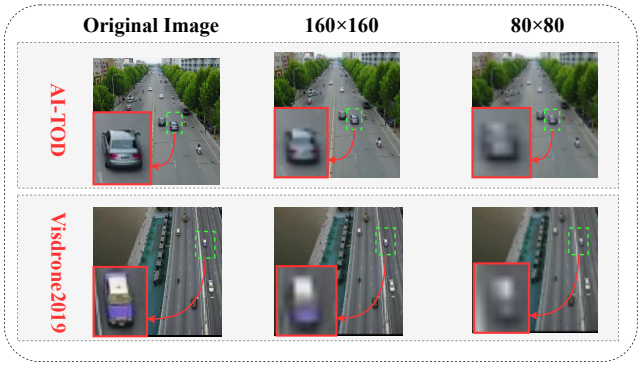


Fig. 1: Fine-grained appearance degradation of small objects during downsampling.Visualization of progressive detail loss of small objects as the feature resolution decreases from the original image to 160×160 and 80×80 on the AI-TOD and VisDrone2019 datasets. Critical appearance cues of small objects are severely blurred or lost under successive downsampling, highlighting the inherent challenge of preserving fine-grained information in conventional feature extraction pipelines.

To alleviate fine-grained information degradation, existing studies have explored architectural modifications from different stages of the detection pipeline. At the feature extraction stage, some methods attempt to preserve spatial details by adjusting downsampling strategies, such as reducing the stride of early convolutional layers (e.g., TY-RIST) or introducing dilated convolutions [10] to enlarge receptive fields without reducing feature resolution. At the feature fusion stage, various feature pyramid–based architectures have been developed to improve multi-scale representations, including the Feature Pyramid Network (FPN) [11], Path Aggregation Network (PAN) [12], Bidirectional Feature Pyramid Network (BiFPN) [13], and Path Aggregation Feature Pyramid Network (PAFPN)) [14].

Despite these efforts, existing methods still encounter three interrelated challenges in aerial small object detection: **(1) Loss of shallow fine-grained features:** During successive downsampling and feature extraction, the critical spatial structures of small objects are progressively weakened, resulting in insufficient feature representation capacity.**(2) Difficulty in same-scale feature modeling and alignment:** Even when high-resolution features are partially preserved, small object detection still relies on effective collaboration between local spatial details and semantic context within the same feature scale. Conventional convolutional blocks employ fixed receptive fields and lack dedicated mechanisms to jointly enhance fine-grained details and contextual information, leading to weak discriminative representations for small objects.**(3) Insufficient reuse of shallow features during multi-scale fusion:** In most FPN/PAN-style necks, shallow high-resolution features are explicitly injected only once into the fusion pathway and are progressively diluted as they are repeatedly aggregated with deeper semantic features. This one-pass fusion strategy limits the sustained contribution of fine-grained information, resulting in suboptimal feature alignment and insufficient utilization of shallow cues across multiple detection scales.

These challenges indicate that preserving fine-grained information at a single stage is insufficient [15, 16] ; instead, small object detection in UAV imagery requires a unified architectural design that preserves, enhances, and fully exploits fine-grained features throughout downsampling, feature modeling, and multi-scale fusion.

Motivated by the above analysis, this paper proposes an aerial small object detection framework termed D3F-Det. As shown in Figure 2, compared with the conventional FPN pipeline, the proposed framework preserves fine-grained features and enables multi-stage feature reuse, leading to more complete and robust small object detection.

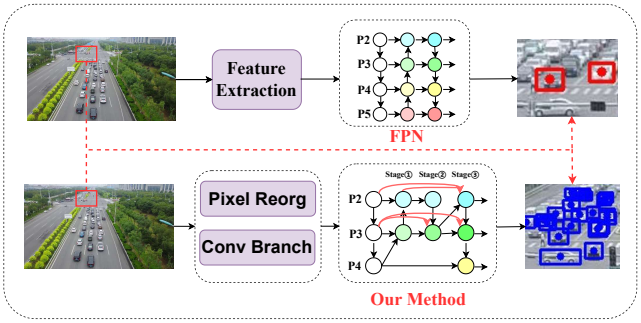


Fig. 2: Illustration of the effectiveness of the proposed multi-stage feature reuse strategy for small object detection. Compared with the conventional FPN pipeline, the proposed method preserves fine-grained features and enables multi-stage feature reuse, leading to more complete and robust small object detection.

The main contributions of this work are summarized as follows:

– We propose a dual-branch detail-preserving downsampling strategy that achieves scale compression while retaining fine-grained features. By exploiting complementary feature extraction pathways, the proposed design mitigates early-stage information degradation and establishes robust shallow representations for subsequent multi-scale modeling.
– We design a same-scale dual-feature collaborative module that integrates multi-receptive-field modeling with feature interaction mechanisms. By jointly modeling local spatial details and semantic context within unified feature layers, the module enhances discriminative representations of small objects while maintaining controlled model complexity.
– We construct a three-stage feature reuse neck that progressively injects shallow high-resolution features into multi-scale fusion. Through staged fine-grained information integration and alignment, the proposed neck strengthens cross-scale feature interaction and improves the sustained contribution of shallow cues to detection heads.
– Extensive experiments on the VisDrone2019, AI-TOD, and TinyPerson datasets demonstrate that D3F-Det not only surpasses the YOLOv12 baseline but also achieves superior performance over several recent state-of-the-art detectors under comparable computational complexity, highlighting the effectiveness and robustness of the proposed framework.

2 Related Work

2.1 Small Object Detection in Aerial Scenarios

Small object detection in UAV-based remote sensing and aerial imagery remains a fundamental challenge [17] due to extremely limited object scales, scarce discriminative features, and complex background interference. To address these issues, recent studies have explored various strategies,including feature enhancement, contextual modeling, and multi-scale perception. For example, HS-FPN [18] enhances small object representations through high-frequency perception and spatial dependency modeling, while S^2FCNet [19] introduces local semantic reactivation and foreground spatial perception mechanisms to compensate for insufficient low-level semantics. HAFNet [20] focuses on infrared small object detection by strengthening cross-scale interactions, whereas CFENet [21] improves localization accuracy through contextual feature enhancement. Although these methods achieve promising improvements in specific scenarios, many of them rely on task-specific components or additional structural complexity, which may limit scalability and real-time deployment in UAV applications. Moreover, most approaches focus on enhancing isolated modules rather than systematically addressing fine-grained information preservation and utilization across the entire detection pipeline. Motivated by this gap, our work proposes a unified architectural framework that consistently preserves, enhances, and reuses fine-grained features to improve aerial small object detection performance.

2.2 Fine-Grained Feature Preservation and Downsampling

Downsampling operations are essential for enlarging receptive fields and reducing computational cost; however, they inevitably degrade fine-grained information such as edges and textures, which is particularly detrimental to small objects. To alleviate this issue, various feature-preserving strategies have been proposed. LSM-YOLO [22] introduces a lightweight adaptive extraction module with dual-branch parameter sharing to retain critical details, while IF-YOLO [23] incorporates an information-preserving feature aggregation module to jointly optimize scale compression and cross-level fusion. SMA-YOLO [24] preserves small object details via parameter-free sliced convolution and further enhances detection performance through cross-scale interaction mechanisms. Despite these improvements, many existing methods introduce additional branches or complex operations to compensate for information loss, which may increase computational burden and structural redundancy. Furthermore, most approaches attempt to mitigate degradation after downsampling rather than redesigning the downsampling mechanism itself. To address this structural limitation, our work develops a detail-preserving downsampling strategy that performs scale reduction and fine-grained retention in a coordinated manner, providing robust shallow representations for subsequent multi-scale modeling.

2.3 Feature Alignment and Scale Modeling

Multi-scale object detection depends on coordinated modeling of fine-grained spatial details and high-level semantic representations, where effective feature alignment and scale awareness are critical. However, achieving precise alignment across heterogeneous feature dimensions remains challenging. For instance, MSHC [25] enhances feature coupling through visual–semantic embedding alignment but primarily emphasizes global similarity modeling and provides limited spatial refinement.

Beyond feature alignment, researchers have also explored advanced fusion architectures and dual-branch designs to boost feature representation. Specifically, FFFN [26] implements iterative feedback fusion to refine feature details across frames, while SCNet [27] adopts a dual-branch structure combining Transformer and ConvNeXt to enhance robust feature modeling under complex disturbances. Other methods aim to align local regions with semantic attributes or strengthen cross-layer correspondence [28], yet often overlook the diverse receptive field requirements of objects at different scales. As a result, these approaches may struggle to simultaneously preserve fine-grained details for small objects while maintaining robust semantic representations for larger ones. This imbalance indicates the necessity of same-scale collaborative modeling mechanisms that can jointly enhance local spatial cues and contextual semantics within unified feature layers. Accordingly, our work emphasizes adaptive receptive field design and same-scale feature collaboration to achieve more effective integration of spatial and semantic information.

2.4 Feature Pyramid and Multi-Scale Fusion Networks

Feature Pyramid Networks (FPNs) have become a canonical architecture for multi-scale object detection by integrating high-level semantic information with low-level spatial details through top-down pathways and lateral connections. To strengthen feature interaction, PANet introduces a bottom-up path augmentation mechanism, while subsequent variants explore bidirectional enhancement strategies [29, 30] to improve semantic representation and detail preservation.Targeting small object detection tasks, recent advanced FPN-based methods have made targeted improvements for diverse visual scenarios. For instance, ISOD designs an extended scale feature pyramid network to boost generic tiny object detection performance [31], while E-FPN further optimizes cross-scale feature fusion specifically for UAV aerial scenes to suppress background interference [32]. Despite their effectiveness, most FPN-based frameworks focus primarily on cross-level semantic aggregation rather than sustained exploitation of shallow high-resolution features. In typical designs, shallow features are incorporated into the fusion pathway only once and are progressively diluted through repeated aggregation with deeper semantic representations. Such one-pass fusion mechanisms may limit the long-term contribution of fine-grained information, which is particularly critical for accurate localization and recognition in aerial small object detection.To address this structural limitation, our work emphasizes progressive reuse and staged integration of shallow features throughout multi-scale fusion, enabling sustained fine-grained information propagation aclross multiple representation levels.

3 Methods

In this section, we present the proposed aerial small object detection framework, termed D3F-Det, which is built upon YOLOv12 [33]. To systematically address fine-grained information degradation and underutilization throughout downsampling, same-scale feature modeling, and multi-scale fusion, D3F-Det adopts a unified architectural design composed of three tightly coupled modules: a dual-branch detail-preserving downsampler (DBDown), a dual-feature multi-scale collaborative module (DFMS), and a three-stage reuse neck (3S-RN).Specifically, DBDown is employed in the early backbone stages to replace conventional stride-2 downsampling operations, enabling effective scale compression while preserving shallow fine-grained features. Building upon the preserved high-resolution representations, DFMS is embedded into same-scale backbone feature layers to enhance spatial and semantic feature modeling through dual-feature collaboration and multi-receptive-field design. Furthermore, a three-stage reuse neck (3S-RN) is designed as an alternative to traditional FPN/PAN structures, progressively reutilizing shallow high-resolution features and continuously injecting fine-grained information into the multi-scale fusion process. As illustrated in Figure 3, these components are integrated into a unified framework to collaboratively enhance small object detection performance.

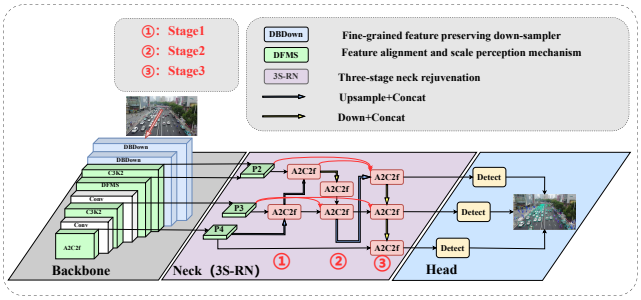


Fig. 3: D3F-Det model architecture.

3.1 Dual-Branch Detail-Preserving Downsampler

Shallow high-resolution features are particularly critical for aerial small object detection, whereas downsampling operations in backbone networks are often a primary

source of fine-grained information degradation. Conventional stride-2 convolutional downsampling achieves scale compression by directly reducing spatial sampling points, which can lead to irreversible loss of key structural information of small objects, such as edges and textures, thereby weakening subsequent feature modeling capability. To alleviate this issue, we propose a dual-branch detail-preserving downsampling module (DBDown) to replace the early convolutional downsampling structures in the backbone network. As shown in Figure 4, DBDown combines pixel-level structured reorganization with parallel dual-branch modeling, enabling spatial scale compression while preserving original spatial structures to the greatest extent. This design further enhances shallow feature representations and provides high-quality inputs for subsequent feature alignment and multi-scale fusion.

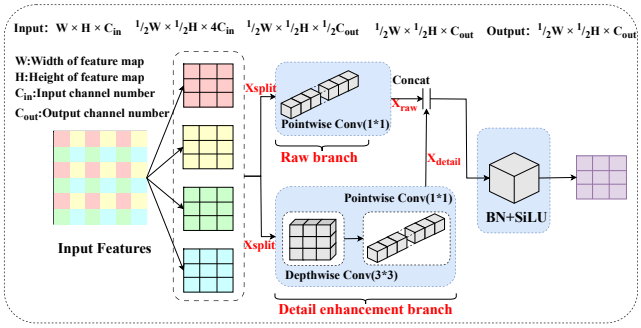


Fig. 4: Structure of the Dual-Branch Fine-grained Feature Preserving Downsampling Module (DBDown).

Pixel-Level Structured Reorganization. Pixel-level structured reorganization serves as the fundamental operation of DBDown, with the core objective of reducing spatial resolution while avoiding the fine-grained information loss caused by direct pixel discarding in conventional downsampling. Unlike stride-based sampling strategies, this operation performs pixel-level spatial rearrangement on the input feature maps, fully preserving the original spatial distribution of pixels and thereby providing a richer fine-grained information basis for subsequent feature modeling. Specifically, a PixelUnShuffle operation is employed to reorganize the input feature map by rearranging each local 2×2 pixel block into the channel dimension, which halves the spatial resolution while correspondingly expanding the channel dimension. The computation can be formulated as follows:

$$X_{\text{split}} = \text{PixelUnShuffle}(2, X_{\text{in}}) \tag{1}$$

where $X_{\text{in}} \in \mathbb{R}^{B\times C_1\times H\times W}$ denotes the input feature map of the module, and $X_{\text{split}} \in \mathbb{R}^{B\times 4C_1\times \frac{H}{2}\times \frac{W}{2}}$ represents the reorganized feature representation after pixel rearrangement. This operation achieves structured compression of spatial information through channel reorganization rather than direct pixel discarding, thereby effectively mitigating the degradation of fine-grained information.

Dual-Branch Parallel Feature Modeling. After pixel reorganization, DBDown employs a dual-branch parallel feature modeling structure to perform complementary modeling on the reorganized features, aiming to preserve fundamental spatial structures while enhancing local discriminative representations. This structure consists of an original information preservation branch and a fine-grained information enhancement branch.

(1) Original information preservation branch. This branch is designed to maximally retain the original spatial structural information after pixel reorganization, while avoiding the introduction of additional changes to the spatial receptive field. Specifically, it applies only a 1×1 pointwise convolution to compress the channel dimension, directly propagating shallow features while maintaining the spatial layout unchanged. The computation process is formulated as follows:

$$X_{\text{raw}} = W_{\text{raw}} * X_{\text{split}} \tag{2}$$

where $W_{\text{raw}} \in \mathbb{R}^{\frac{C_2}{2}\times 4C_1\times 1\times 1}$ denotes a pointwise convolution kernel used to perform channel compression without introducing changes to the spatial receptive field.

(2) Fine-grained information enhancement branch. This branch focuses on modeling local fine-grained features to enhance the discriminative capability of small objects. To maintain computational efficiency, a depthwise separable convolutional structure is adopted in this branch, where depthwise convolution is employed to capture local spatial patterns, followed by pointwise convolution for channel mapping. The computation process is formulated as follows:

$$X_{\text{detail}} = W_{\text{pw}} * \varphi\big(\text{BN}(W_{\text{dw}} * X_{\text{split}})\big) \tag{3}$$

where W_{dw} and W_{pw} denote the depthwise and pointwise convolution kernels, respectively, BN(·) represents the batch normalization layer, and $\varphi(\cdot)$ denotes the SiLU activation function.

Feature Fusion and Output Generation. The outputs of the two branches are concatenated along the channel dimension and further integrated through batch normalization and nonlinear activation to generate the final downsampled feature representation. The computation process is formulated as follows:

$$X_{\text{out}} = \varphi\big(\text{BN}(\text{Concat}(X_{\text{raw}}, X_{\text{detail}}))\big) \tag{4}$$

where Concat(·) denotes channel-wise concatenation of feature maps, BN(·) represents batch normalization, and $\varphi(\cdot)$ denotes the SiLU activation function.

Through the above design, DBDown achieves spatial resolution reduction while effectively preserving and enhancing shallow fine-grained features, thereby providing more stable and discriminative feature representations for subsequent feature alignment modules and the three-stage reuse neck network.

3.2 Dual-Feature Multi-Scale Collaborative Module

In aerial small object detection, same-scale features often simultaneously encode precise spatial localization cues and high-level semantic information. However, these two types of features exhibit inherently different representational emphases, and their direct fusion may lead to feature misalignment. Moreover, a single receptive field is insufficient to accommodate the representation requirements of objects at different scales. To address these issues, this paper proposes a dual-feature multi-scale collaborative module (DFMS), which is designed to achieve spatial–semantic alignment within same-scale feature layers while enhancing multi-scale perception capability. As shown in Figure 5, the overall structure of DFMS follows a unified design that integrates dual-feature collaboration with multi-receptive-field modeling.

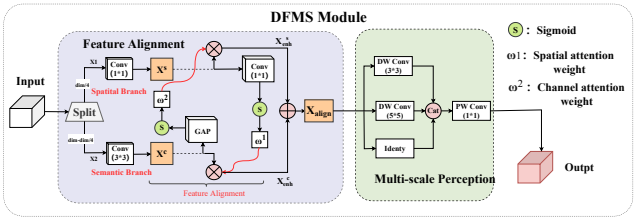


Fig. 5: Structure of the Dual-Feature Multi-Scale Collaboration Module (DFMS).

Same-scale Feature Alignment. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, DFMS first splits the feature along the channel dimension into a semantic branch and a spatial branch. To enable directional modeling of the two feature types, the semantic and spatial branches are transformed using different convolutional operators, while their channel dimensions are unified.

$$(X^1, X^2) = \mathrm{Split}(X_{\mathrm{in}}) \tag{5}$$

$$X^c = \mathcal{C}_{3\times3}(X^1), \quad X^s = \mathcal{C}_{1\times1}(X^2) \tag{6}$$

where $X_{\mathrm{in}} \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map. X^1 and X^2 are the two feature sub-branches obtained by channel-wise splitting, and X^c and X^s denote the outputs of the two branches after convolutional transformations. $\mathcal{C}_{3\times3}(\cdot)$ and $\mathcal{C}_{1\times1}(\cdot)$ represent standard convolution operators with kernel sizes of 3×3 and 1×1, respectively.

Building upon the branch-specific feature transformations, DFMS further introduces a lightweight bidirectional complementary mechanism to enable collaborative alignment and enhancement between semantic and spatial features. Specifically, semantic features are used to reweight the spatial branch along the channel dimension, while spatial features are employed to refine the semantic branch in the spatial dimension, resulting in mutually enhanced feature representations. The process can be formulated as follows:

$$\omega_1 = \sigma\big(\mathcal{C}_{1\times1}(\mathcal{C}_{1\times1}(\mathrm{GAP}(X^c)))\big) \tag{7}$$

$$X^s_{\mathrm{enh}} = X^s \otimes \omega_1 \tag{8}$$

$$\omega_2 = \sigma\big(\mathcal{C}_{1\times1}(X^s)\big) \tag{9}$$

$$X^c_{\mathrm{enh}} = X^c \otimes \omega_2 \tag{10}$$

where GAP(·) denotes global average pooling, $\sigma(\cdot)$ represents the sigmoid activation function, $\omega_1 \in \mathbb{R}^{C\times1\times1}$ and $\omega_2 \in \mathbb{R}^{1\times H\times W}$ denote the channel-wise and spatial attention weights, respectively, and $\otimes$ denotes element-wise multiplication.

Finally, the two enhanced feature representations are aggregated via element-wise addition, yielding a same-scale feature representation with well-aligned spatial and semantic information.

$$X_{\mathrm{align}} = X^c_{\mathrm{enh}} \oplus X^s_{\mathrm{enh}} \tag{11}$$

where $\oplus$ denotes element-wise addition, and X_{align} denotes the final same-scale feature representation with aligned spatial and semantic information.

Multi-scale Perception and Feature Fusion. After obtaining the aligned feature representation X_{align}, DFMS further enhances multi-scale perception at the same feature scale by employing a multi-branch convolutional structure. To capture both small-scale local details and mid-scale contextual information while controlling computational cost, two depthwise separable convolution branches and one identity mapping branch are adopted.

$$Y_1 = \mathcal{C}^{3\times3}_{\mathrm{dw}}(X_{\mathrm{align}}), \quad Y_2 = \mathcal{C}^{5\times5}_{\mathrm{dw}}(X_{\mathrm{align}}), \quad Y_3 = I(X_{\mathrm{align}})$$

(12)

where $\mathcal{C}_{\mathrm{dw}}^{3\times3}(\cdot)$ and $\mathcal{C}_{\mathrm{dw}}^{5\times5}(\cdot)$ denote depthwise separable convolutions with kernel sizes of 3×3 and 5×5, respectively, and $I(\cdot)$ denotes the identity mapping used to preserve the original aligned features.

Subsequently, the outputs of all branches are concatenated along the channel dimension and integrated through a 1×1 pointwise convolution for channel compression and feature reorganization. The resulting feature is then residually fused with X_{align} to produce the final output of the DFMS module.

$$Y = X_{\mathrm{align}} \oplus \mathcal{C}_{1\times1}(\mathrm{Concat}(Y_1, Y_2, Y_3)) \tag{13}$$

where $\mathrm{Concat}(\cdot)$ denotes channel-wise concatenation, $\mathcal{C}_{1\times1}(\cdot)$ denotes a 1×1 pointwise convolution, $\oplus$ represents element-wise addition, and $Y \in \mathbb{R}^{C\times H\times W}$ denotes the output feature of the DFMS module.

3.3 Three-Stage Reuse Neck

Conventional neck networks based on FPN or PAN typically perform a one-time fusion of shallow high-resolution features output by the backbone. During successive scale transformations and feature propagation, shallow fine-grained information is prone to gradual degradation, which limits small object detection performance. To address this issue, this paper proposes a Three-Stage Reuse Neck (3S-RN). As shown in Figure 6, 3S-RN adopts a multi-stage cascading design that repeatedly introduces and reuses original shallow features from the backbone across different scale levels, enabling high-resolution spatial details to be continuously preserved throughout the feature fusion process and progressively aligned with deep semantic information. As a result, spatial–semantic consistency is effectively enhanced.

Stage 1: Initial Feature Alignment and Basic Reuse. The first stage aims to establish an initial correspondence between deep semantic features and shallow high-resolution features. Specifically, the intermediate-to-deep features output by the backbone are taken as the starting point and upsampled to match the spatial resolution of the shallow features. They are then fused with the original backbone features through channel-wise concatenation, resulting in an initially enhanced feature representation. This process can be formulated as follows:

$$P_3^{(1)} = \mathcal{F}(\mathrm{Up}(P_4),\, P_3) \tag{14}$$

$$P_2^{(1)} = \mathcal{F}\Big(\mathrm{Up}(P_3^{(1)}),\, P_2\Big) \tag{15}$$

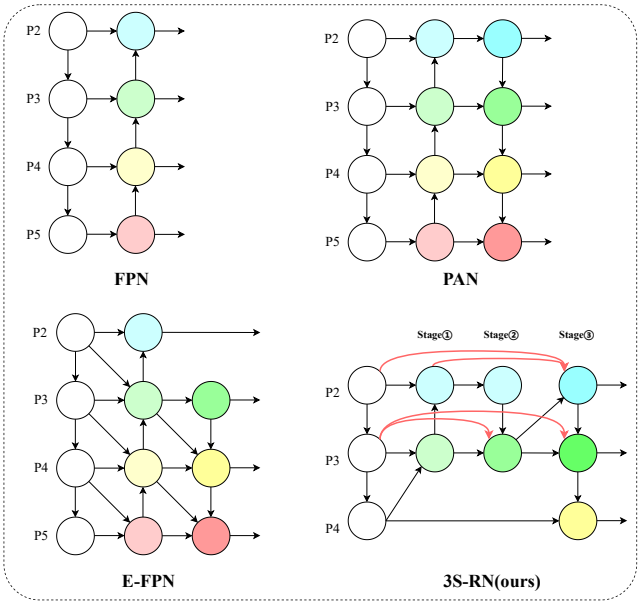


Fig. 6: Structural comparison between conventional feature pyramids and the proposed three-stage reuse neck (3S-RN).

where P_2, P_3, and P_4 denote the shallow, intermediate, and deep feature maps output by the backbone network, respectively; $\mathrm{Up}(\cdot)$ denotes the upsampling operation; $\mathcal{F}(\cdot)$ represents the feature fusion operator composed of channel-wise concatenation followed by the lightweight feature extraction module A2C2f; and $P_i^{(1)}$ denotes the enhanced feature generated at the first stage.

Stage 2: Cross-Stage Interaction and Secondary Reuse. The second stage further strengthens the propagation of shallow fine-grained information toward intermediate semantic features through cross-stage feature interaction. Specifically, the shallow enhanced features generated in Stage 1 are downsampled to the intermediate scale and jointly modeled with the original intermediate features from the backbone as well as the enhanced intermediate features obtained in Stage 1. Subsequently, the refined intermediate features are fed back to the shallow scale, enabling secondary reuse of shallow features.

$$P_3^{(2)} = \mathcal{F}\Big(\mathrm{Down}(P_2^{(1)}),\, P_3^{(1)},\, P_3\Big) \tag{16}$$

$$P_2^{(2)} = \mathcal{F}\Big(\mathrm{Down}(P_3^{(2)}),\, P_2^{(1)},\, P_2\Big) \tag{17}$$

where $\mathrm{Down}(\cdot)$ denotes the downsampling operation, $\mathcal{F}(\cdot)$ represents the feature fusion operator, and $P_i^{(2)}$ denotes the enhanced feature generated at the second stage.

Table 1: Experimental environment and training configurations.

Parameter	Value
Framework	PyTorch 2.1, Python 3.10, CUDA 12.1
GPU	NVIDIA GeForce RTX 4090
Epochs	200
Batch size	12
Input size	640×640
Optimizer	SGD
Learning rate	0.01 (cosine scheduler, warmup 10 epochs)
Momentum	0.937
Weight decay	5×10^{-4}

Stage 3: Multi-scale Integration and Final Output. The third stage performs unified integration of the enhanced features generated in the previous two stages to construct multi-scale feature representations suitable for the detection head. Specifically, the shallow enhanced features obtained in Stage 2 are downsampled and fused with the intermediate enhanced features as well as the original intermediate features from the backbone. The resulting representations are then propagated toward deeper feature levels to further enhance semantic information.

$$P_3^{(3)} = \mathcal{F}\Big(\mathrm{Down}(P_2^{(2)}),\, P_3^{(2)},\, P_3\Big) \tag{18}$$

$$P_4^{(3)} = \mathcal{F}\Big(\mathrm{Down}(P_3^{(3)}),\, P_4\Big) \tag{19}$$

Finally, 3S-RN outputs a set of multi-scale feature representations, which are used by the subsequent detection head for prediction.

$$\Big\{P_2^{(2)},\, P_3^{(3)},\, P_4^{(3)}\Big\} \tag{20}$$

4 Experiments

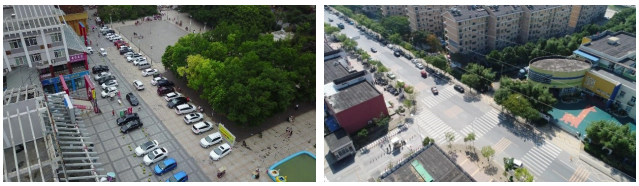
4.1 Implementation Details

All experiments are conducted based on the Ultralytics YOLO framework. The experimental environment and training configurations are summarized in Table 1. All models are trained and evaluated under identical hardware platforms and training settings to ensure fair and comparable evaluation.

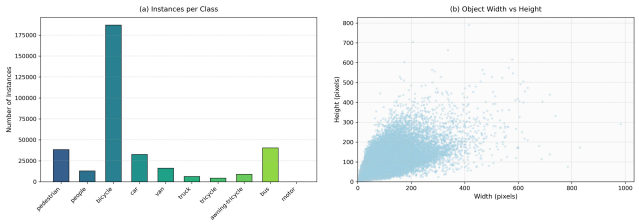
4.2 Datasets

Three representative UAV-based aerial object detection datasets are employed to systematically evaluate the proposed method, including VisDrone2019-DET [34], AI-TOD [35], and TinyPerson [36]. Among these datasets, VisDrone2019-DET is adopted as the core benchmark for performance comparison and ablation analysis, whereas AI-TOD and TinyPerson are further utilized to validate the generalization ability of our method under diverse scene conditions and extremely small object scenarios.

VisDrone2019. The VisDrone2019 dataset, collected by the AISKYEYE team at Tianjin University, is a widely used benchmark for UAV-based small object detection. It contains a total of 8,629 high-resolution images, with 6,471 images for training, 548 images for validation, and the remaining images reserved for testing. The images are captured across diverse urban environments and flight altitudes, covering typical UAV scenarios such as city streets, traffic intersections, and complex backgrounds. Due to variations in aerial viewpoints and imaging distances, object scales exhibit significant changes, with small objects accounting for a large proportion of the dataset, which poses considerable challenges for multi-scale modeling and fine-grained feature representation. The dataset includes annotations for 10 common object categories, such as pedestrians, vehicles, and cyclists. As shown in Figure 7, representative images and the distribution of annotations illustrate the characteristics of the dataset.



(a) VisDrone2019 dataset samples.



(b) Category distribution and object size density statistics of VisDrone2019.

Fig. 7: Overview and statistical characteristics of the VisDrone2019 dataset. (a) Representative aerial images from the VisDrone2019 dataset illustrating complex urban scenes. (b) Category distribution and object size density, highlighting the predominance of small-scale objects.

AI-TOD. AI-TOD is a publicly available dataset specifically designed for tiny object detection in remote

sensing scenarios, consisting of a large number of high-resolution remote sensing images. The dataset contains 11,214 training images, 2,804 validation images, and 14,018 test images, with a unified image resolution of 800×800 pixels, covering eight typical remote sensing object categories. Compared with most existing remote sensing datasets, objects in AI-TOD are significantly smaller in scale, with an average object size of approximately 12.8 pixels, and a large proportion of instances falling into the extremely small-scale regime. This characteristic poses substantial challenges for fine-grained feature modeling and tiny object perception. Therefore, AI-TOD is well suited for evaluating the effectiveness of the proposed method in extremely small object detection scenarios. As shown in Figure 8(a), representative examples highlight the prevalence of extremely small and densely distributed objects in this dataset.

TinyPerson. TinyPerson is a dataset designed for extremely small-scale object detection, primarily collected from long-distance aerial video scenarios and containing a large number of small and densely distributed pedestrian instances. The dataset is constructed from frames extracted from the same video sequences, exhibiting characteristics such as small object sizes, a high number of instances, and a relatively uniform scale distribution. These properties pose significant challenges for small object perception and fine-grained feature modeling under complex backgrounds. In this work, the proposed method is evaluated on the TinyPerson dataset following its official data split. As shown in Figure 8(b), representative examples illustrate the dense distribution and extreme scale characteristics of pedestrian instances in real-world aerial scenes.



(a) AI-TOD dataset samples.



(b) TinyPerson dataset samples.

Fig. 8: Representative examples from the AI-TOD and TinyPerson datasets. (a) Example images from the AI-TOD dataset. (b) Example images from the TinyPerson dataset.

4.3 Evaluation Indicators

To comprehensively evaluate the performance of the proposed method on UAV-based small object detection tasks, several standard evaluation metrics are adopted, including Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) for detection accuracy assessment. In addition, the number of parameters (Params) and computational complexity measured by GFLOPs are reported to evaluate model complexity and inference efficiency. Among these metrics, mAP@0.5 and mAP@0.5:0.95 are used as the primary evaluation criteria. The definitions of these evaluation metrics are given as follows:

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$AP = \int_0^1 P(R)\, dR \tag{23}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{24}$$

$$\text{Parameters} = \sum_{l=1}^{L} \left(C_{\text{in}}^{(l)} \times K_h^{(l)} \times K_w^{(l)} \times C_{\text{out}}^{(l)} + C_{\text{out}}^{(l)} \right) \tag{25}$$

$$\text{GFLOPs} = \frac{1}{10^9} \sum_{l=1}^{L} \left(2 \times C_{\text{in}}^{(l)} \times K_h^{(l)} \times K_w^{(l)} \times C_{\text{out}}^{(l)} \times H_{\text{out}}^{(l)} \times W_{\text{out}}^{(l)} \right) \tag{26}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively; N is the number of object categories; L denotes the number of network layers; $C_{\text{in}}^{(l)}$ and $C_{\text{out}}^{(l)}$ represent the input and output channel numbers of the l-th layer; $K_h^{(l)} \times K_w^{(l)}$ denotes the kernel size; and $H_{\text{out}}^{(l)} \times W_{\text{out}}^{(l)}$ denotes the spatial resolution of the output feature map.

4.4 Results on VisDrone2019 Dataset.

Comparison with State-of-the-art Methods. As shown in Table 2, the proposed D3F-Det is compared with mainstream YOLO-series real-time detectors as well as representative end-to-end DETR-based methods on the VisDrone2019 dataset to validate its effectiveness. Taking YOLOv12-s as the baseline, D3F-Det improves mAP@0.5:0.95 from 24.1% to 30.3%, achieving a gain of 6.2 percentage points, and boosts mAP@0.5 from 39.2% to 49.5%, corresponding to an improvement of 10.3 percentage points. Meanwhile, the number of parameters is significantly reduced from 9.3M to 3.7M, resulting in a substantially lighter model while maintaining superior detection accuracy. Compared with end-to-end DETR-based detectors, D3F-Det also demonstrates strong competitiveness on VisDrone2019, consistently outperforming the vanilla DETR and its representative variants in terms of both mAP@0.5 and mAP@0.5:0.95. In addition, D3F-Det preserves clear advantages in parameter size and computational complexity, indicating a favorable balance between detection accuracy and efficiency for UAV-based scenarios.

Convergence Analysis on VisDrone2019. To further analyze the optimization behavior, Figure 9 shows the training curves of Precision, Recall, mAP@0.5, and mAP@0.5:0.95 on the VisDrone2019 dataset. Compared with the baseline, D3F-Det converges faster and consistently achieves higher performance across all metrics. Notably, it exhibits steeper early-stage improvement and converges to a higher performance plateau, indicating improved optimization efficiency and training stability.

Qualitative Results. To further validate the effectiveness of the proposed method for UAV-based small object detection, qualitative visualizations of detection results and the corresponding multi-scale detection head feature response heatmaps are provided on the VisDrone2019 dataset. As shown in Figure 10, compared with the baseline model, the proposed method exhibits more compact and cleaner feature response distributions, while spurious activations in complex background regions are effectively suppressed. This reduction in background noise enables the model to focus more accurately on target regions, resulting in more precise object localization and bounding box regression that are highly consistent with the ground-truth annotations. These observations indicate that the proposed method enhances background suppression and improves the reliability of small object perception and localization in complex UAV-captured scenes.

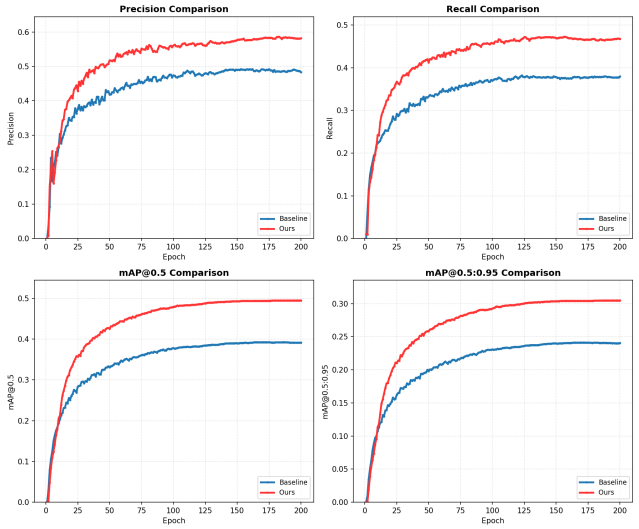


Fig. 9: Training dynamics comparison between the baseline model and D3F-Det on the VisDrone2019 dataset. The curves show the evolution of Precision, Recall, mAP@0.5, and mAP@0.5:0.95 over training epochs. D3F-Det demonstrates faster convergence and consistently higher performance across all metrics, indicating improved optimization efficiency and training stability.

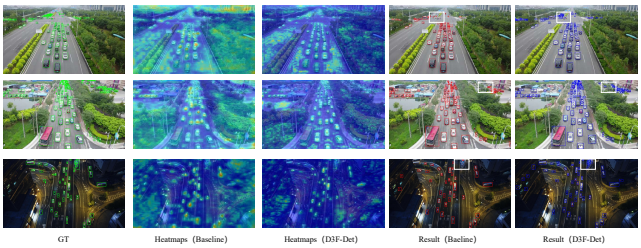


Fig. 10: Qualitative comparison of feature response heatmaps and detection results on the VisDrone2019 dataset. Compared with the baseline, D3F-Det produces more compact and target-focused feature responses while effectively suppressing background noise, leading to more accurate localization and detection of small objects in complex UAV-captured scenes.

Table 2: Quantitative comparison of the proposed D3F-Det with mainstream YOLO-series real-time detectors and representative end-to-end detectors on the VisDrone2019 dataset.

Model	Input size	mAP@0.5↑	mAP@0.5:0.95↑	Params (M)↓	GFLOPs↓
Real-time object detectors					
YOLOv8-s [37]	640 × 640	39.6	23.6	11.2	28.6
YOLOv8-m [37]	640 × 640	40.7	24.6	25.9	78.9
YOLOv9-s [38]	640 × 640	38.3	22.7	7.2	26.7
YOLOv9-m [38]	640 × 640	46.7	25.2	20.1	76.8
YOLOv10-s [39]	640 × 640	39.3	23.8	7.2	21.6
YOLOv10-m [39]	640 × 640	44.8	27.5	15.4	59.1
YOLOv11-s [40]	640 × 640	40.4	24.2	9.4	21.3
YOLOv11-m [40]	640 × 640	46.7	28.9	20.1	68.0
YOLOv12-s(2025) [33]	640 × 640	39.2	24.1	9.3	21.7
YOLOv13-s(2025) [41]	640 × 640	39.1	23.4	9.0	20.8
FO-YOLO(2025) [42]	640 × 640	43.9	24.6	24.5	76.8
MFA-YOLO-m(2025) [43]	640 × 640	45.0	25.5	24.2	100.6
LCW-YOLO(2025) [44]	640 × 640	49.3	30.6	19.8	68.5
End-to-end object detectors					
DETR [45]	1333 × 750	40.1	24.1	60.0	187
Deformable-DETR [28]	1333 × 800	42.2	27.1	40.0	173
Sparse-DETR [46]	1333 × 800	42.5	27.3	40.9	121
RT-DETR-R18(2024) [47]	640 × 640	44.6	26.7	20.0	60
RT-DETR-R50(2024) [47]	640 × 640	47.0	28.4	42.0	136
UAV-DETR-R18(2025) [48]	640 × 640	48.8	29.8	20.0	77
ALDNet(2025) [49]	640 × 640	48.1	29.6	15.1	51.0
SO-DETR(2025) [50]	640 × 640	49.0	29.9	19.9	53.7
Proposed UAV detector					
D3F-Det (ours)	640 × 640	**49.5**	**30.3**	**3.7**	**44.0**

4.5 Results on AI-TOD Dataset.

Comparison with State-of-the-art Methods. The AI-TOD dataset contains a large number of extremely small objects and high-density scenes, posing higher demands on the fine-grained feature representation capability of detection models. To further evaluate the effectiveness of the proposed method for aerial tiny object detection, we conduct comparative experiments on the AI-TOD dataset, and the results are reported in Table 3. As observed, D3F-Det achieves 54.9% mAP@0.5 and 28.2% mAP@0.5:0.95, yielding the best performance among all compared methods and consistently outperforming existing YOLO-series models as well as other advanced detectors. These results indicate that the proposed method exhibits stronger robustness and feature modeling capability in densely distributed tiny object scenarios.

Qualitative Results. To further analyze the performance of the proposed method in extremely small object detection scenarios, qualitative visualizations of detection results on the AI-TOD dataset are provided. As shown in Figure 11, D3F-Det accurately detects the majority of tiny objects in challenging scenes characterized by extremely small object sizes and dense distributions. Compared with the baseline model, the proposed method exhibits more stable detection results in local regions, effectively reducing missed detections while achieving more consistent bounding box localization in densely populated and low-contrast areas. These qualitative observations demonstrate that D3F-Det exhibits enhanced robustness and discriminative capability in densely distributed tiny object scenarios.

4.6 Results on Tinyperson Dataset

Comparison with State-of-the-art Methods. As reported in Table 4, the proposed D3F-Det is compared with various mainstream object detection methods on the TinyPerson dataset. The results show that D3F-Det consistently outperforms most existing YOLO-series detectors in terms of detection accuracy. Taking YOLOv12-s as the baseline, D3F-Det improves mAP@0.5 from 26.0% to 33.0% and mAP@0.5:0.95 from 8.1% to 10.5%, achieving clear performance gains in extremely small-scale pedestrian detection scenarios. Meanwhile, the number of parameters is reduced from 9.2M to 3.7M, resulting in a more favorable accuracy–complexity trade-off under comparable computational cost. These results demonstrate the effectiveness and practicality of the proposed method on the TinyPerson dataset.

Table 3: Quantitative comparison of different methods on the AI-TOD dataset.

Model	mAP@0.5↑	mAP@0.5:0.95↑	AP_s↑	AP_m↑
YOLOv12-s [33]	35.6	16.2	5.4	10.1
YOLOv12-l [33]	51.1	23.1	28.6	36.4
Faster R-CNN [51]	27.1	9.5	8.8	21.4
Swin Transformer [52]	34.6	13.5	5.7	9.2
CDMNet [53]	41.2	16.8	16.8	29.4
GLSAN [54]	42.7	17.7	18.9	30.8
RT-DETR(2024) [47]	48.3	19.8	19.7	32.8
PP-YOLOE-SOD [55]	50.1	21.2	20.2	35.2
HANet [56]	53.7	22.1	27.3	36.8
FSANet-Swin-T [57]	52.8	22.6	29.1	38.5
SCASNet(2025) [58]	54.2	24.6	21.2	38.2
MD-DRIFPN(2025) [59]	50.9	24.8	33.5	42.2
HAD-YOLO(2025) [60]	52.6	21.4	32.3	33.6
FANet(2025) [61]	52.4	21.4	26.4	33.8
TBNet(2025) [62]	54.3	23.0	28.5	32.6
D3F-Det (ours)	**54.9**	**28.2**	**25.6**	**39.7**

Table 4: Quantitative comparison with state-of-the-art methods on the TinyPerson dataset.

Model	mAP@0.5↑	mAP@0.5:0.95↑	Params (M)↓	GFLOPs↓
YOLOv5-s [63]	23.6	7.3	9.1	23.8
YOLOv8-s [37]	23.9	7.5	11.1	28.5
YOLOv9-s [38]	24.6	7.4	7.2	26.7
YOLOv9-m [38]	26.3	8.6	20.1	76.6
YOLOv10-s [39]	20.3	6.3	8.1	24.5
YOLOv10-m [39]	26.0	8.1	15.3	58.9
YOLOv11-s [40]	25.6	8.0	9.4	21.6
YOLOv11-m [40]	27.3	8.8	20.0	68.2
YOLOv12-s(2025) [33]	26.0	8.1	9.2	21.5
YOLOv12-m(2025) [33]	26.6	8.4	20.1	67.7
MSEF-YOLO11s(2026) [64]	28.1	9.2	13.8	34.5
D3F-Det (ours)	**33.0**	**10.5**	**3.7**	**44.0**

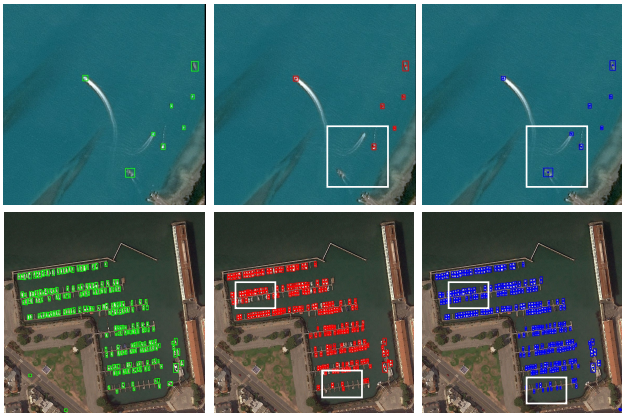


Fig. 11: Qualitative comparison of detection results on the AI-TOD dataset. Compared with the baseline, D3F-Det detects more tiny objects with fewer missed detections and achieves more consistent localization in dense scenes.

Qualitative Results. To further validate the effectiveness of the proposed method in extremely small-scale pedestrian detection scenarios, qualitative visualizations are conducted on the TinyPerson dataset. As shown in Figure 12, under long-distance UAV viewpoints and highly crowded scenes, the baseline model tends to suffer from missed detections and unstable localization, particularly when pedestrians are extremely small or partially occluded. In contrast, D3F-Det achieves more accurate localization and exhibits improved detection completeness and consistency in densely populated regions.

4.7 Ablation experiments

We conduct ablation experiments on the VisDrone2019-DET dataset using YOLOv12-s as the baseline to validate the core components of the proposed method.

Effect of Key Components. In this ablation study, the proposed modules are progressively integrated into the baseline model to evaluate their individual contributions. As shown in Table 5, the detection perfor-

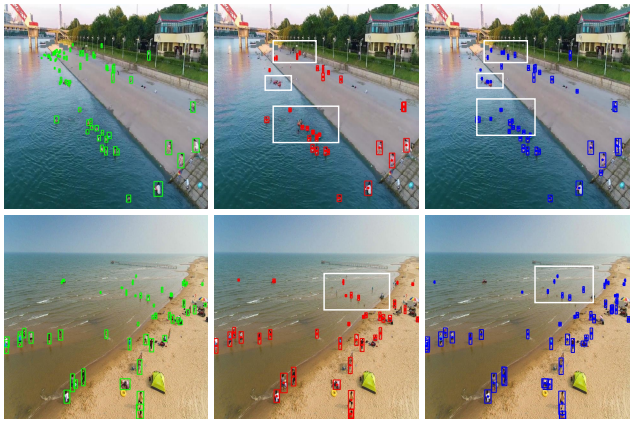


Fig. 12: Qualitative comparison of detection results on the TinyPerson dataset. Compared to the baseline model, our D3F - Det method demonstrates a notable capability in detecting a larger quantity of minuscule pedestrian objects, effectively minimizing missed detections.

mance consistently improves in terms of both mAP@0.5 and mAP@0.5:0.95 as DBDown, DFMS, and 3S-RN are successively introduced, indicating that these components provide complementary performance gains for small object detection.Notably, the improvement in detection accuracy is achieved with only a marginal increase in model parameters and computational cost, demonstrating that the proposed design achieves a favorable balance between accuracy and efficiency.

Effect of DFMS Placement. To further investigate the effectiveness of DFMS at different feature scales, ablation experiments are conducted by inserting DFMS into same-scale feature layers of 160×160, 80×80, and 40×40, while keeping the rest of the network unchanged. The results are reported in Table 6.It can be observed that deploying DFMS at the high-resolution 160×160 feature level achieves the best overall performance across all evaluation metrics. This indicates that performing spatial–semantic alignment and multi-scale perception at higher-resolution feature maps is more effective for preserving fine-grained details of small objects, thereby maximizing the benefits of the DFMS module.

5 Conclusion

This paper addresses the challenges of extremely small object scales, fine-grained feature degradation, and multi-scale feature misalignment in UAV aerial imagery, and proposes an efficient small object detection framework termed D3F-Det. Built upon YOLOv12, the proposed method systematically enhances small object representation and multi-scale perception from three key stages: downsampling, same-scale feature modeling, and neck-level feature fusion.Specifically, a dual-branch detail-preserving downsampling module (DBDown) is introduced to alleviate the loss of fine-grained information during early downsampling. A dual-feature multi-scale synergy module (DFMS) is then designed to achieve spatial–semantic alignment within same-scale feature layers while enhancing multi-scale modeling capability. Furthermore, a three-stage reuse neck network (3S-RN) is proposed to maximize the utilization of preserved shallow high-resolution features through a progressive cross-stage reuse strategy, thereby continuously reinforcing spatial–semantic consistency during multi-scale feature fusion.Experimental results on the VisDrone2019, AI-TOD, and TinyPerson datasets demonstrate that D3F-Det achieves consistent improvements in key metrics such as mAP@0.5 and mAP@0.5:0.95, while maintaining a favorable balance between detection accuracy and computational complexity. Ablation studies further validate the effectiveness of each proposed component in fine-grained feature preservation, feature alignment, and multi-scale modeling. Future work will explore more lightweight alignment and multi-scale modeling strategies and extend the framework to more complex aerial scenarios and extreme small-object distributions.

Acknowledgements This work was supported by Key Scientific Research Project Plan of Colleges and Universities in Henan Province (25CY039); the Science and Technology Innovation Leading Talent Support Program of Henan Province, China (254200510017); Research Programs of Henan Science and Technology Department (252102210034, 252102210039, 252102220053); Henan Province key research and development project (231111212000); Supported by Henan Center for Outstanding Overseas Scientists (GZS2022011); Youth Fund Project of National Natural Science Foundation of China (62202434); Zhengzhou University of Aeronautics Graduate Education Innovation Program Fund (2025CX121); Collaborative Innovation Center of Aeronautics and Astronautics Electronic Information Technology, Henan Province, Zhengzhou, 450046, Henan, China; Henan Province Aerospace Intelligent Technology and Systems International Joint Laboratory, Zhengzhou 450046, Henan, China; Henan Key Laboratory of General Aviation Technology, Zhengzhou 450046, Henan, China.Open Fund Project of Henan General Aviation Technology Key Laboratory(ZHKF-240205).

References

1. Xiaoling Wu, Li Zhang, Weidong Li, and Jie Wang. Deep learning for UAV-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2022.

Table 5: Ablation study of the proposed components on the VisDrone2019-DET dataset using YOLOv12-s as the baseline.

Setting	DBDown	DFMS	3S-RN	mAP@0.5↑	mAP@0.5:0.95↑	Params (M)↓	GFLOPs↓
Baseline	-	-	-	39.2	24.1	9.3	21.5
+ DBDown	✓	-	-	39.9	24.3	9.3	22.1
+ DFMS	-	✓	-	41.5	25.1	9.5	31.9
+ 3S-RN	-	-	✓	46.4	30.4	3.5	33.0
+ DBDown + DFMS	✓	✓	-	41.9	25.4	9.5	32.5
D3F-Det (full)	✓	✓	✓	**49.5**	**30.3**	**3.7**	**44.0**

Table 6: Ablation study of DFMS placement at different feature scales on the VisDrone2019-DET dataset.

Setting	40×40	80×80	160×160	mAP@0.5↑	mAP@0.5:0.95↑	Params (M)↓	GFLOPs↓
Baseline	-	-	-	39.2	24.1	9.3	21.5
+ DFMS (40×40)	✓	-	-	40.0	23.7	11.3	29.3
+ DFMS (80×80)	-	✓	-	41.1	24.6	11.3	36.9
+ DFMS (160×160)	-	-	✓	**41.5**	**25.1**	9.5	31.9

2. Jiaxu Leng, Yongming Ye, Mengjingcheng Mo, Chenqiang Gao, Ji Gan, Bin Xiao, and Xinbo Gao. Recent advances for aerial object detection: a survey. *ACM Computing Surveys*, 56(12):1–36, 2024.
3. Nadin Habash, Ahmad Abu Alqumsan, and Tao Zhou. Recent real-time aerial object detection approaches, performance, optimization, and efficient design trends for onboard performance: A survey. *Sensors*, 25(24):7563, 2025.
4. Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.
5. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
6. Wei Hua and Qili Chen. A survey of small object detection based on deep learning in aerial images. *Artificial Intelligence Review*, 58(6):1–67, 2025.
7. Junhyung Kang, Shahroz Tariq, Han Oh, and Simon S Woo. A survey of deep learning-based object detection methods and datasets for overhead imagery. *Ieee Access*, 10:20118–20134, 2022.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
9. Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.
10. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
11. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
12. Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.
13. Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.
14. Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
15. Javed Sayyad, Khush Attarde, and Khaled Mohamad Almustafa. A systematic literature review on deep learning approaches for small object detection. *Array*, page 100615, 2025.
16. Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.
17. Zhicheng Zhao, Jiaxin Du, Chenglong Li, Xiang Fang, Yun Xiao, and Jin Tang. Dense tiny object detection: A scene context guided approach and a unified benchmark. *IEEE transactions on geoscience and remote sensing*, 62:1–13, 2024.
18. Zican Shi, Jing Hu, Jie Ren, Hengkang Ye, Xuyang Yuan, Yan Ouyang, Jia He, Bo Ji, and Junyu Guo. Hs-fpn: High frequency and spatial perception fpn for tiny object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6896–6904, 2025.
19. Qin Xu, Yuhui Zhang, Zhicheng Zhao, and Jin Tang. S2fcnet: Semantic and spatial feature compensation network for tiny-object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

20. Peiran Peng, Tingfa Xu, Bo Huang, and Jianan Li. Hafnet: hierarchical attentive fusion network for multispectral pedestrian detection. *Remote Sensing*, 15(8):2041, 2023.
21. Heng Hu, Si-Bao Chen, and Jin Tang. Cfenet: Contextual feature enhancement network for tiny object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.
22. Zhongwen Yu, Qiu Guan, Jianmin Yang, Zhiqiang Yang, Qianwei Zhou, Yang Chen, and Feng Chen. Lsm-yolo: A compact and effective roi detector for medical detection. In *International Conference on Neural Information Processing*, pages 30–44. Springer, 2024.
23. Ying Li, Changfei Zhu, Qiang Zhang, Jianing Zhang, and Guifang Wang. If-yolo: An efficient and accurate detection algorithm for insulator faults in transmission lines. *IEEE Access*, 12:167388–167403, 2024.
24. Shenming Qu, Chaoxu Dang, Wangyou Chen, and Yanhong Liu. Sma-yolo: An improved yolov8 algorithm based on parameter-free attention mechanism and multi-scale feature fusion for small object detection in uav images. *Remote Sensing*, 17(14):2421, 2025.
25. Hongxu Ma, Chenbo Zhang, Lu Zhang, Jiaogen Zhou, Jihong Guan, and Shuigeng Zhou. Fine-grained zero-shot object detection. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.
26. Jian Zhu, Qingwu Zhang, Lunke Fei, Ruichu Cai, Yuan Xie, Bin Sheng, and Xiaokang Yang. Fffn: Frame-by-frame feedback fusion network for video super-resolution. *IEEE Transactions on Multimedia*, 25:6821–6835, 2022.
27. Chuchao Lin, Changjun Zou, and Hangbin Xu. Scnet: A dual-branch network for strong noisy image denoising based on swin transformer and convnext. *Computer Animation and Virtual Worlds*, 36(3):e70030, 2025.
28. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.
29. Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.
30. Hou-I Liu, Yu-Wen Tseng, Kai-Cheng Chang, Pin-Jyun Wang, Hong-Han Shuai, and Wen-Huang Cheng. A denoising fpn with transformer r-cnn for tiny object detection. *IEEE transactions on geoscience and remote sensing*, 62:1–15, 2024.
31. Ping Ma, Xinyi He, Yiyang Chen, and Yuan Liu. Isod: Improved small object detection based on extended scale feature pyramid network. *The Visual Computer*, 41(1):465–479, 2025.
32. Zhongxu Li, Qihan He, and Wenyuan Yang. E-fpn: an enhanced feature pyramid network for uav scenarios detection. *The Visual Computer*, 41(1):675–693, 2025.
33. Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.
34. Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
35. Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *2020 25th international conference on pattern recognition (ICPR)*, pages 3791–3798. IEEE, 2021.
36. Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020.
37. Glenn Jocher, Ayush Chaurasia, Jing Qiu, and Ultralytics Team. Ultralytics yolov8, 2023.
38. Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pages 1–21. Springer, 2024.
39. Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
40. Glenn Jocher and Ultralytics Team. Ultralytics yolov11, 2024.
41. M Lei, S Li, Y Wu, H Hu, Y Zhou, X Zheng, G Ding, S Du, Z Wu, Y Gao, et al. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. arxiv 2025. *arXiv preprint arXiv:2506.17733*, 10, 2025.
42. Huaping Zhou, Wei Yin, Kelei Sun, Tao Wu, and Bin Deng. Fo-yolo for small object detection in drone aerial imagery: H. zhou et al. *The Journal of Supercomputing*, 81(12):1208, 2025.
43. Shuo Li and Chong Chen. Mfa-yolo: a multi-feature aggregation approach for small-object detection method in drone imagery. *Scientific Reports*, 2025.
44. Dan Liao, Rengui Bi, Yubi Zheng, Cheng Hua, Liangqing Huang, Xiaowen Tian, and Bolin Liao. Lcw-yolo: An explainable computer vision model for small object detection in drone images. *Applied Sciences*, 15(17):9730, 2025.
45. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
46. Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.
47. Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024.
48. Huaxiang Zhang, Hao Zhang, Kai Liu, Zhongxue Gan, and Guo-Niu Zhu. Uav-detr: efficient end-to-end object detection for unmanned aerial vehicle imagery. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 15143–15149. IEEE, 2025.
49. Hao Cai, Jinhong Zhang, and Jianlong Xu. Aldnet: a lightweight and efficient drone detection network. *Measurement Science and Technology*, 36(2):025402, 2025.
50. Huaxiang Zhang, Hao Zhang, Aoran Mei, Zhongxue Gan, and Guo-Niu Zhu. So-detr: leveraging dual-domain features and knowledge distillation for small object detection. *arXiv preprint arXiv:2504.11470*, 2025.
51. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

52. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
53. Jia Shi and Chunyu Wang. Cdmnet: Cloud detection in remote sensing images based on cnn. In *Journal of Physics: Conference Series*, volume 2640, page 012013. IOP Publishing, 2023.
54. Sutao Deng, Shuai Li, Ke Xie, Wenfeng Song, Xiao Liao, Aimin Hao, and Hong Qin. A global-local self-adaptive network for drone-view object detection. *IEEE Transactions on Image Processing*, 30:6248–6260, 2021.
55. Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022.
56. Chengxi Han, Chen Wu, Haonan Guo, Meiqi Hu, and Hongruixuan Chen. Hanet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:3867–3878, 2023.
57. Jixiang Wu, Zongxu Pan, Bin Lei, and Yuxin Hu. Fsanet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.
58. Zhenkuan Wang, Xue-Mei Dong, and Yongli Xu. Scasnet: Spatial context-aware selection network for small object detection in aerial imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
59. Houyu Luan, Shuobo Xu, Dishi Xu, Lele Liu, Mengwei Guo, and Shaoqing Huang. Md-drifpn: dilated multi-directional fpn for small drone object detection. *The Journal of Supercomputing*, 81(18):1646, 2025.
60. Yiting Zheng, Xiaobo Liu, Dongsen Zhang, Hongbo Xiang, Lianyi He, and Yan Wang. Had-yolo: Small object detection in remote sensing images with hatafpn and dynamic receptive fields. In *2025 44th Chinese Control Conference (CCC)*, pages 8755–8760. IEEE, 2025.
61. Wen Zixiao, Peifeng Li, Yuhan Liu, Jingming Chen, Xiang Xiantai, Yuan Li, Wang Huixian, Zhao Yongchao, and Zhou Guangyao. Fanet: Frequency-aware attention-based tiny-object detection in remote sensing images. *Remote Sensing*, 17(24):4066, 2025.
62. Zheng Li, Yongcheng Wang, Dongdong Xu, Yunxiao Gao, and Tianqi Zhao. Tbnet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery. *Pattern Recognition*, 158:110976, 2025.
63. Glenn Jocher and Ultralytics Team. Yolov5, 2020. Accessed: 2026-02-05.
64. Kai Zhang, Pengcheng Zhang, Farhan Ullah, and Yue Zhao. Msef-yolo11s: a multi-scale extraction and fusion network for small target detection in drone imagery. *The Journal of Supercomputing*, 82(2):72, 2026.