

Article

FDE-YOLO: An Improved Algorithm for Small Target Detection in UAV Images

Jialiang Li, Xu Guo, Xu Zhao and Jie Jin *

College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; m320123316@sues.edu.cn (J.L.)

* Correspondence: 02140001@sues.edu.cn

Abstract

Accurate small object detection in unmanned aerial vehicle (UAV) imagery is fundamental to numerous safety-critical applications, including intelligent transportation, urban surveillance, and disaster assessment. However, extreme scale compression, dense object distributions, and complex backgrounds severely constrain the feature representation capability of existing detectors, leading to degraded reliability in real-world deployments. To overcome these limitations, we propose FDE-YOLO, a lightweight yet high-performance detection framework built upon YOLOv11 with three complementary architectural innovations. The Fine-Grained Detection Pyramid (FGDP) integrates space-to-depth convolution with a CSP-MFE module that fuses multi-granularity features through parallel local, context, and global branches, capturing comprehensive small target information while avoiding computational overhead from layer stacking. The Dynamic Detection Fusion Head (DDFHead) unifies scale-aware, spatial-aware, and task-aware attention mechanisms via sequential refinement with DCNv4 and FReLU activation, adaptively enhancing discriminative capability for densely clustered targets in complex scenes. The EdgeSpaceNet module explicitly fuses Sobel-extracted boundary features with spatial convolution outputs through residual connections, recovering edge details typically lost in standard operations while reducing parameter count via depthwise separable convolutions. Extensive experiments on the VisDrone2019 dataset demonstrate that FDE-YOLO achieves 53.6% precision, 42.5% recall, 43.3% mAP50, and 26.3% mAP50:95, surpassing YOLOv11s by 2.8%, 4.4%, 4.1%, and 2.8% respectively, with only 10.25 M parameters. The proposed approach outperforms UAV-specialized methods including Drone-YOLO and MASF-YOLO while using significantly fewer parameters (37.5% and 29.8% reductions respectively), demonstrating superior efficiency. Cross-dataset evaluations on UAV-DT and NWPU VHR-10 further confirm strong generalization capability with 1.6% and 1.5% mAP50 improvements respectively, validating FDE-YOLO as an effective and efficient solution for reliable UAV-based small object detection in real-world scenarios.



Academic Editors: Jorge Brieva and Ernesto Moya-Albor

Received: 19 January 2026

Revised: 8 February 2026

Accepted: 11 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: YOLOv11; UAV aerial images; small target detection; multi-scale feature fusion; object detection

MSC: 68T45; 68T07

1. Introduction

With the rapid advancement of computer technology, particularly artificial intelligence (AI) and machine learning (ML), data-driven approaches have been widely adopted

across various industries to improve efficiency, prediction accuracy, and decision-making processes [1]. Building upon these technological advancements, unmanned aerial vehicle (UAV) technology has experienced rapid development, with applications spanning agricultural monitoring [2], disaster relief, traffic inspection [3], and border security [4]. UAV aerial imagery has emerged as a critical data source for large-scale surface observation due to its flexibility, wide field of view, and cost-effectiveness [5]. However, aerial images often suffer from challenges such as small target sizes, complex backgrounds, and densely distributed objects. Accurate detection of small targets—particularly vehicles, pedestrians, and small facilities—remains a fundamental challenge that limits the intelligent perception capabilities of UAV systems [6,7].

Currently, object detection algorithms can be divided into two categories: traditional detection algorithms based on manual feature extraction [8,9] and object detection algorithms based on deep learning, which possess powerful feature expression capabilities and generalization, representing the current mainstream research direction. Deep learning-based object detection algorithms can be subdivided into two-stage algorithms (representative algorithms include Fast R-CNN [10,11] and Mask R-CNN [12,13]) and one-stage algorithms (representative algorithms include the YOLO [14–16] series and SSD [17]). Among these, one-stage algorithms with fast detection speed, high computational efficiency, and good edge deployment capabilities have become common algorithms for small target detection in UAV aerial images. The YOLO series algorithms are typical representatives of this category, widely applied in UAV aerial image small target detection due to their fast, accurate, and efficient characteristics, with extensive research conducted by scholars worldwide.

Recent advances in YOLOv11-based detection systems have demonstrated significant improvements in various UAV applications [18–20]. However, existing UAV aerial image detection algorithms still suffer from the following problems to varying degrees: First, although the above-mentioned improved algorithms can improve small target detection performance to a certain extent, they lead to dramatic increases in computational cost, affecting overall model performance improvement while being difficult to apply to edge devices [21,22]. Additionally, insufficient utilization of feature information remains a problem. When facing scenarios where highly similar small targets appear in large numbers within limited areas, the above-mentioned improved methods perform poorly, resulting in missed detections [23,24].

To address these fundamental challenges in UAV aerial image small target detection, we propose FDE-YOLO, an enhanced YOLOv11-based detection framework incorporating three complementary architectural innovations for robust small target localization under challenging aerial imaging conditions. The proposed framework comprises a reconstructed neck architecture through the Fine-Grained Detection Pyramid (FGDP), which employs space-to-depth convolution and cross-stage partial multi-scale feature extraction to preserve small target information while avoiding the computational burden of excessive detection layer stacking. To enhance discriminative capability for densely clustered similar targets, we introduce a Dynamic Detection Fusion Head (DDFHead) that integrates scale-aware, spatial-aware, and task-aware attention mechanisms through sequential refinement, enabling adaptive focus on multi-scale targets across complex spatial distributions. Furthermore, the framework incorporates EdgeSpaceNet, a lightweight dual-branch feature extraction module that explicitly captures edge and spatial information through Sobel operators and standard convolutions with residual fusion, strengthening boundary discrimination for blurred small targets while reducing model complexity. These synergistic components collectively enhance feature representation quality, localization accuracy, and computational efficiency for UAV-based small target detection tasks.

In summary, the main contributions of this paper are as follows:

- To address the problem of excessive computational cost introduced by stacking detection layers, a Fine-Grained Detection Pyramid (FGDP) is designed with space-to-depth convolution and CSP-MFE modules to enhance multi-scale feature fusion capability. The pyramid structure adopts space-to-depth transformation strategy to preserve small target information through lossless downsampling while avoiding excessive computational overhead, achieving improved detection accuracy with controlled model complexity.
- To address the problem of missed detections caused by small target aggregation, a Dynamic Detection Fusion Head (DDFHead) incorporating DCNv4 and FReLU activation is developed to strengthen multi-dimensional perception through unified spatial-aware, scale-aware, and task-aware attention mechanisms. This detection head dynamically combines attention mechanisms to enhance discriminative capability for densely distributed similar small targets and improve localization precision in complex backgrounds.
- To address the problem of missed detections caused by blurred small target boundaries, an EdgeSpaceNet module combining Sobel edge extraction and spatial convolution is introduced to explicitly capture boundary and spatial features for small targets. The dual-branch architecture with residual fusion combines edge information and spatial information extraction, enhancing feature representation while reducing parameter count to facilitate deployment on edge devices.

The remainder of this paper is organized as follows. Section 2 reviews related work on YOLO-based small target detection, UAV detection systems, and attention mechanisms. Section 3 presents the detailed methodology of FDE-YOLO including network architecture and module designs. Section 4 reports experimental results including ablation studies, comparative experiments, and visualization analyses. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. YOLO-Based Small Target Detection

The YOLO series has evolved significantly since its introduction [14], with subsequent versions progressively improving detection accuracy and speed. YOLOv3 [25] introduced multi-scale prediction through feature pyramid networks, while YOLOv4 [15] incorporated various training techniques and architectural innovations. Recent iterations including YOLOX [26], YOLOv6 [27], YOLOv7 [28], and YOLOv9 [16] have continued this trajectory of improvement, demonstrating the framework's adaptability to diverse detection scenarios. However, these general-purpose architectures often struggle with small target detection in aerial imagery due to limited feature resolution and insufficient multi-scale representation.

To address small target detection challenges, researchers have proposed various architectural modifications focusing on feature enhancement and multi-scale fusion. Zhang et al. [29] proposed feature enhancement modules and spatial context awareness modules to enrich local and global context features respectively, effectively enhancing small target perception by fully utilizing local and global context information. Qi et al. [29] designed efficient multi-scale attention modules in the neck network to achieve cross-channel information intersection and cross-spatial learning, improving the feature extraction capability of the pyramid network and enhancing local feature correlation. Xiong et al. [30] learned weight parameters to adjust the degree of feature fusion, fusing large target information from deep features and small target information from shallow features by weight, thereby enhancing the final fusion effect.

Detection head innovations have also proven crucial for small target localization. Qin et al. [31] added a detection head to low-level high-resolution feature maps, which contains more small target information, improving the model's small target detection capability. Xie et al. [32] proposed an adaptive local-global feature aggregation layer to fully utilize the complementary correlation between local and global features, adaptively combining their advantages through attention-based fusion modules. Despite these advances, balancing detection accuracy with computational efficiency remains challenging, particularly for resource-constrained UAV platforms requiring real-time processing capabilities.

2.2. UAV-Based Detection Systems

UAV-based object detection has attracted substantial research attention due to its critical applications in surveillance, disaster response, and infrastructure inspection. Traditional two-stage detectors including Faster R-CNN [10,33] and its variants [11] provide high accuracy but suffer from computational inefficiency unsuitable for real-time UAV applications. Mask R-CNN [12] and its adaptations [13,34,35] have been applied to aerial image segmentation but face similar latency constraints. Single-stage detectors, particularly YOLO variants, have thus become preferred choices for UAV deployment due to their favorable speed-accuracy trade-offs [17].

Recent YOLOv11-based systems have demonstrated significant improvements through specialized architectural designs for aerial imagery. Bi et al. [4] proposed DRM-YOLO with MetaDWBlock integrating multi-branch depthwise separable convolutions and Cross-scale Feature Fusion Module employing CARAFE upsampling, achieving 21.4% improvement in mAP@0.5 on VisDrone2019. Liu et al. [19] introduced PD-YOLOv11 with C3K2_Sc backbone and CARAFE neck for power transmission tower detection, demonstrating superior accuracy and robustness in infrastructure inspection tasks. Tian et al. [18] developed SCDF-YOLOv11 with parallel branch feature extraction and selection-compensation feature integration mechanisms, maintaining lightweight design while significantly improving detection accuracy on UAV datasets [36,37].

Application-specific optimizations have further enhanced UAV detection capabilities across diverse domains. Huo et al. [5] combined UAV-borne and handheld LiDAR with YOLOv11 for precise urban tree species identification and biomass estimation, achieving 87.3% classification accuracy. You et al. [6] developed a lightweight insulator defect detection network based on improved YOLOv7 with rotating frame detection for aerial images, while Ji et al. [7] proposed WDI-YOLO for steel bridge weld defect detection using UAV images, achieving 93.7% precision with only 7.1 GB FLOPs. These domain-specific adaptations demonstrate the versatility of YOLO architectures for UAV-based detection tasks [2,20,38,39].

2.3. Attention Mechanisms and Dynamic Heads

Attention mechanisms have emerged as powerful tools for enhancing feature representation in object detection networks by enabling adaptive focus on informative regions while suppressing irrelevant background information. The Dynamic Head framework [40] introduced unified attention mechanisms operating across scale, spatial, and task dimensions, providing a principled approach to enhance detection head expressiveness. This concept has been successfully adapted for various detection tasks, demonstrating consistent improvements in handling multi-scale objects and complex spatial distributions [41,42].

Recent UAV detection systems have integrated advanced attention mechanisms to address small target localization challenges. Liang et al. [37] proposed YOLOv11-RAH with recurrent attention-enhanced edge intelligence networks, achieving approximately 5% mAP improvement over YOLOv11 baseline through iterative feature refinement with

depthwise gated convolutions. Hou et al. [21] developed MFEL-YOLO with multi-stage feature enhancement and context-enhanced heads employing multi-scale convolutions to capture contextual information, improving detection and classification accuracy for small aerial targets. These attention-based enhancements demonstrate the effectiveness of adaptive feature weighting for UAV detection tasks [22,24].

Lightweight attention mechanisms have been specifically designed for edge deployment on UAV platforms with limited computational resources. Liang et al. [23] combined Slicing Aided Hyper Inference (SAHI) with Focal loss and Convolutional Block Attention Module (CBAM) to address dense small target detection, achieving 85.17% precision while maintaining efficient processing of high-resolution UAV imagery. Qi et al. [43] integrated depth-based ROI extraction with improved YOLOv11 and SeaFormer lightweight segmentation for crack detection in bridge inspection, demonstrating 93.2% identification accuracy with efficient 3D reconstruction capabilities. These developments highlight the critical balance between detection performance and computational efficiency for practical UAV deployment [3,21,44].

3. Materials and Methods

3.1. YOLOv11 Architecture

YOLOv11 is the latest object detection algorithm model that demonstrates excellent performance with significant improvements over previous YOLO series algorithms. The YOLOv11 object detection algorithm model is primarily divided into three parts: feature extraction network (backbone), feature fusion network (neck), and detection network (head), as illustrated in Figure 1.

The backbone utilizes convolutional neural networks to transform raw image data into multi-scale feature maps, introducing the novel C3k2 block to optimize feature extraction efficiency while reducing computational cost. Additionally, it retains the Fast Spatial Pyramid Pooling (SPPF) module from previous versions and adds the C2PSA block to enhance the model's spatial attention capability. The neck serves as an intermediate processing stage, using specialized layers to aggregate and enhance feature representations at different scales, comprehensively utilizing upsampling and multi-scale feature concatenation methods to strengthen aggregation capabilities for information at different resolutions while integrating C3k2 and C2PSA blocks. The head functions as the prediction mechanism, combining C3k2 and CBS through multi-scale paths to achieve efficient feature processing and fine-grained target prediction [4,18].

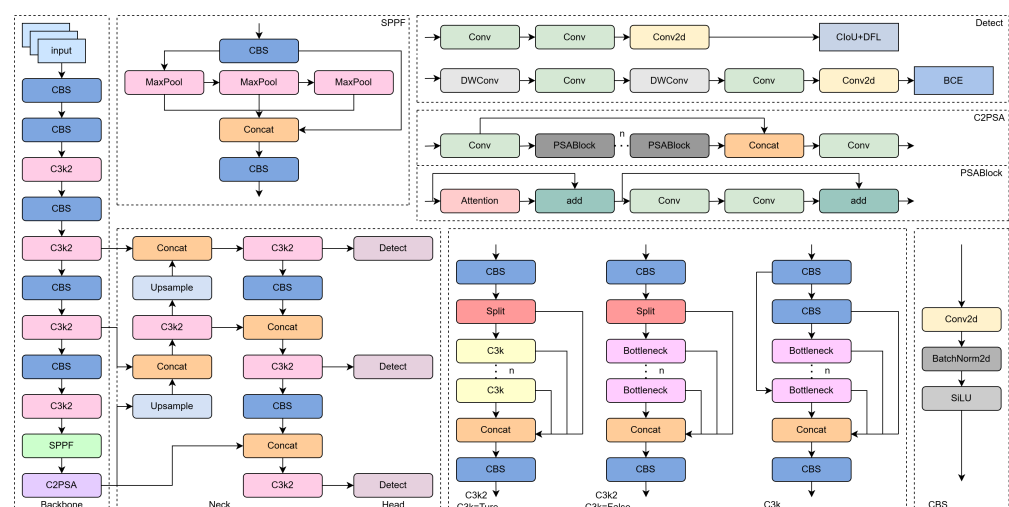


Figure 1. Overall architecture of the proposed FDE-YOLO network.

3.2. Fine-Grained Detection Pyramid (FGDP)

Existing YOLO series models still suffer from low sensitivity to small targets. Traditional optimization methods improve small target detection capability by stacking detection layers, but simultaneously bring problems such as excessive computational cost and time-consuming post-processing. Therefore, it is necessary to develop new feature pyramids effective for small targets. This paper proposes an improved Fine-Grained Detection Pyramid (FGDP), improving upon the original neck structure and combining innovative methods of feature extraction and multi-scale feature fusion to increase model attention to small targets while avoiding excessive computational cost. The architecture of FGDP is shown in Figure 2.

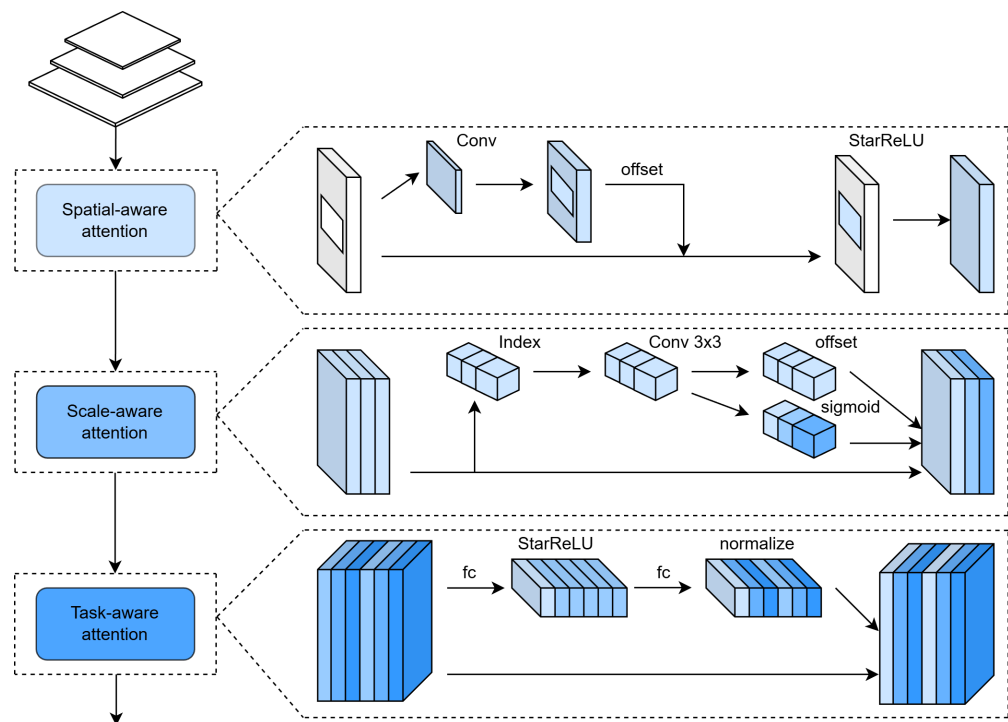


Figure 2. Architecture of the Fine-Grained Detection Pyramid (FGDP).

SPDConv [45] adopts a space-to-depth conversion strategy, avoiding information loss in traditional downsampling operations and obtaining features rich in small target information, as shown in Equation (1):

$$F_{p2}^{enh} = \sigma(BN(Conv(W, b, k)(SPD(F_{p2}, s)))) \quad (1)$$

where σ is the SiLU activation function, BN is the standard batch normalization layer, W is the weight matrix, b is the bias vector, k is the convolution kernel size, and s is the stride.

The CSP-MFE module is the core innovation of FGDP, combining the gradient shunting advantages of the CSP [46] structure with the multi-dimensional perception capability of our self-designed MFE module. The CSP structure divides the original input into two branches through channel splitting, enabling the model to learn richer gradient information. The MFE module enables the model to possess extreme capture capability for aerial small targets while maintaining lightweight characteristics through multi-granularity perspective adaptive feature recalibration, which adaptively reweights channel-wise features by learning importance coefficients to enhance discriminative representations.

After feature fusion, the input features first undergo channel adjustment through 1×1 convolution, then the input feature channels are allocated at a ratio of 25% and 75% through the Split operation. The shallow branch directly transmits 25% of features to

preserve detail information and enhance gradient flow, while the deep branch utilizes 75% of features for multi-scale feature extraction and modeling. This design effectively balances computational complexity and feature expression capability, as expressed in Equation (2):

$$[F_{light}, F_{dense}] = Split(Con_{1 \times 1}(F_{fused}), [0.25C, 0.75C]) \quad (2)$$

where F_{light} represents shallow branch features, F_{dense} represents deep branch features, F_{fused} represents fused features, $Split$ represents the channel splitting operation, and C is the total number of input feature channels.

The MFE module innovatively designs a three-branch parallel structure to solve the computational redundancy problem of traditional large kernel convolutions, reconstructing features from three dimensions: local details, anisotropic context, and global semantics, as shown in Figure 3.

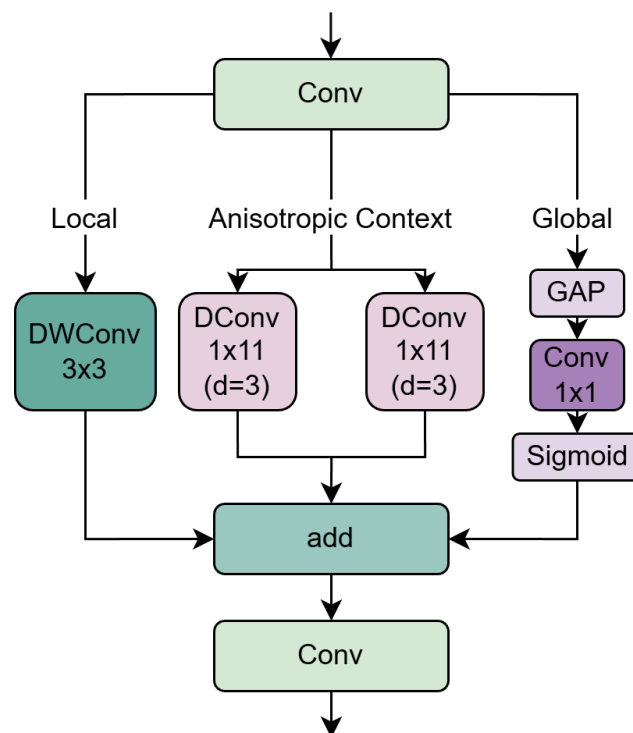


Figure 3. Structure of the multi-scale feature extraction (MFE) module.

In the global branch, to avoid excessive computational cost, complex self-attention mechanisms are abandoned in favor of a global gated aggregation strategy. First, global average pooling is performed on input features to obtain global distribution information, then channel weight coefficients are generated through 1×1 convolution, and finally coefficients ranging from 0 to 1 are generated through the Sigmoid activation function to gate information and recalibrate features, as formulated in Equation (3):

$$F_{global} = \sigma(Con_{1 \times 1}(GAP(F_{dense}))) \cdot F_{dense} \quad (3)$$

where F_{global} represents global branch output features, GAP denotes global average pooling operation, and σ is the Sigmoid activation function.

In the context branch, targeting elongated targets common in UAV aerial images, cross-shaped dilated convolution (CCDC) is designed. Two orthogonal depthwise separable strip convolutions ($1 \times K$ and $K \times 1$) are used with dilation rate d introduced, enabling capture

of contextual information spanning entire targets while drastically reducing parameter count, as shown in Equation (4):

$$F_{context} = DWConv_{1 \times 1}^{d=3}(F_{dense}) + DWConv_{11 \times 1}^{d=3}(F_{dense}) \quad (4)$$

where $F_{context}$ represents context branch output features, $DWConv$ represents depthwise separable convolution, and $d = 3$ represents the dilation rate. This design simulates human visual perception of long-range dependencies.

In the local branch, 3×3 depthwise separable convolution is used to extract texture and edge details of small targets, supplementing high-frequency information, as given in Equation (5):

$$F_{local} = DWConv_{3 \times 3}(F_{dense}) \quad (5)$$

where F_{local} represents local branch output features.

Finally, the MFE module adaptively fuses features from three dimensions, as expressed in Equation (6):

$$F_{out} = Concat(F_{light}, F_{global} + F_{context} + F_{local}) \quad (6)$$

The complete architecture of the CSP-MFE module is illustrated in Figure 4.

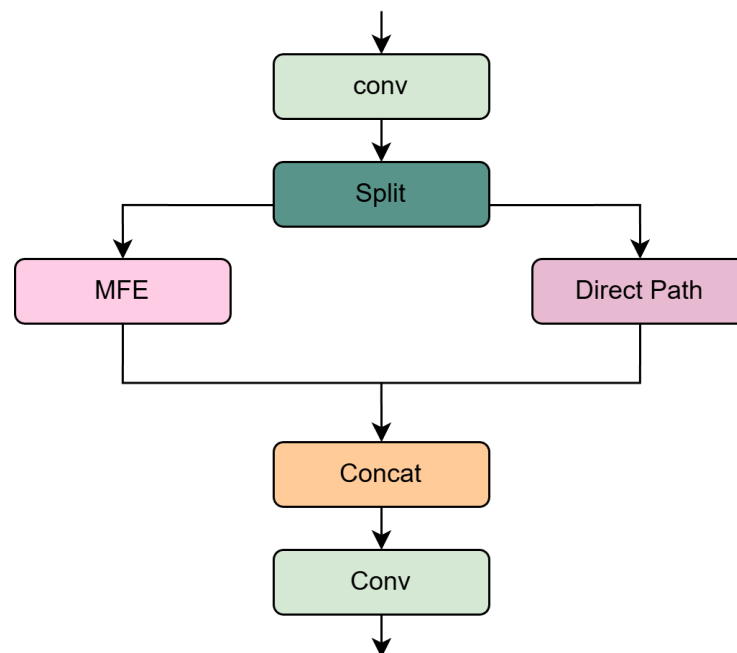


Figure 4. Architecture of the CSP-MFE module.

3.3. Dynamic Detection Fusion Head (DDFHead)

UAV aerial images often present scenarios where highly similar and spatially dense small targets appear in large numbers within limited areas. The original algorithm's insufficient focus on inter-class differences in small target feature maps leads to reduced target detection accuracy and missed detections. Traditional detection head optimization methods often only study one dimension of the network without fully utilizing information from scale, channel, and spatial dimensions. To address this problem, we designed a detection head called DDFHead that integrates scale, channel, and spatial attention mechanisms.

Dynamic Head [40] is an innovative dynamic detection head design for object detection tasks that improves scene adaptability and detection accuracy by introducing dynamic adjustment mechanisms. However, when small targets are densely clustered or targets are too similar, the original detection head struggles to accurately extract target infor-

mation. Therefore, we enhance information extraction capability in Dynamic Head by introducing DCNv4 as the convolution module to replace the original DCNv2 and introducing the FReLU adaptive activation function. DCNv4 removes softmax normalization in the spatial aggregation process, enhancing the network's dynamic adaptability and representation capability. Removing softmax overcomes its inherent limitations, including limited convergence speed and reduced operator expression capability. By adopting adaptive windows and dynamic unconstrained weights, DCNv4 provides greater flexibility in feature processing. Additionally, DCNv4 optimizes memory access patterns through analytical-level kernel analysis, reducing redundant operations and significantly improving computational efficiency. The optimized memory access substantially enhances overall operator performance.

The structure of DDFHead is shown in Figure 5.

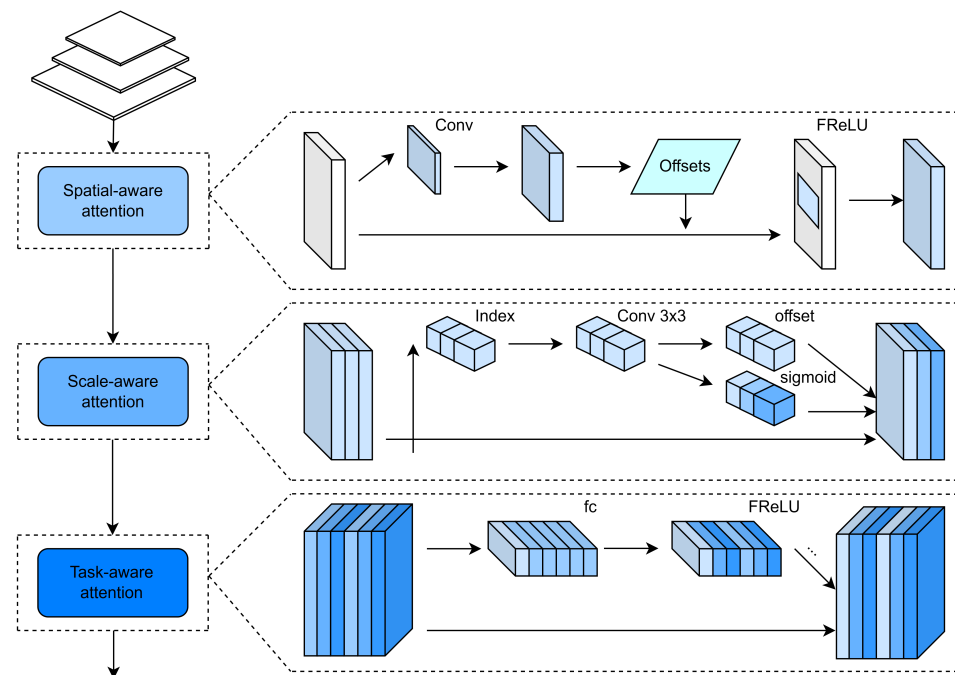


Figure 5. Architecture of the Dynamic Detection Fusion Head (DDFHead).

To improve image contrast and help the detection head more accurately locate and recognize targets, a spatial attention mechanism is constructed to focus on target regions. First, deformable convolution (DCNv4) learns offsets in feature maps, enabling features to adaptively focus on the shape and position of small targets. Then, features are aggregated across hierarchy at the same spatial position to enhance feature map discriminability. The spatial attention is formulated in Equation (7):

$$\pi_s(F) \cdot F = \sum_{l=1}^L \sum_{k=1}^K W_{l,k} \cdot F(l; p_k + \Delta p_k; c) \cdot \Delta m_k \quad (7)$$

where L represents the hierarchy of feature maps, K is the number of sparse sampling positions, and $p_k + \Delta p_k$ is the spatial offset learned by the network.

To address the problem of dense quantity and high similarity in small target detection, a scale attention mechanism is constructed to extract multi-scale features from different hierarchies, helping the detection head better capture targets. First, 1×1 convolutional layers are used to adaptively learn attention weights for each scale. Then, FReLU and

hard-sigmoid activation functions are used to assign weights to features at each scale. Finally, the normalized attention weight vector is applied to multi-scale feature maps to generate corresponding output results. The scale attention is expressed in Equation (8):

$$\pi_L(F) \cdot F = \sigma \left(f \left(\frac{1}{SC} \sum_{S,c} F \right) \right) \cdot F \quad (8)$$

where σ is the hard-sigmoid activation function, S is the number of scales, and C is the number of channels.

Under some complex environmental conditions, image contrast is significantly reduced. To ensure the model accurately focuses on local feature information, a channel attention mechanism is constructed to perform weighted selection on feature channels, emphasizing feature channels more important for small targets. By integrating information from these channels and filtering out irrelevant information, feature representativeness is improved. First, global average pooling is performed on input feature maps in the spatial dimension to obtain global feature representations for each channel. Then, global feature representations are processed through two fully connected layers, and a normalization layer is used to learn parameters controlling activation thresholds for each channel. Finally, FReLU adaptively adjusts the nonlinear activation degree. The task attention is formulated in Equation (9):

$$\pi_C(F) \cdot F = \max(\alpha_1(F) \cdot F_c, \beta_1(F)) \cdot F, (\alpha_2(F) \cdot F_c, \beta_2(F)) \quad (9)$$

where $[\alpha_1, \alpha_2, \beta_1, \beta_2]^T$ is the parameter vector controlling activation thresholds learned by the hyperfunction $\theta(\cdot)$.

This paper simultaneously introduces scale-aware attention, spatial position-aware attention, and task-aware attention in the detection head. Given a three-dimensional feature tensor $F \in R^{L \times S \times C}$ on the detection layer, this attention function is shown in Equation (10):

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \quad (10)$$

Through sequential stacking of $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ modules, the detection head can better integrate different information and flexibly adjust weights of different feature layers, enhancing integration capability for information from different dimensions.

3.4. EdgeSpaceNet Module

In UAV aerial image small target detection tasks, small targets often exhibit blurred boundaries. The original algorithm's insufficient accuracy in small target feature discrimination leads to reduced target detection accuracy and missed detections. Traditional optimization methods often cannot effectively extract edge and spatial features when increasing model feature extraction capability. To solve this problem, this research proposes a new efficient front-end feature extraction module called EdgeSpaceNet, integrating it into YOLOv11 to replace the traditional C3k2 convolution structure. The EdgeSpaceNet module combines edge information extraction and spatial information extraction, enabling the model to learn image features more comprehensively in complex scenes to improve small target feature discrimination accuracy, while reducing parameter count for convenient edge device deployment. The structure of EdgeSpaceNet is shown in Figure 6.

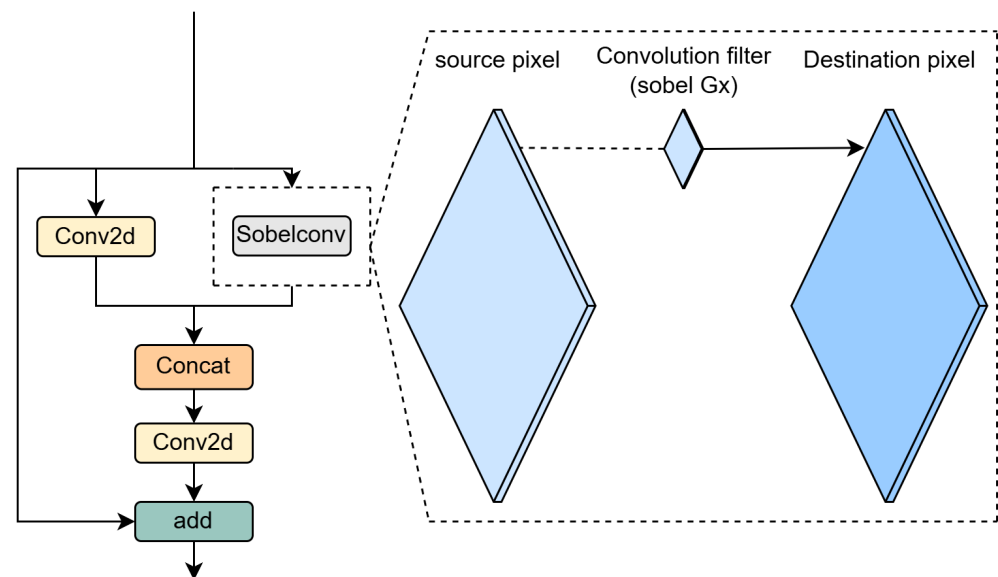


Figure 6. Architecture of the EdgeSpaceNet module.

EdgeSpaceNet adopts a dual-branch parallel structure design that enhances image feature representation capability by combining edge information and spatial information extraction capabilities, enabling the model to better discriminate features of different small targets. This module consists of two parts: a SobelConv branch for extracting image edge information and a convolution branch for extracting spatial information. The two branches' extracted edge and spatial information are then integrated through a feature fusion part.

In the SobelConv branch, the Sobel operator [47] contains convolution kernels in horizontal and vertical directions, effectively capturing intensity changes in images to obtain important edge information. Unlike learnable convolutions that may converge to arbitrary feature extractors during training, the fixed Sobel operator provides mathematically guaranteed gradient detection with stable boundary extraction properties, ensuring that the edge branch maintains its intended architectural complementarity with the spatial convolution branch. This non-learnable design reduces trainable parameters and enhances generalization across diverse aerial imaging conditions where edge patterns vary significantly due to atmospheric effects, illumination changes, and target appearance variations. Convolutional neural networks typically excel at learning spatial information but are somewhat insufficient in extracting edge information from images. The EdgeSpaceNet module can explicitly extract edge features of images through the SobelConv branch.

For input feature map X , gradients in horizontal and vertical directions are calculated respectively in Equations (11) and (12):

$$G_x = S_x * X \quad (11)$$

$$G_y = S_y * X \quad (12)$$

where S_x and S_y are Sobel operators in horizontal and vertical directions respectively, as defined in Equations (13) and (14):

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (13)$$

$$S_Y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (14)$$

Edge features can be expressed as shown in Equation (15):

$$F_{edge} = \sqrt{G_x^2 + G_y^2 + \epsilon} \quad (15)$$

where ϵ is a small positive number for numerical stability.

In addition to edge information, spatial information in images is equally important. The EdgeSpaceNet module extracts spatial information through an additional convolution branch. Unlike the SobelConv branch, the convolution branch extracts features from original images, preserving rich spatial details. The spatial feature extraction branch adopts a standard convolution structure, and its output features can be expressed in Equation (16):

$$F_{spatial} = \text{ReLU}(\text{BN}(W * X + b)) \quad (16)$$

where W and b are the weight and bias of the convolutional layer respectively, BN represents batch normalization operation, and ReLU represents the ReLU activation function.

The EdgeSpaceNet module finally achieves deep fusion and output of edge features and spatial features through two stages: feature concatenation and feature fusion. Edge features output by the SobelConv branch and spatial features output by the convolution branch undergo feature concatenation, as shown in Equation (17):

$$F_{cat} = [F_{edge}, F_{spatial}] \quad (17)$$

The concatenated features undergo convolution processing followed by residual connection for final output, as formulated in Equation (18):

$$F_{out} = W * [(W * F_{cat} + b) + F_{spatial}] + b \quad (18)$$

where F_{edge} and $F_{spatial}$ represent edge feature extraction branch and spatial feature extraction branch and mapping functions respectively, and F_{fusion} represents the feature fusion function.

4. Experiments

4.1. Datasets

The ablation and comparison experiments in this paper use the VisDrone2019-DET [48] dataset to verify model detection performance. This dataset was collected by the Machine Learning and Data Mining Laboratory team at Tianjin University and contains a total of 10,209 static images, including 6471 training images, 3190 test images, and 548 validation images. The dataset includes 10 categories: cars, pedestrians, people, trucks, vans, buses, motorcycles, bicycles, tricycles, and awning tricycles.

Generalization experiments in this paper are verified through experiments on the UAV aerial photography dataset UAV-DT [49] and the remote sensing image dataset NWPU VHR-10 [50]. The UAV-DT dataset was released by Zhejiang University and is specifically used for target detection and tracking tasks in UAV images, covering various scenarios such as urban streets and open spaces, including pedestrians and vehicles. The NWPU VHR-10 dataset was released by Northwestern Polytechnical University and is a commonly used benchmark for high-resolution remote sensing image target detection, containing 800 high-resolution satellite images covering 10 target categories including aircraft, ships, and storage

tanks, characterized by complex backgrounds, dense target distribution, and large intra-class variability.

4.2. Experimental Setup

4.2.1. Implementation Details

The experimental environment for this paper is Operating System: Windows 11, CPU: Intel(R) Core(TM) i9-10900X @ 3.70 GHz, running memory 64 GB, GPU: NVIDIA Tesla P40 with 48 GB memory. The deep learning framework uses PyTorch 2.2, the programming language uses Python 3.10, and the experimental environment is built in PyCharm 2025.2. Network training sets Epoch size to 200, and image input size to 640×640 and other key training hyperparameters are summarized in Table 1.

Table 1. Training hyperparameter settings.

Parameter Name	Parameter Value
Initial Learning Rate	0.01
Final Learning Rate	0.0001
Optimizer	SGD
Patience	100
Weight Decay	0.0005
Batch Size	32
Epochs	200
Input Size	640×640

4.2.2. Comparison Methods

To comprehensively evaluate the performance of the proposed FDE-YOLO, we compare it with multiple state-of-the-art methods including both two-stage detectors and one-stage detectors. For two-stage methods, we select Faster R-CNN [51], which represents the classical region-based detection paradigm. For one-stage methods, we select representative models from the YOLO series including YOLOv5s [52], YOLOX [26], YOLOv8s [53], YOLOv10s [54], YOLOv11s [55], YOLOv12s [56], and YOLO26 [57], as well as the transformer-based RT-DETR [58] and D-Fine [59]. Additionally, we compare with UAV-specialized detection methods including Drone-YOLO [60] and MASF-YOLO [61], which are specifically designed for real-time aerial image target detection. For detection head comparison, we evaluate our proposed DDFHead against the original Dynamic Head [40]. For edge detection operators, we compare Sobel [47], Prewitt, and Roberts operators within the EdgeSpaceNet module.

4.2.3. Evaluation Metrics

The main indicators adopted in this paper are precision (Pre), recall (R), mean average precision at IoU threshold mAP50 and mAP50:95, frames per second (FPS), and network model parameter count (P). Precision is the proportion of actual positive samples among samples predicted as positive, calculated as shown in Equation (19); recall is the proportion of samples correctly predicted as positive among all samples that are actually positive, calculated as shown in Equation (20); mAP is used to evaluate model detection accuracy across different categories, calculated as shown in Equations (21) and (22). mAP50 and mAP50:95 represent the mAP value when the IoU threshold is 0.50 and the mAP value calculated when the IoU threshold varies in the range of 0.50 to 0.95, respectively; Params

is the number of parameters in the model; FPS is the number of images processed by the model per second, used to measure detection speed.

$$P = \frac{TP}{TP + FP} \quad (19)$$

$$R = \frac{TP}{TP + FN} \quad (20)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP(i) \quad (21)$$

$$AP = \int P dR \quad (22)$$

where TP is the number of correctly predicted positive samples, FP is the number of negative samples predicted as positive, FN is the number of positive samples predicted as negative, n is the number of detection categories, and AP is the area contained by the precision–recall (PR) curve.

4.3. Results and Analysis

4.3.1. Performance Assessment

To comprehensively evaluate the effectiveness of FDE-YOLO, we conduct extensive comparisons with state-of-the-art object detection methods on the VisDrone2019 dataset under identical experimental conditions. The baseline methods include representative two-stage detectors (Faster R-CNN), classical one-stage detectors from the YOLO family (YOLOv5s, YOLOX, YOLOv8s, YOLOv10s, YOLOv11s, YOLOv12s, and YOLO26), transformer-based architectures (RT-DETR and D-Fine), and UAV-specific detection methods (Drone-YOLO and MASF-YOLO). All models are evaluated using consistent metrics including precision, recall, mAP50, mAP50:95, parameter count, and inference speed. The comprehensive comparison results are presented in Table 2.

Table 2. Performance comparison of algorithms on VisDrone2019 dataset.

Model	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)	P (M)	FPS
Faster R-CNN	40.4	28.9	29.9	14.2	41.5	12.1
YOLOv5s	46.9	35.2	34.9	19.4	7.0	86.4
YOLOX	44.3	37.3	37.8	22.1	9.1	74
YOLOv8s	47.5	35.2	38.8	22.4	11.4	118.2
YOLOv10s	49.7	37.7	38.0	21.6	7.2	130.3
YOLOv11s	50.8	38.1	39.2	23.5	9.5	122.5
YOLOv12s	49.5	37.5	38.4	23.3	9.1	106.3
YOLO26	50.5	37.5	39.0	22.9	9.4	126.2
RT-DETR	60.5	43.6	44.8	27.4	32.8	62.1
D-Fine	63.9	48.0	49.3	30.3	28.7	65.1
Drone-YOLO	52.6	40.3	40.5	21.7	16.4	70.6
MASF-YOLO	53.8	41.9	43.1	25.9	14.6	72.0
FDE-YOLO	53.6	42.5	43.3	26.3	10.25	69.8

The experimental results demonstrate that FDE-YOLO achieves superior detection performance while maintaining a favorable balance between accuracy and computational efficiency. Specifically, FDE-YOLO attains 53.6% precision, 42.5% recall, 43.3% mAP50, and 26.3% mAP50:95, representing substantial improvements of 2.8%, 4.4%, 4.1%, and 2.8% respectively over the baseline YOLOv11s. Compared with the UAV-specialized Drone-YOLO and MASF-YOLO, FDE-YOLO achieves 2.8% and 0.2% higher mAP50 respectively, while significantly reducing model parameters by 37.5% (10.25 M vs. 16.4 M) compared to

Drone-YOLO and 29.8% (10.25 M vs. 14.6 M) compared to MASF-YOLO, demonstrating the effectiveness of our lightweight architectural innovations. Against other YOLO variants including YOLOv5s, YOLOX, YOLOv8s, YOLOv10s, YOLOv12s, and the recently proposed YOLO26, FDE-YOLO consistently outperforms across all evaluation metrics, validating its robustness for small target detection in aerial imagery. When compared with transformer-based architectures, D-Fine exhibits the highest overall performance with 49.3% mAP50 and 30.3% mAP50:95, followed by RT-DETR with 44.8% mAP50 and 27.4% mAP50:95. While FDE-YOLO achieves slightly lower accuracy (6.0% mAP50 gap with D-Fine and 1.5% gap with RT-DETR), it requires only 35.7% of D-Fine's parameters (10.25 M vs. 28.7 M) and 31.2% of RT-DETR's parameters (10.25 M vs. 32.8 M), highlighting the efficiency advantages of our approach for resource-constrained UAV deployment scenarios where computational budget and real-time performance are critical constraints.

4.3.2. Ablation Study

To verify the performance of the improved modules proposed in this paper, ablation experiments were conducted on the VisDrone2019 dataset using the YOLOv11s model as baseline. The effects of the improved Fine-Grained Detection Pyramid (FGDP), self-developed detection head DDFHead, and EdgeSpaceNet replacing C3k2 were further analyzed. The experimental results are shown in Table 3.

Adding the algorithm YOLOv11s-F that improves the original Neck structure to enhance small target detection accuracy, results show that precision (Pre), recall (R), mAP50, and mAP50:95 increased by 2.3%, 3.2%, 3.4%, and 2.2% respectively, while network size only increased by 0.66 M. On this basis, introducing the dynamic detection head algorithm YOLOv11s-FD to enhance feature extraction and improve recognition capability for targets in different complex environments, precision (Pre) and recall (R) increased by 0.1% and 0.4%, and mean average precision mAP50 and mAP50:95 also increased by 0.5% and 0.4% respectively. On this basis, introducing the algorithm YOLOv11s-FDE that replaces the traditional C3k2 convolution structure, by combining edge information and spatial information extraction to enhance image feature representation capability, precision (Pre), recall (R), mAP50, and mAP50:95 increased by 0.4%, 0.8%, 0.2%, and 0.2% respectively, while simultaneously reducing network model parameter count (P) by 0.25 M. The final improved algorithm FDE-YOLO compared to the original algorithm YOLOv11s shows precision (Pre) increased by 2.8%, recall (R) increased by 4.4%, and mean average precision mAP50 and mAP50:95 increased by 4.1% and 2.8% respectively, with relatively obvious improvement in detection performance.

Table 3. Ablation experiment results on VisDrone2019 dataset.

Model	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)	P (M)	FPS
YOLOv11s	50.8	38.1	39.2	23.5	9.5	122.5
YOLOv11s-F	53.1	41.3	42.6	25.7	10.16	104.5
YOLOv11s-FD	53.2	41.7	43.1	26.1	10.48	77.6
FDE-YOLO	53.6	42.5	43.3	26.3	10.25	69.8

Comparative analysis of experimental result figures for YOLOv11s and the improved algorithm shows that the improved model achieves significant detection accuracy improvements across all categories, especially with Bus class improving by 9.5%. Additionally, although the Bicycle and Awning Tricycle classes have limited feature information making them difficult to detect, detection accuracy for these two classes also improved by 5.5% and 3.6%, respectively. The precision–recall curves for both models are presented separately in Figures 7 and 8 for better clarity and detailed analysis.

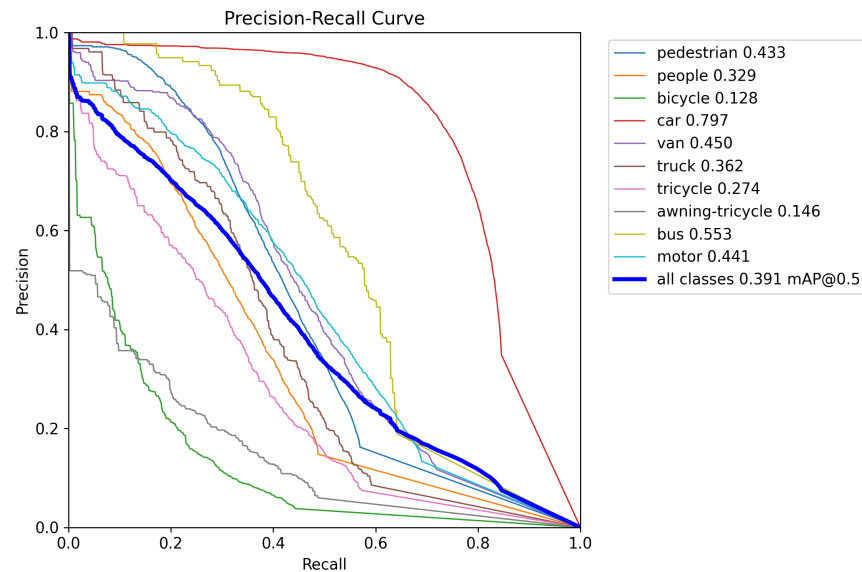


Figure 7. Precision–recall curves of YOLOv11s baseline model on VisDrone2019 dataset.

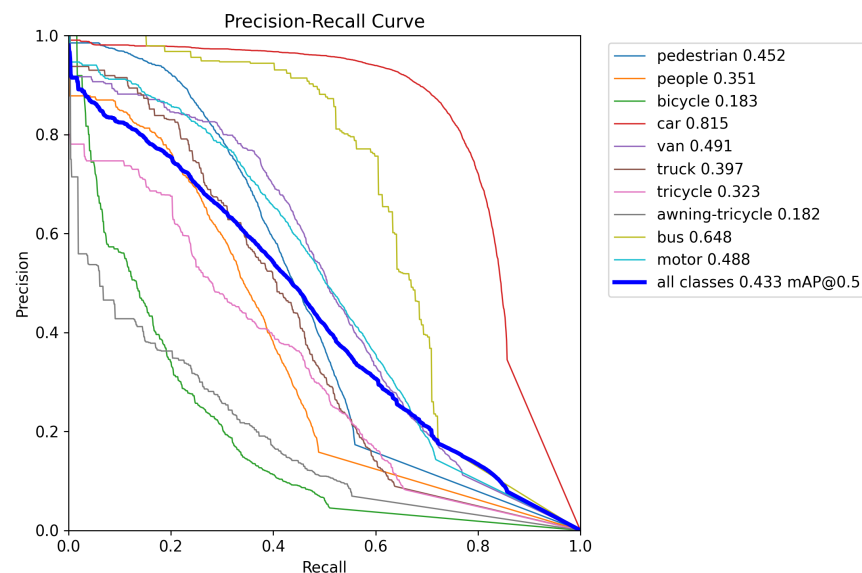


Figure 8. Precision–recall curves of FDE-YOLO model on VisDrone2019 dataset.

The ablation experiments demonstrate that feature fusion mechanisms fundamentally contribute to improved object detection performance in aerial images. Among the proposed modules, the Fine-Grained Detection Pyramid (FGDP) exhibits the most substantial impact on detection accuracy. As shown in Table 3, introducing FGDP alone (YOLOv11s-F) improves mAP50 by 3.4% and mAP50:95 by 2.2%, accounting for 82.9% and 78.6% of the total improvement, respectively. This substantial gain stems from FGDP’s multi-scale feature fusion strategy, which integrates features across multiple granularity levels—local details through 3×3 depthwise convolutions, anisotropic context via cross-shaped dilated convolutions, and global semantics through gated aggregation. Given that small targets in UAV aerial images often span only 5–20 pixels, they are highly sensitive to feature scale. The multi-scale fusion architecture ensures that subtle appearance cues at fine scales are preserved while semantic context from coarse scales provides discrimination against background clutter, particularly critical in complex aerial scenes with dense small objects.

Beyond multi-scale integration, the hierarchical feature fusion architecture in CSP-MFE demonstrates how gradient flow optimization enhances representation learning.

By splitting features into 25% shallow and 75% deep branches, the module enables parallel processing of features at different abstraction levels before fusion. This design alleviates gradient vanishing issues while allowing the network to jointly optimize multi-granularity features during backpropagation. The adaptive feature recalibration mechanism in the global branch further refines the fused features by learning channel-wise importance weights, suppressing irrelevant channels while amplifying discriminative ones. In aerial images where background patterns such as roads and buildings often exhibit similar textures to small vehicles or pedestrians, this channel-wise recalibration following fusion significantly improves feature discriminability, contributing to the 3.2% recall improvement observed in the ablation study.

The fusion of edge and spatial features through EdgeSpaceNet provides another dimension of complementary information essential for precise localization. The ablation results indicate that adding EdgeSpaceNet (YOLOv11s-FDE) contributes an additional 0.8% recall improvement beyond FGDP and DDFHead. This gain is attributed to the explicit extraction and fusion of boundary information via Sobel operators with spatial contextual features. Small targets in aerial images frequently suffer from blurred boundaries due to motion blur, atmospheric effects, or low resolution. By fusing edge-enhanced features with spatial convolution outputs through residual connections, EdgeSpaceNet recovers boundary details typically lost in standard convolutions, enabling more accurate bounding box regression. The synergistic integration of multi-level feature fusion (FGDP), hierarchical channel fusion (CSP-MFE), and boundary-spatial fusion (EdgeSpaceNet) ultimately achieves 4.1% mAP50 improvement, demonstrating that comprehensive feature fusion across scales, channels, and feature types is fundamental to robust small target detection in challenging aerial scenarios.

To verify the precision and efficiency of the detection head DDFHead designed in this paper, comparison experiments with the Dynamic Head detection head were designed. The experimental data uses the VisDrone2019 dataset, with fixed test image size of 640×640 , baseline model YOLOv11s, and sequential replacement with our designed detection head and Dynamic Head detection head. Experimental results are shown in Table 4. From the table, it can be seen that the detection head designed in this paper achieves improvements of 0.8%, 0.5%, 1.3%, and 1.3% in precision (Pre), recall (R), mAP50, and mAP50:95, respectively, compared to the Dynamic Head detection head. Experimental results demonstrate that the DDFHead detection head achieves overall detection effect improvements based on the Dynamic Head detection head.

Table 4. Performance comparison of detectors with different detection heads.

Model	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)
YOLOv11s	50.8	38.1	39.2	23.5
+Dynamic Head	52.0	40.5	40.6	24.4
+DDFHead	52.8	41.0	41.9	25.7

To verify the superiority of the Sobel operator in the EdgeSpaceNet module designed in this paper, comparison experiments were conducted by integrating EdgeSpaceNet with different edge detection operators (Sobel, Prewitt, and Roberts) to replace the traditional C3k2 convolution module in YOLOv11s. The experimental data uses the VisDrone2019 dataset, with fixed test image size of 640×640 . Experimental results are shown in Table 5. From the table, it can be seen that the Sobel-based EdgeSpaceNet achieves improvements of 2.6%, 2.8%, 2.4%, and 2.4% in precision (Pre), recall (R), mAP50, and mAP50:95, respectively, compared to the Prewitt-based variant, and improvements of 5.2%, 5.4%, 5.1%, and 4.2%,

respectively, compared to the Roberts-based variant. Experimental results demonstrate the superiority of the Sobel operator for edge feature extraction in the EdgeSpaceNet module.

Table 5. Performance comparison with different operators.

Operator Type	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)
Sobel	51.5	39.5	40.5	24.3
Prewitt	48.9	36.7	38.1	21.9
Roberts	46.3	34.1	35.4	20.1

4.3.3. Parameter Sensitivity Analysis

To validate the robustness of FDE-YOLO and justify the parameter choices adopted in our experiments, we conduct comprehensive sensitivity analysis on four critical parameters: CSP branch ratio in the CSP-MFE module, dilation rate in the cross-shaped dilated convolution, initial learning rate, and batch size. These parameters directly influence the model's feature representation capability, gradient flow optimization, and convergence behavior. For each parameter, we systematically vary its value while keeping all other configurations constant, evaluating the impact on detection performance using mAP50 and mAP50:95 metrics on the VisDrone2019 validation set. The comprehensive sensitivity analysis results are presented in Figure 9, which reveals the relationship between parameter configurations and detection accuracy across different architectural and training dimensions.

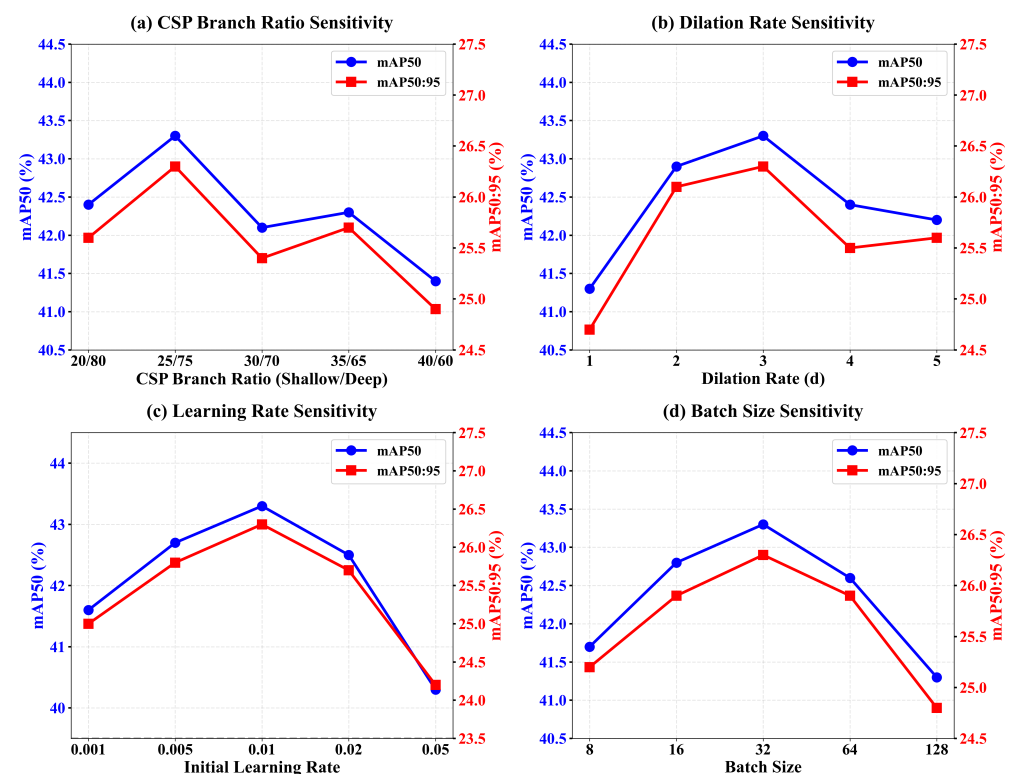


Figure 9. Parameter sensitivity analysis for FDE-YOLO: (a) CSP branch ratio, (b) dilation rate, (c) initial learning rate, and (d) batch size.

The sensitivity curves presented in Figure 9 demonstrate that our adopted parameter configurations achieve optimal or near-optimal performance across all evaluated settings. For the CSP branch ratio, the 25/75 allocation achieves peak performance, substantially outperforming alternative configurations, with the 20/80 ratio showing marginally lower performance due to insufficient shallow feature representation, while higher ratios (30/70, 35/65, and 40/60) exhibit progressive degradation as reduced deep branch capacity limits

multi-scale extraction. For the dilation rate in CCDC, $d = 3$ achieves optimal performance with $d = 2$ yielding highly competitive results, while smaller rates provide insufficient contextual coverage for elongated targets and larger rates introduce excessive spatial gaps degrading localization accuracy. The learning rate analysis reveals that 0.01 optimally balances convergence speed and final accuracy, with 0.005 producing comparable but slightly lower results requiring longer training, whereas higher rates cause training instability and notable performance degradation. Similarly, batch size 32 achieves peak performance, with size 16 maintaining competitive accuracy at the cost of noisier gradient estimates, while larger batches exhibit diminishing returns. The relatively smooth sensitivity curves around optimal configurations indicate stable performance under moderate parameter variations, suggesting practical robustness without requiring exhaustive hyperparameter tuning. Notably, all evaluated configurations consistently surpass the baseline YOLOv11s performance across both metrics, validating that the architectural innovations in FGDP, DDFHead, and EdgeSpaceNet provide fundamental improvements beyond mere hyperparameter optimization, while also providing practitioners with guidance for adapting FDE-YOLO to specific application requirements or hardware constraints.

4.3.4. Visualization Analysis

Images of dense targets and complex backgrounds in different scenarios were selected from the experimental dataset to test the effectiveness of the FDE-YOLO algorithm, as shown in Figure 10. From the comparison in the figure, it can be seen that in scenarios with dense targets, a large number of similar small targets appear clustered together. Compared to the baseline algorithm YOLOv11s, the FDE-YOLO algorithm reduces missed detections and false positives of small targets, with overall better detection performance. In scenarios with complex backgrounds, fast-moving vehicles highly overlap with roads, and basketball courts have high similarity with surrounding environments, causing target boundaries to blur. The baseline algorithm YOLOv11s struggles to detect and recognize targets, while the FDE-YOLO algorithm can still accurately identify targets.

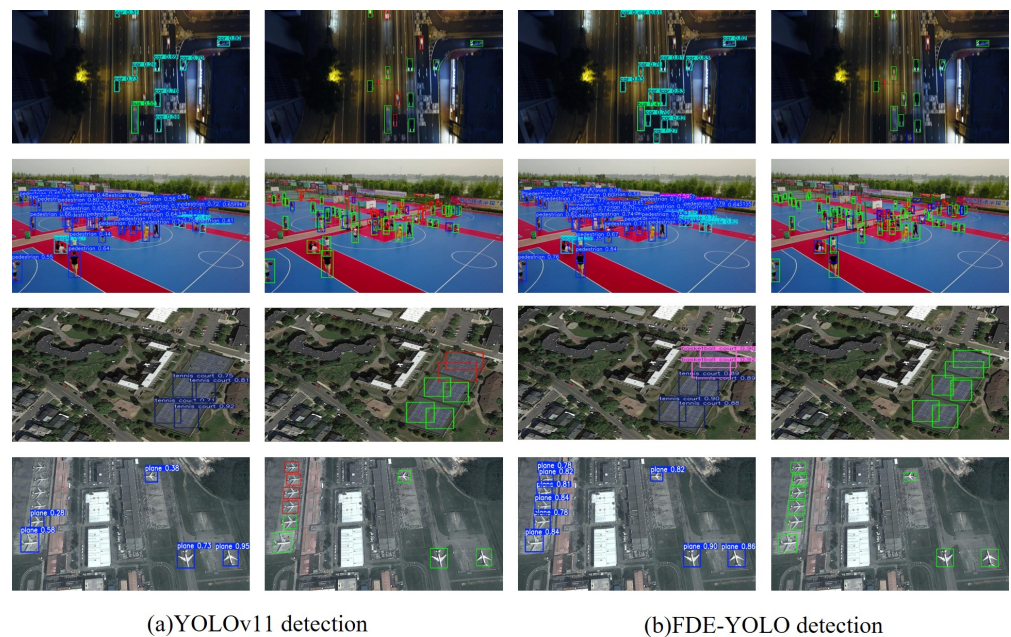


Figure 10. Visual comparison of detection results between YOLOv11 and FDE-YOLO.

To verify and analyze the reasons for FDE-YOLO algorithm's improved detection effect on small targets, this paper uses heat maps for analysis and comparison, as shown in

Figure 11. From the figure, it can be seen that the baseline model YOLOv11s ignores some clustered small targets and difficult-to-identify targets, and focuses on useless positions. The improved algorithm model focuses more carefully, with more accurate coverage positions, reducing attention to uninteresting targets while also being able to focus on distant smaller targets.

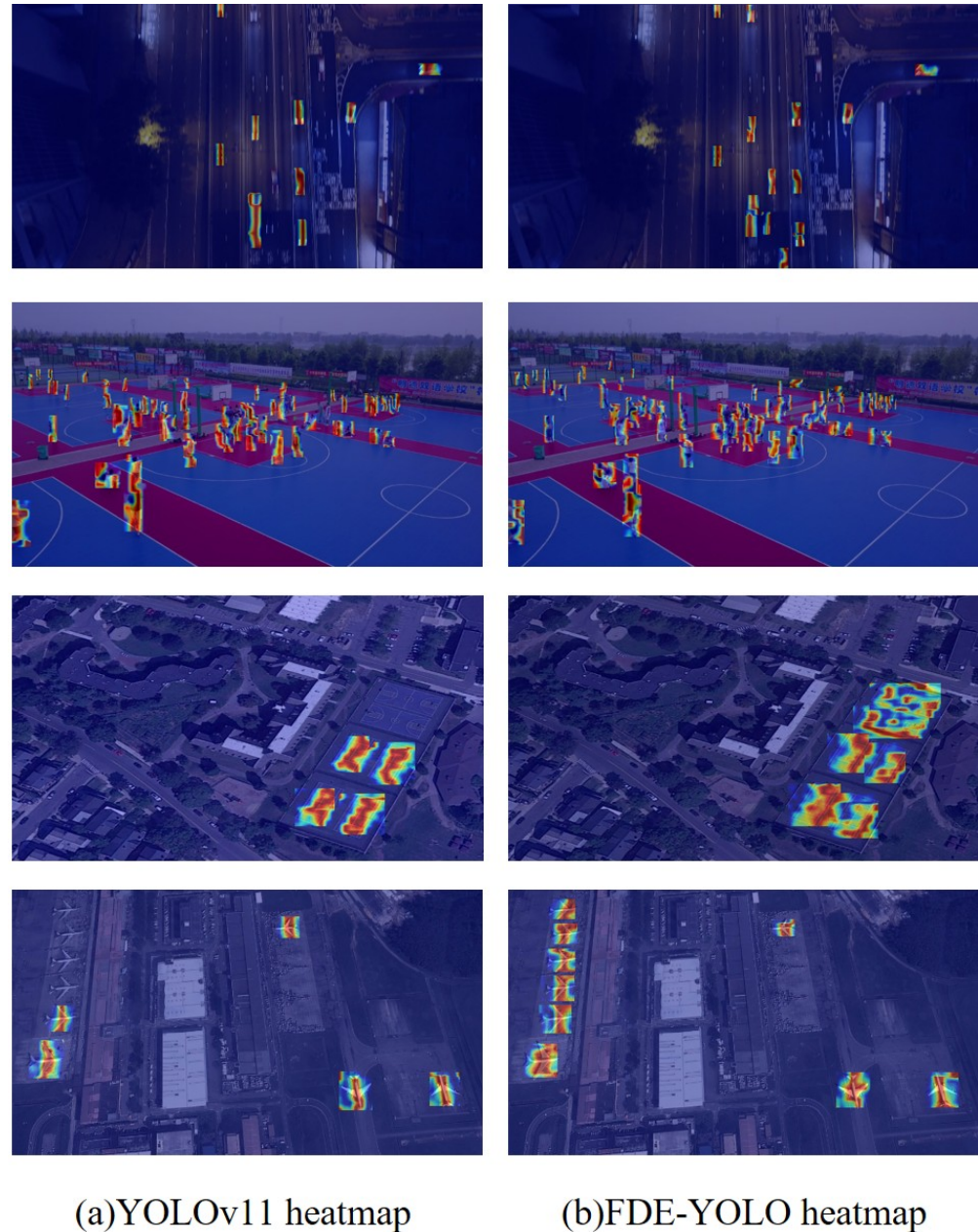


Figure 11. Attention heatmap visualization comparison between YOLOv11 and FDE-YOLO.

4.3.5. Generalization Analysis

To rigorously assess the generalization capability and cross-domain robustness of FDE-YOLO, we conduct comprehensive validation experiments on two additional benchmark datasets with distinct characteristics from the primary VisDrone2019 training corpus. The UAV-DT dataset, released by Zhejiang University, comprises UAV imagery captured under diverse urban and open-space scenarios with varying illumination conditions, camera angles, and target densities, providing a stringent test for model adaptability to heterogeneous aerial imaging conditions. The NWPU VHR-10 dataset, established by Northwestern

Polytechnical University, contains high-resolution satellite remote sensing images covering ten object categories including aircraft, ships, and storage tanks, characterized by significantly different imaging modalities, target scales, and background complexity compared to low-altitude UAV footage. These cross-dataset evaluations enable assessment of whether the architectural innovations in FDE-YOLO learn generalizable feature representations rather than dataset-specific patterns, a critical consideration for real-world deployment where imaging conditions inevitably deviate from training distributions.

The generalization performance on the UAV-DT dataset, presented in Table 6, demonstrates FDE-YOLO's robust transferability to unseen aerial imaging scenarios. FDE-YOLO achieves 93.1% precision, 90.8% recall, 96.0% mAP50, and 67.1% mAP50:95, substantially outperforming the baseline YOLOv11s by 0.5%, 3.0%, 1.6%, and 4.9%, respectively, across all metrics. Notably, the 3.0% recall improvement and 4.9% mAP50:95 enhancement are particularly significant, indicating that FDE-YOLO's multi-scale feature fusion and edge-spatial feature extraction mechanisms effectively capture diverse target appearances and localization cues that generalize beyond the training domain. When compared with the UAV-specialized Drone-YOLO method, FDE-YOLO achieves 1.9% higher mAP50 (96.0% vs. 94.1%), validating that our architectural designs yield superior cross-scenario adaptability. The consistent performance gains across precision, recall, and mAP metrics suggest that FDE-YOLO successfully balances false positive suppression and missed detection reduction in varying environmental conditions, demonstrating robust generalization to diverse UAV imaging scenarios beyond the VisDrone2019 training distribution.

Table 6. Generalization experiments based on UAV-DT dataset.

Model	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)
YOLOv5s	90.2	84.9	89.3	55.8
YOLOv8s	91.7	86.5	92.6	60.5
YOLOv10s	91.5	86.1	92.4	60.1
YOLOv11s	92.6	87.8	94.4	62.2
YOLOv12s	92.1	86.3	93.0	61.0
Drone-YOLO	92.4	87.3	94.1	61.9
FDE-YOLO	93.1	90.8	96.0	67.1

The cross-domain generalization to satellite remote sensing imagery, evaluated on the NWPU VHR-10 dataset (Table 7), further substantiates FDE-YOLO's capability to transfer learned representations across drastically different imaging modalities. Despite the substantial domain shift between low-altitude UAV footage (training domain) and high-altitude satellite imagery (test domain), which encompasses differences in spatial resolution, viewing geometry, atmospheric effects, and target appearance, FDE-YOLO maintains strong detection performance with 95.6% precision, 82.6% recall, 94.6% mAP50, and 63.9% mAP50:95. Compared to the baseline YOLOv11s, FDE-YOLO achieves improvements of 0.7%, 0.8%, 1.5%, and 3.0% across all evaluation metrics, with the 3.0% mAP50:95 enhancement being particularly noteworthy as it reflects improved localization accuracy under stricter IoU thresholds. The 2.2% mAP50 advantage over Drone-YOLO (94.6% vs. 92.4%) underscores the superior generalizability of FDE-YOLO's feature extraction and fusion strategies. Collectively, the consistent performance improvements across both UAV-DT and NWPU VHR-10 datasets, spanning diverse imaging conditions, target categories, and domain characteristics, provide compelling evidence that FDE-YOLO's architectural innovations—particularly the multi-granularity feature fusion in FGDP, adaptive attention mechanisms in DDFHead, and edge-spatial feature integration in EdgeSpaceNet—learn robust, generalizable representations rather than overfitting to specific dataset charac-

teristics, thereby ensuring reliable deployment across heterogeneous real-world aerial detection scenarios.

Table 7. Generalization experiments based on NWPU VHR-10 dataset.

Model	Pre (%)	R (%)	mAP50 (%)	mAP50:95 (%)
YOLOv5s	94.1	80.8	90.8	59.4
YOLOv8s	94.3	80.7	91.6	59.8
YOLOv10s	94.5	81.1	92.2	60.5
YOLOv11s	94.9	81.8	93.1	60.9
YOLOv12s	93.9	79.3	90.0	60.2
Drone-YOLO	94.9	81.5	92.4	60.5
FDE-YOLO	95.6	82.6	94.6	63.9

5. Conclusions

This paper addresses the issues of missed detections, false positives, and low accuracy in UAV aerial image small target detection algorithms by proposing the FDE-YOLO algorithm based on improved YOLOv11. By introducing the improved Fine-Grained Detection Pyramid (FGDP), the algorithm extracts composite features of targets to provide rich target feature information for the neck structure, improving small target detection performance while avoiding excessive computational cost. By designing the dynamic detection head DDFHead, the algorithm enhances perception of different scales, spatial positions, and tasks, further improving the accuracy and robustness of small target detection. Finally, by designing the efficient front-end feature module EdgeSpaceNet to replace the traditional convolution structure C3k2, the algorithm improves learning capability for edge and spatial information, enhancing model detection effect while reducing model complexity.

Experimental results on the VisDrone2019 dataset show that the proposed FDE-YOLO algorithm achieves precision of 53.6%, and mAP50 and mAP50:95 of 43.3% and 26.3% respectively, surpassing the original YOLOv11 algorithm and most mainstream YOLO-based object detection algorithms including YOLOv5s, YOLOX, YOLOv8s, YOLOv10s, YOLOv12s, and YOLO26, while maintaining competitive performance compared to transformer-based methods with significantly fewer parameters, with good performance. The algorithm effectively completes UAV aerial image small target detection tasks with superior detection performance for small targets compared to other mainstream models. Meanwhile, generalization experiments on the UAV-DT and NWPU VHR-10 datasets demonstrate that FDE-YOLO detection accuracy mAP50 improved by 1.6% and 1.5%, respectively, compared to the baseline algorithm.

Although FDE-YOLO maintains a relatively compact model size, practical deployment on resource-constrained edge devices such as NVIDIA Jetson series or other embedded platforms commonly mounted on UAV systems requires further validation and optimization. Future research will focus on comprehensive edge device deployment verification, including inference speed benchmarking, memory consumption profiling, and power efficiency analysis on actual hardware platforms. Additionally, adopting model compression techniques such as pruning, quantization, or knowledge distillation could further reduce computational requirements while preserving detection accuracy. Exploring multi-modal fusion approaches combining RGB and thermal imaging could enhance detection robustness in challenging environmental conditions. Integration of attention mechanisms at multiple scales and investigation of transformer-based architectures may provide additional performance improvements while maintaining computational efficiency for practical edge deployment scenarios.

Author Contributions: Conceptualization, J.L. and J.J.; methodology, J.L. and X.G.; software, J.L. and X.Z.; validation, J.L., X.G. and J.J.; formal analysis, J.L. and X.G.; investigation, J.L., X.G. and X.Z.; resources, J.J.; data curation, J.L. and X.G.; writing—original draft preparation, J.L. and X.G.; writing—review and editing, J.L., X.G., X.Z. and J.J.; visualization, J.L. and X.G.; supervision, J.J.; project administration, J.J.; funding acquisition, J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The VisDrone2019 dataset utilized in this study is publicly available at <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 18 January 2026). The UAV-DT dataset is available at <https://sites.google.com/view/grli-uavdt> (accessed on 18 January 2026). The NWPU VHR-10 dataset is publicly available and can be downloaded from <https://gcheng-nwpu.github.io/> (accessed on 18 January 2026). The code and trained models are available from the corresponding author upon reasonable request.

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

References

1. Feng, X.; Liu, L.; Ye, M.; Mašek, O.; Gouda, S.; Chang, K.; Wang, X.; Huang, Q. Unveiling and interpreting the relationships among multi-pollutant emission factors in municipal solid waste incineration by machine learning. *Waste Manag.* **2026**, *210*, 115256. [CrossRef] [PubMed]
2. Al-Obeidat, N.; Li, Z. A Multimodal UAV-Based Pipeline for Precision Agriculture: Aerial Stress Detection with YOLO and High-Fidelity Disease Classification Using DeiT. *Procedia Comput. Sci.* **2025**, *257*, 921–926. [CrossRef]
3. Mujtaba, G.; Liu, W.; Alshehri, M.; AlQahtani, Y.; Almujaally, N.A.; Liu, H. Aerial Images for Intelligent Vehicle Detection and Classification via YOLOv11 and Deep Learner. *Comput. Mater. Contin.* **2025**, *86*, 1–19. [CrossRef]
4. Bi, H.; Dai, R.; Han, F.; Zhang, C. DRM-YOLO: A YOLOv11-based structural optimization method for small object detection in UAV aerial imagery. *Image Vis. Comput.* **2026**, *167*, 105894. [CrossRef]
5. Huo, Z.; Fang, L.; Chu, Y.; Dang, S.; Yang, J.; Li, L.; Li, X.; Ren, S.; Chen, J.; Peng, Y.; et al. Precise urban tree species identification and biomass estimation using UAV–Handheld LiDAR Synergy and YOLOv11 deep learning. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105049. [CrossRef]
6. You, X.; Zhao, X. A insulator defect detection network based on improved YOLOv7 for UAV aerial images. *Measurement* **2025**, *253*, 117410. [CrossRef]
7. Ji, W.; Liu, S.; Deng, L.; Li, J.; Liu, Y.; Xiong, Z. WDI-YOLO: A lightweight steel bridge weld defect detection algorithm using UAV images. *J. Constr. Steel Res.* **2025**, *235*, 109833. [CrossRef]
8. Oyallon, E.; Rabin, J. An analysis of the SURF method. *Image Process. Line* **2015**, *5*, 176–218. [CrossRef]
9. Konstantinidis, D.; Stathaki, T.; Argyriou, V.; Grammalidis, N. Building detection using enhanced HOG–LBP features and region refinement processes. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *10*, 888–905. [CrossRef]
10. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS’16, Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016*; Curran Associates Inc.: Red Hook, NY, USA, 2016.
11. Wan, J.; Zhang, B.; Zhao, Y.; Du, Y.; Tong, Z. Vistrongerdet: Stronger visual information for object detection in visdrone images. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 2820–2829.
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2961–2969.
13. Bharati, P.; Pramanik, A. Deep learning techniques—R-CNN to mask R-CNN: A survey. In *Computational Intelligence in Pattern Recognition, Proceedings of CIPR 2019*; Springer: Singapore, 2019; pp. 657–668.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
16. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. Yolov9: Learning what you want to learn using programmable gradient information. In *Computer Vision—ECCV 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 1–21.

17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
18. Tian, M.; Cui, M.; Yu, S.; Chen, Z.; Song, Y. SCDF-YOLOv11: Selection and Compensation Detailed Feature Integration Network for UAV Detection. *Chin. J. Aeronaut.* **2025**, *in press*. [[CrossRef](#)]
19. Liu, L.; Meng, L.; Li, A.; Lv, Y.; Zhao, B. PD-YOLOv11: A power distribution enabled YOLOv11 algorithm for power transmission tower component detection in UAV inspection. *Alex. Eng. J.* **2025**, *131*, 312–324. [[CrossRef](#)]
20. Cai, H.; Cai, M.; Hua, J. Design of Unmanned Aerial Vehicle Image Intelligent Recognition System Based on Machine Learning Algorithm. *Int. J. Intell. Inf. Technol.* **2025**, *21*, 1–20. [[CrossRef](#)]
21. Hou, T.; Leng, C.; Wang, J.; Pei, Z.; Peng, J.; Cheng, I.; Basu, A. MFEL-YOLO for small object detection in UAV aerial images. *Expert Syst. Appl.* **2025**, *291*, 128459. [[CrossRef](#)]
22. Chen, X.; Lin, C. EVMNet: Eagle visual mechanism-inspired lightweight network for small object detection in UAV aerial images. *Digit. Signal Process.* **2025**, *158*, 104957. [[CrossRef](#)]
23. Liang, X.; Xiang, J.; Qin, S.; Xiao, Y.; Chen, L.; Zou, D.; Ma, H.; Huang, D.; Huang, Y.; Wei, W. Small target detection algorithm based on SAHI-Improved-YOLOv8 for UAV imagery: A case study of tree pit detection. *Smart Agric. Technol.* **2025**, *12*, 101181. [[CrossRef](#)]
24. Xie, J.; Han, D.; Cheng, T.; Niu, Z.; Li, W.; Su, Y.; Yu, L.; Yuan, F.; Wang, D.; Zhang, D. Accurate UAV-based detection of planting pits via spectral-spatial dual-domain collaboration. *Smart Agric. Technol.* **2025**, *12*, 101384. [[CrossRef](#)]
25. Redmon, J.; Farhadi, A. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [[CrossRef](#)]
26. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YoloX: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [[CrossRef](#)]
27. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [[CrossRef](#)]
28. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.
29. Qi, S.; Song, X.; Shang, T.; Hu, X.; Han, K. Msfe-yolo: An improved yolov8 network for object detection on drone view. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6013605. [[CrossRef](#)]
30. Xiong, X.; He, M.; Li, T.; Zheng, G.; Xu, W.; Fan, X.; Zhang, Y. Adaptive feature fusion and improved attention mechanism-based small object detection for UAV target tracking. *IEEE Internet Things J.* **2024**, *11*, 21239–21249. [[CrossRef](#)]
31. Qin, Z.; Chen, D.; Wang, H. MCA-YOLOv7: An improved UAV target detection algorithm based on YOLOv7. *IEEE Access* **2024**, *12*, 42642–42650. [[CrossRef](#)]
32. Xie, T.; Wang, L.; Wang, K.; Li, R.; Zhang, X.; Zhang, H.; Yang, L.; Liu, H.; Li, J. FARP-Net: Local-global feature aggregation and relation-aware proposals for 3D object detection. *IEEE Trans. Multimed.* **2023**, *26*, 1027–1040. [[CrossRef](#)]
33. Purkait, P.; Zhao, C.; Zach, C. SPP-Net: Deep absolute pose regression with synthetic views. *arXiv* **2017**, arXiv:1712.03452. [[CrossRef](#)]
34. Ren, X.; Sun, M.; Zhang, X.; Liu, L.; Zhou, H.; Ren, X. An improved mask-RCNN algorithm for UAV TIR video stream target detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102660. [[CrossRef](#)]
35. Zhang, X.; Wang, C.; Jin, J.; Huang, L. Object detection of VisDrone by stronger feature extraction FasterRCNN. *J. Electron. Imaging* **2023**, *32*, 013018. [[CrossRef](#)]
36. Hu, Q.; Wang, L. HF-D-FINE: High-resolution features enhanced D-FINE for tiny object detection in UAV image. *Image Vis. Comput.* **2026**, *165*, 105834. [[CrossRef](#)]
37. Liang, Y.; Yang, L.; Sun, S.; Li, Z.; Shi, Y.; Zhang, Z.; Zhang, H.; Li, Z.; Zhou, L.; Zhang, Z.; et al. YOLOv11-RAH: A recurrent attention-enhanced edge intelligence network for UAV-based power transmission line insulator inspection. *Int. J. Intell. Netw.* **2025**, *6*, 244–252. [[CrossRef](#)]
38. Hnida, Y.; Mahraz, M.A.; Riffi, J.; Achebour, A.; Yahyaouy, A.; Tairi, H. Transfer learning-enhanced deep learning for tree crown geometric analysis and crop yield estimation using UAV imagery. *Remote Sens. Appl. Soc. Environ.* **2025**, *39*, 101663. [[CrossRef](#)]
39. Patel, M.K.; Bull, G.; Egan, L.M.; Swain, N.; Rolland, V.; Stiller, W.N.; Conaty, W.C. High-throughput Verticillium wilt detection in cotton: A comparative study of faster R-CNN and YOLOv11. *Biosyst. Eng.* **2026**, *263*, 104379. [[CrossRef](#)]
40. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 7373–7382.
41. Islam Nayeem, N.; Mahbuba, S.; Disha, S.I.; Buiyan, M.R.H.; Rahman, S.; Abdullah-Al-Wadud, M.; Uddin, J. A YOLOv11-Based Deep Learning Framework for Multi-Class Human Action Recognition. *Comput. Mater. Contin.* **2025**, *85*, 1541–1557. [[CrossRef](#)]
42. Albright, P.C.; Martin, S.M.; Rose, C.G. Evaluating 3D Gesture Recognition in UAV-Based Human-Robot Interaction. *IFAC-PapersOnLine* **2025**, *59*, 371–376. [[CrossRef](#)]

43. Qi, Y.; Lin, P.; Yang, G.; Liang, T. Crack detection and 3D visualization of crack distribution for UAV-based bridge inspection using efficient approaches. *Structures* **2025**, *78*, 109075. [\[CrossRef\]](#)
44. Haque, K.M.S.; Joshi, P.; Subedi, N. Integrating UAV-based multispectral imaging with ground-truth soil nitrogen content for precision agriculture: A case study on paddy field yield estimation using machine learning and plant height monitoring. *Smart Agric. Technol.* **2025**, *12*, 101542. [\[CrossRef\]](#)
45. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Machine Learning and Knowledge Discovery in Databases*; Springer Nature: Cham, Switzerland, 2022; pp. 443–459.
46. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 390–391.
47. Gao, W.; Zhang, X.; Yang, L.; Liu, H. An improved Sobel edge detection. In *Proceedings of the 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, 9–11 July 2010*; IEEE: Piscataway, NJ, USA, 2010; Volume 5, pp. 67–71.
48. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Republic of Korea, 27–28 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 213–226.
49. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 370–386.
50. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [\[CrossRef\]](#)
51. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [\[CrossRef\]](#)
52. Jocher, G.; Stoken, A.; Borovec, J.; Liu, C.; Hogan, A.; Diaconu, L.; Dave, P. *ultralytics/yolov5*, version 3.0; Zenodo: Geneva, Switzerland, 2020.
53. Varghese, R.; Sambath, M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, 18–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
54. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. YOLOv10: Real-time end-to-end object detection. In *NIPS '24, Proceedings of the 38th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024*; Curran Associates Inc.: Red Hook, NY, USA, 2024; pp. 107984–108011.
55. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [\[CrossRef\]](#)
56. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
57. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv* **2025**, arXiv:2509.25164. [\[CrossRef\]](#)
58. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 16–22 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.
59. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. *arXiv* **2024**, arXiv:2410.13842.
60. Zhang, Z. Drone-YOLO: An efficient neural network method for target detection in drone images. *Drones* **2023**, *7*, 526. [\[CrossRef\]](#)
61. Lu, L.; He, D.; Liu, C.; Deng, Z. MASF-YOLO: An Improved YOLOv11 Network for Small Object Detection on Drone View. *arXiv* **2025**, arXiv:2504.18136. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.