

MDFR-Net: Towards Enhanced Feature Refinement for Aerial Small Object Detection

Jing Wang

wj@hebust.edu.cn

Hebei University of Science and Technology

Jia Su

Hebei University of Science and Technology

Yanli Hou

Hebei University of Science and Technology

Boning Hu

Hebei University of Science and Technology

Article

Keywords: Deep learning, Complex background, Object detection, Aerial image, Small target, YOLOv11

Posted Date: September 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9719013/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MDFR-Net: Towards Enhanced Feature Refinement for Aerial Small Object Detection

Jing Wang^a, Jia Su^a, Yanli Hou^a, Boning Hu^a

^a School of Information Science and Engineering, Hebei University of Science and Technology, shijiazhuang, 050018, hebei, China
Jing Wang(corresponding author) E-mail: [wj@hebut.edu.cn]

Abstract

Detecting tiny objects in aerial imagery remains challenging due to their minimal scale, diverse orientations, and complex background interference, limiting the ability of existing methods to capture discriminative features. To address these limitations, we propose MDFR-Net with the following key contributions. Firstly, The Multi-scale Hierarchical Attentional Fusion (MHAF) module leverages hierarchical convolution layers and dual-path attention to jointly extract salient features while suppressing background noise. Next, The Orientation Decoupled Feature Enhancer (ODFE) module incorporates direction-separated paths with lightweight attention to focus on orientation-specific target details. Additionally, The Hierarchical Convolution Pooling Fusion (HCPF) module employs parallel dilated convolutions and pooling operations to capture cross-scale contextual information. Finally, A high-resolution P2 detection layer maintains focus on tiny objects in deep network stages. Experiments on benchmark aerial datasets VisDrone and AI-TOD demonstrate that MDFR-Net significantly advances small object detection accuracy in aerial images.

Keywords: Deep learning, Complex background, Object detection, Aerial image, Small target, YOLOv11

1. Introduction

With the rapid development of drone platforms and remote sensing technology, aerial image target detection has become increasingly vital in critical areas such as urban management, land resource exploration, and disaster emergency response. In urban management, drones can monitor illegal constructions and traffic flow in real-time; in resource exploration [1], they can effectively identify surface mineral deposits and vegetation coverage; in disaster relief scenarios, they can swiftly locate trapped individuals and damaged facilities. These applications require high precision and real-time performance from detection algorithms. Aerial images, which are characterized by complex terrains, variable lighting conditions, and high oblique viewing angles, pose challenges such as the coexistence of multi-scale targets, severe occlusions, and a high proportion of small targets, thereby further increasing the demands on detection algorithms [2].

Traditional object detection algorithms, such as the Scale-Invariant Feature Transform (SIFT) [3], primarily rely on manually designed feature operators to extract local features. These features are then used by machine learning methods like Support Vector Machines (SVM) [4] and Random Forests to perform classification tasks. While these methods have improved the accuracy and efficiency of aerial image detection, they suffer from high computational complexity, struggle to capture advanced semantic information in complex images, and lack robustness due to their poor adaptability to diverse environmental conditions and background changes.

With the rapid advancement of deep learning technology, neural network-based object detection methods have supplanted traditional approaches that relied on manually designed feature extraction, becoming the dominant paradigm in the

field. Deep learning-based object detection algorithms are primarily divided into two categories: two-stage and one-stage algorithms. Two-stage algorithms, such as R-CNN [5], Fast R-CNN [6], and Faster R-CNN [7], improve detection accuracy by first identifying candidate Regions of Interest (RoIs) and subsequently performing detailed classification and localization. This multi-stage processing ensures high detection accuracy but demands substantial computational resources and is relatively slower. In contrast, one-stage detection algorithms, such as SSD [8] and the YOLO series, bypass the complex process of generating candidate boxes. By directly applying predefined anchor boxes at the image level, they streamline the object detection task, reducing computational steps and enabling efficient real-time detection.

Redmon et al. introduced the YOLO (You Only Look Once) [9] algorithm, which partitions the entire image into a grid, with each grid cell predicting the position and category of the target box. Building upon YOLOv1, Redmon et al. further developed YOLOv2 [10], enhancing detection accuracy through batch normalization, multi-scale training, and the incorporation of prior boxes. Redmon et al. proposed YOLOv3 algorithm [11], which constructed a more concise and efficient backbone network structure with the idea of residual structure, introduced feature pyramid network to realize multi-scale feature fusion, and achieved a better balance between accuracy and speed. Bochkovskiy et al. introduced YOLOv4 [12], which leveraged Mosaic and Mixup data augmentation techniques to further boost detection accuracy and speed. The Ultralytics team launched YOLOv5, which accelerated model training and convergence while providing models of varying sizes for diverse tasks. Wang et al. introduced YOLOv7 [13], which incorporated the concept of model re-parameterization and a dynamic label allocation mechanism to enhance detec-

tion accuracy without increasing inference load. The Ultralytics team then introduced YOLOv8, which decouples the head to separately handle classification and regression tasks and employs the Anchor-Free method for regression, thereby improving both detection accuracy and generalization. Wong et al. introduced YOLOv9 [14], which reduced the model's parameters and computational load through programmable gradient information and generalized efficient layer aggregation networks. Finally, Wang et al. introduced YOLOv10 [15], which achieved real-time end-to-end object detection by eliminating the need for Non-Maximum Suppression (NMS) during both training and inference via dual label allocation.

As the latest version of the object detection algorithm released by Ultralytics in 2024, YOLOv11 [16] has significantly improved detection accuracy while maintaining high real-time performance by enhancing the model's neck and backbone modules and optimizing its architecture. However, it still struggles to effectively extract key features and accurately identify targets in aerial images with dense backgrounds and complex environments.

To address these challenges, this paper proposes the MDRF-Net algorithm based on YOLOv11, which improves the detection accuracy of small targets while keeping the model lightweight. The main contributions of this study are as follows.

1. We designed the Multi-scale Hierarchical Attentional Fusion (MHAF) module, which constructs a three-level depthwise separable convolution pathway to extract multi-scale features. It optimizes feature fusion through dual channel-spatial attention mechanisms, effectively suppressing background interference while highlighting critical information of small targets.
2. An Orientation Decoupled Feature Enhancer (ODFE) module was proposed, enhancing small target features via convolution stacking and directional path enhancement. A lightweight feature refinement layer applies importance weighting, significantly improving the model's capture capability for directionally sensitive features of aerial small targets.
3. By developing the Hierarchical Convolution Pooling Fusion (HCPF) module, which cascades convolution groups with different dilation rates for multi-receptive-field feature extraction and concatenates these features with multi-level pooled features, we formed a robust cross-scale contextual representation that enhances model adaptability to drastic target size variations.

2. Related work

This section focuses on reviewing the latest research in the field of small target detection in UAV images, particularly in relation to our proposed method. Many scholars have studied issues such as the high proportion of small targets, their dense distribution, complex backgrounds, and diverse scales in UAV datasets. The main topics covered include small target detection in UAV images, feature expression capabilities, and attention mechanisms in this section.

2.1. Small target detection in UAV images

Drone aerial image target detection represents a classic small object detection task. Within drone imagery, small objects are typically numerous, densely distributed, and exhibit weak semantic features due to challenges like low resolution and occlusion. To address these difficulties, researchers have conducted extensive studies. Wang et al. introduced AODet [17], a Transformer-based remote sensing target detection model. Leveraging the relatively simple backgrounds common in remote sensing imagery, AODet identifies background regions early, drastically reducing computational redundancy and improving small target detection accuracy. However, its performance degrades significantly when small targets dominate the scene or foreground objects occupy most of the image. Chen et al. proposed the Hyneter model [18], which employs a hybrid network backbone to integrate and transmit local and global information in parallel, markedly enhancing small target detection. This improvement, however, comes at the cost of increased parameters and network size. Li et al. developed OGMN [19], a multi-task guided network for detecting occluded small objects. It resolves feature confusion caused by overlapping targets through an occlusion localization module and a decoupling head, effectively suppressing background interference and optimizing detection performance for occluded objects. The model's high computational complexity, however, hinders practical deployment. Chen et al. introduced a semantic information-guided fusion module for small targets [20]. This approach uses high-level semantic features to guide and align low-level texture information, enhancing small targets' feature-level semantic representation and boosting detection accuracy. The added complexity slows inference speed. Li et al. designed AD-Det [21], a detection network focused on enhancing small target features. Its Adaptive Small Object Enhancement (ASOE) module leverages global features for accurate detection but overlooks interactions between global image context and local regions, potentially limiting detection robustness. Muzamul et al. utilized the Slice Assisted Hyper Inference (SAHI) method [22], applying slicing and stitching techniques to improve detection of small, dense objects. While effective, its robustness varies significantly across environmental conditions. Zhou et al. proposed KLDet [23], which employs Kullback-Leibler divergence (KL-divergence) as a positioning metric to precisely measure tiny bounding box localization. This enhances model performance but may increase missed detections and false alarms.

2.2. Feature expression ability

UAV images frequently exhibit uneven feature scale variations and significant inter-object proportional differences, posing substantial challenges to the feature representation capabilities of target detection models. To address these limitations, enhancing feature extraction capacity, fully exploiting effective feature information, and improving multi-scale fusion have emerged as critical research priorities. Xiao et al. introduced a UAV detection method using the Feature Complementary Mapping (FCM) module [24], which integrates rich semantic in-

formation with precise spatial localization to enhance small-target feature matching in deep networks, though at the cost of reduced processing speed and compromised real-time performance. Min et al. proposed LWUAVDet [25], extending the Feature Pyramid Network to incorporate large-scale feature maps to promote multi-scale information exchange and strengthen positional awareness of small targets, albeit with computational redundancy in dense scenes. Duan et al. developed a multi-channel weight-learnable recursive aggregation network [26] that dynamically fuses multi-level scale features to boost multi-scale extraction efficacy, yet it suffers persistent false positives and missed detections under variable lighting in complex environments. Xiao et al. designed an Feature Complementary Mapping Module (FCM) [27] using grouped feature focusing units and multi-level reconstruction to enhance semantic representation, though detection accuracy remains inadequate in low-light conditions. Yin et al. created SDPDet for end-to-end UAV detection [28], employing Scale-separated Activation Pyramid (SAP) and Scale-separated Learnable Proposals (SLP) mechanisms to dynamically adjust proposals per scale, effectively mitigating uneven scale distribution despite reduced classification accuracy for targets with subtle visual differences. Mao et al. devised ESPP [29], an progressive Spatial Pyramid Pooling method with Progressive PSP and Multi-scale Feature Enhancement Module (MFEM) to capture richer nonlinear features and establish long-range dependencies, yet it remains susceptible to false positives in low-quality images. Jiang et al. proposed MFFSODNet [30], a multi-scale fusion network enhanced by a Multi-scale Feature Extraction Module (MSFEM) to refine fine-grained small-target information, though parallel extraction across branches with differing convolutional kernels increases computational load.

2.3. Attention mechanism

In recent years, attention modules have gained widespread adoption across diverse deep learning domains, including natural language processing and computer vision. Their efficacy is particularly pronounced in object detection, where they selectively emphasize regions of interest by reweighting features to suppress noise, offering a powerful solution for mitigating background interference. Researchers have consequently refined remote sensing object detection through continuous enhancements to attention mechanisms. Ma et al. introduced a Scale Decoupling Module (SDM) [31] employing Sparse Non-Local Attention (SNLA) to adaptively capture sparse contextual location information, thereby strengthening small target feature representation, though it tends to treat dense targets of identical categories as single entities. Song et al. developed the Bottleneck Attention Based Fusion Network (BAFNet) [32], which integrates high-level semantic and low-level spatial features via a dual-stream attention mechanism alongside boundary-aware supervision, capturing distinctive foreground-background cues to highlight small target saliency, albeit with diminished performance on medium and large targets. Huang et al. proposed the BRSTD tiny target detection network [33], incorporating a Top-Down Suppression Attention (TDSA) that utilizes depth information to modulate shallow features, better pre-

serving small object details in early layers while suppressing larger objects and backgrounds, though introducing considerable computational overhead. Ge et al. presented the Multi-attention and Adaptive Reparameterized Feature Pyramid Network (MARFPNet) [34], embedding multi-dimensional and multi-scale attention within the feature extractor to focus on small targets, yet its high complexity hinders deployment on resource-constrained drones. Lin et al. devised a high-precision detection method based on PFRNet [35], fusing multi-scale features via spatial pyramid pooling and a spatial pyramid pooling fusion with spatial attention (SPPFSPA) to concentrate on salient regions, yet it contends with imprecise bounding box localization and inter-object interference affecting feature extraction accuracy. Ding et al. [36] introduced a real-time small object detection algorithm leveraging a lightweight orthogonally compressed attention module to generate semantically rich attention maps via weighted feature integration, yet detection suffers under low-exposure conditions due to object feature interference. In summary, while existing work enhances feature expression and refines attention to suppress background noise, improving feature extraction and fusion capabilities, the resultant increases in model parameters and computational complexity impede deployment on computationally limited drone platforms. Furthermore, current models continue to face difficulties in achieving effective feature extraction and precise target recognition within dense, complex environments.

3. Proposed method

YOLOv11 significantly enhances feature extraction capabilities and improves object detection accuracy as well as performance in complex scenarios by refining both its backbone network and neck architecture. In the backbone, the model replaces YOLOv8's C2f structure with a C3k2 structure and adds a C2PSA module after the SPPF layer to strengthen feature extraction. In the detection head, Depthwise separable Convolution (DWConv) [37] is introduced, and the regression branch loss function is replaced with a combination of Distribution Focal Loss (DFL) [38] and CIoU [39], effectively improving the accuracy of the regression of the bounding box. YOLOv11 offers five model variants with different complexities (YOLOv11n/s/m/l/x), allowing researchers to flexibly select the most suitable model according to their specific application requirements.

To address key challenges in aerial image object detection—such as the loss of small target features, limited direction sensitivity, and weak multi-scale adaptability, this paper introduces an enhanced model called MDR-Net, as shown in Fig. 1. We propose the Multi-scale Hierarchical Attentional Fusion module (MHAF), which combines multi-branch depthwise separable convolutions and a dual-path attention mechanism to better extract target feature information; an Orientation Decoupled Feature Enhancer module (ODFE) that employs parallel main and directional pathways with separable convolutions to bolster direction-sensitive feature representation; a Hierarchical Convergence Pooling Fusion module (HCPF), which fuses multi-scale dilated convolutions and hierarchical convergence to cap-

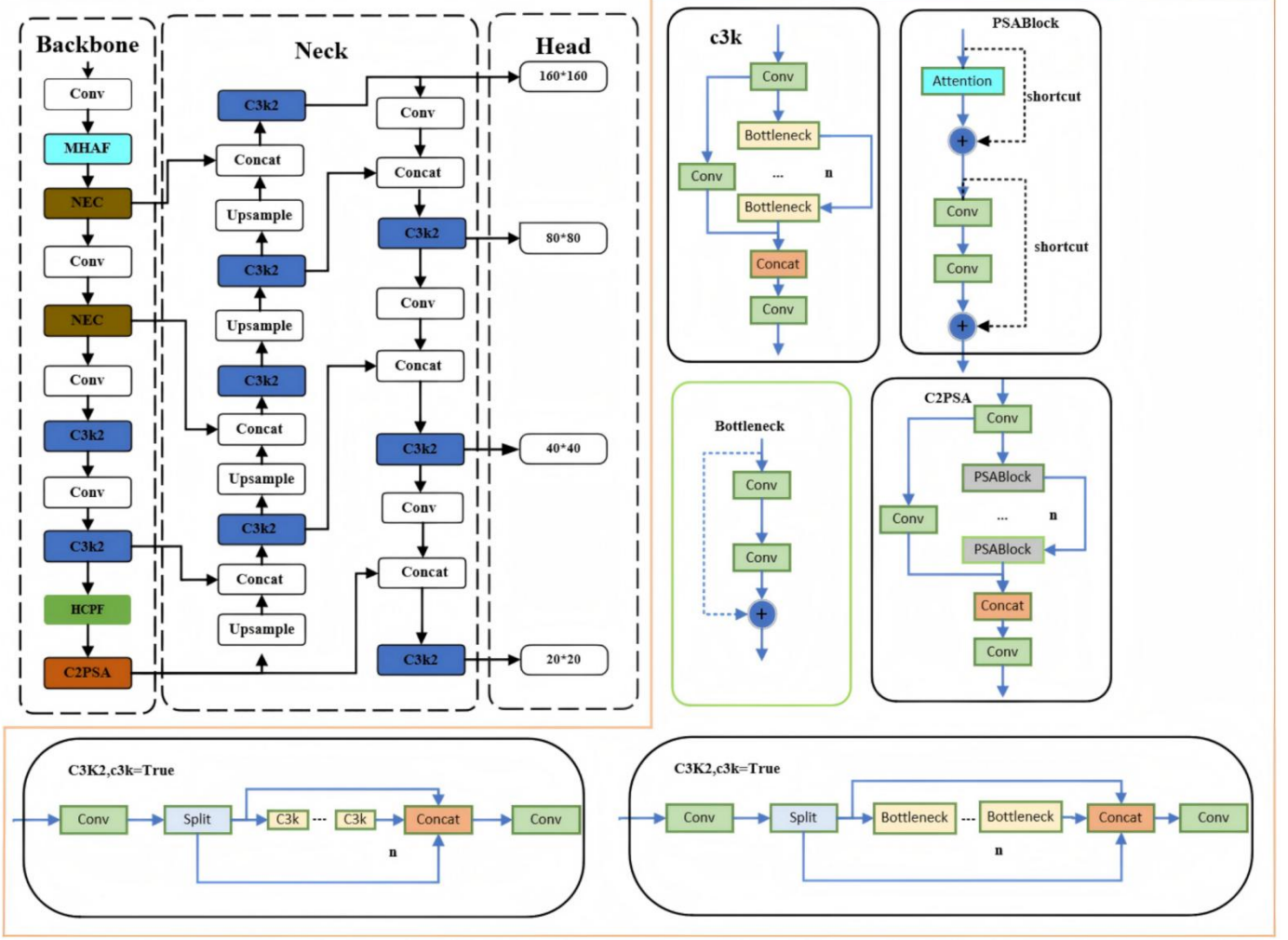


Figure 1: MDFR-Net network structure

ture cross-scale contextual information; and a high-resolution P2 detection layer [40] that effectively preserves the spatial structure of small targets. These modules work together synergistically to create a robust feature enhancement mechanism, significantly boosting the accuracy of small target detection in aerial imagery while keeping the increase in model complexity reasonable.

3.1. MHAF

To address the challenge of losing edge detail features and effectively capturing them in aerial images [41], this paper introduces a Multi-scale Hierarchical Attentional Fusion module (MHAF), as shown in Fig. 2. This module improves feature extraction by leveraging hierarchical feature decoupling combined with an attention mechanism based on deep separable convolutions. The design begins with a Multi-stage Decomposition Convolution (MDC) featuring a three-level branch architecture. In the first stage, a 3×3 depthwise separable convolution [37] is applied to model global semantic information from input features, splitting the output channels into a global semantic branch and a medium-scale enhancement branch. The

second stage applies another 3×3 depthwise separable convolution, further dividing the channels into a medium-scale contour branch and a high-resolution feature branch. In the final stage, a 5×5 large-kernel depthwise separable convolution is introduced, which preserves the spatial resolution of the feature map while using an expanded convolution kernel to capture the edge details of small targets. This enhances cross-scale feature representation, spanning from global semantics to local details. By progressively compressing channel dimensions and increasing the convolution kernel size, this hierarchical division strategy not only reduces computational complexity but also constructs a feature pyramid that covers different levels. This effectively addresses the limitation of traditional single-scale convolutions, which often fail to simultaneously represent both the overall structure and the fine local details of the targets.

To further strengthen the capacity for multi-scale feature discrimination, the MHAF module incorporates a Dual-Path Feature Fusion Module (DPFM) following the fusion of convolution branches. This module consists of two parallel paths: the upper path implements a channel attention mechanism [42], generating channel weights by extracting inter-channel rela-

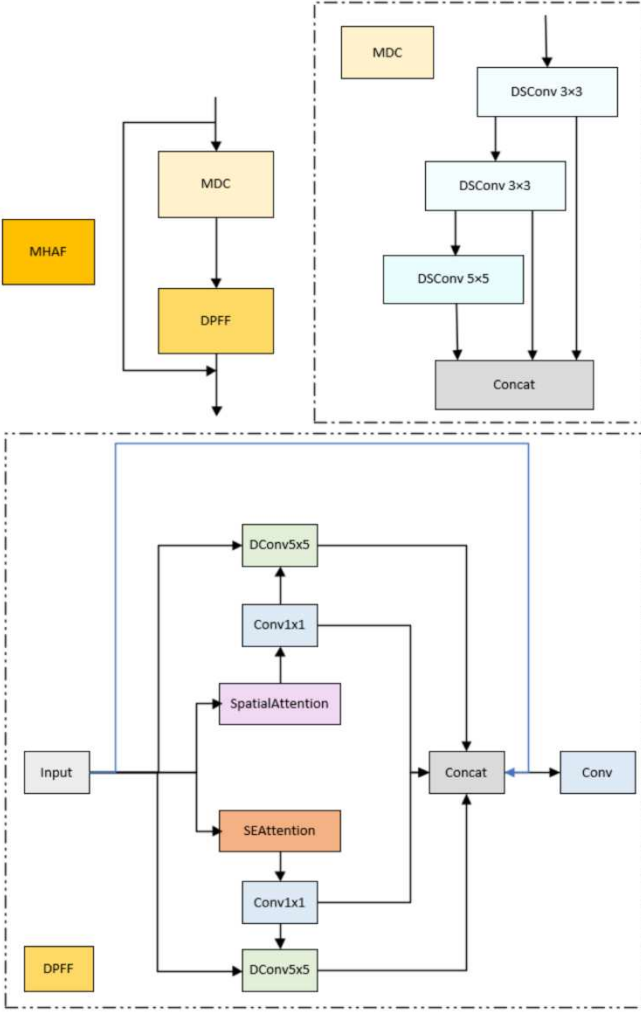


Figure 2: Multi-scale Hierarchical Attentional Fusion module

tionships, thereby enhancing target-related features and suppressing background interference. The lower path employs spatial attention, modeling spatial positional relationships to produce a spatial weight matrix that amplifies activation in target regions while diminishing responses in the background. After each path refines its respective features, 1×1 convolutions facilitate cross-channel information interaction, and a 5×5 dilated convolution [43] layer is applied to expand the receptive field. Together, these steps substantially enhance the module’s ability to capture multi-scale contextual information, thereby improving the adaptability to targets of varying scales.

The optimization of the dual-path output features is achieved through element-wise addition, which not only preserves critical semantic information across the channel dimension but also enhances spatial positioning accuracy. The fused features are then combined with the original input features via a residual connection, introducing deeper-level features while retaining shallow-level details. Finally, feature integration is completed through a 3×3 convolution layer. Positioned at the third level of the main network, this layer leverages the synergistic effects of the channel and spatial dimensions to integrate initial semantic

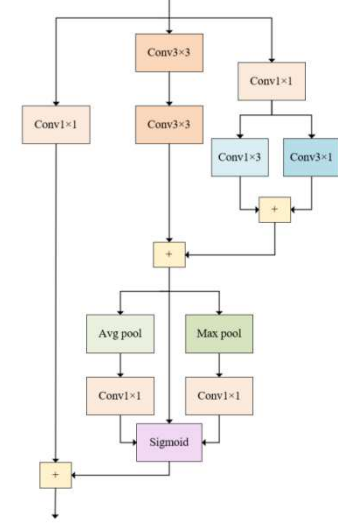


Figure 3: Orientation Decoupled Feature Enhancer module

information while preserving sufficient spatial detail, thereby effectively mitigating interference from complex backgrounds.

3.2. ODFE

Since the C3k2 module of YOLOv11 relies on standard convolution and bottleneck operations for feature extraction, its ability to capture features of small targets is limited when applied in the shallow layers of the main network, resulting in early loss or dilution of critical feature information [44]. To overcome this limitation, this paper introduces an Orientation Decoupled Feature Enhancer (ODFE) module, which replaces the first two C3k2 modules in the main network, as shown in Fig. 3. The dual-path parallel structure more effectively extracts and enhances fundamental features containing directional details. The main path consists of two consecutive 3×3 convolutions, primarily responsible for extracting the main contours, textures, and other basic feature information. This stacked design progressively captures an increasingly abstract yet stable spatial context, providing reliable global contextual information for subsequent processing. Meanwhile, the direction path employs a 1×1 convolution to reduce feature dimensions to one-eighth of the original channels, followed by parallel 1×3 horizontal and 3×1 vertical convolution branches that independently extract features in the horizontal and vertical directions [45]. This directional decomposition improves sensitivity to multi-directional targets in aerial images, facilitating their identification amid complex backgrounds. Finally, features from both paths are fused by concatenation along the channel dimension, combining the global representation from the main path with the detailed directional information from the direction path, thus forming a richer and more discriminative feature representation.

The fused features are then passed to the feature optimization layer, which aims to emphasize key feature information while suppressing redundant data. First, global maximum and average pooling [46] operations are applied to extract the global

statistical information of the feature map, capturing both prominent responses and overall distribution. The outputs from these pooling operations are then independently processed through separate 1×1 convolution layers to learn nonlinear channel-wise interactions. These processed results are combined via element-wise addition and passed through a Sigmoid function [47] to generate a feature weight map with values ranging between 0 and 1. This weight map automatically highlights the relative importance of different spatial positions and channels within the feature map, emphasizing areas containing small targets or critical directional details. By multiplying the fused features with this weight map channel-wise, the expression of small targets and key details is strengthened while background interference is attenuated. Finally, the optimized features are added to a residual connection formed by a 1×1 convolution of the original input, producing the final output. This residual design ensures effective preservation of the original input information, stabilizes the training process, and mitigates the vanishing gradient problem. Ultimately, the ODFE module is able to focus on learning the most informative aspects of the input features, significantly enhancing the detection performance for aerial image targets through direction-sensitive feature enhancement coupled with a lightweight attention mechanism.

3.3. HCPF

While traditional spatial pyramid pooling methods like the SPPF module have optimized multi-scale feature fusion, they remain limited by insufficient receptive field diversity and shallow cross-level interaction. To overcome these constraints, this paper proposes the Hierarchical Convolution Pooling Fusion (HCPF) module, depicted in Fig. 4, as an effective SPPF alternative. The HCPF's core innovation lies in its multi-level extraction mechanism and cross-scale fusion capability.

Central to this architecture is the Hierarchical Dilated Convolution (HDC) component, which executes cross-receptive-field extraction and fusion. An initial 1×1 convolution transforms features and adjusts channel dimensions to condition the input space. Three parallel dilated convolution groups [43] then operate simultaneously: the first group with dilation rate 1 captures finer details, the second with dilation rate 2 extracts intermediate spatial relationships, and the third with dilation rate 3 acquires broader contextual associations. These independently processed feature streams undergo channel-wise concatenation to form an integrated multi-scale tensor. A concluding 3×3 convolution layer facilitates cross-channel interaction, achieving deep multi-scale integration with dimensionality reduction while preserving computational efficiency through optimized spatial aggregation.

The HCPF module commences processing by applying the initial HDC convolution to input features for preliminary multi-scale enhancement. Subsequent max pooling generates primary pooled outputs that intensify salient feature responses. This output then feeds dual parallel average pooling branches, producing secondary and tertiary pooled outputs to capture comprehensive data characteristics. These three pooling operations establish a multi-scale extraction hierarchy, each isolating distinct spatial context levels. The module then concate-

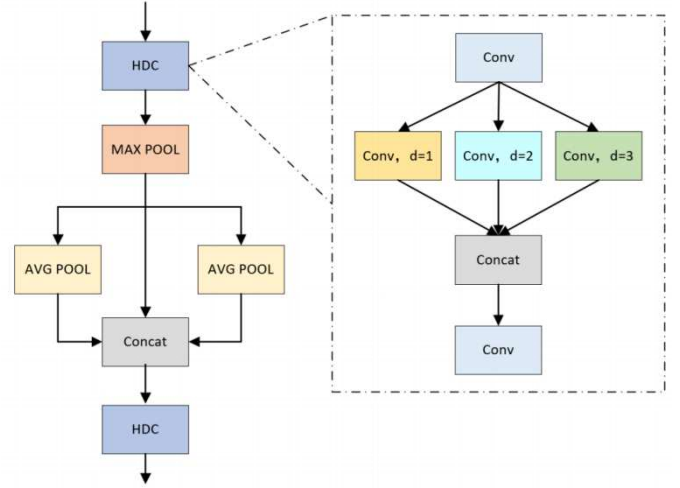


Figure 4: Hierarchical Convolution Pooling Fusion module

nates four components along the channel dimension (the original HDC output, primary, secondary, and tertiary pooled outputs) forming an enriched representation of multi-level spatial information. This composite tensor is processed through a secondary HDC convolution for deep feature refinement and dimensionality compression. Through this structured hierarchical processing pipeline, the HCPF module substantially improves model resilience to extreme scale variance while enhancing feature representation capabilities.

3.4. P2

In the aerial image detection task, to address the issue of missing detections of small targets (such as vehicles and pedestrian) due to the loss of pixel details, a high-resolution P2 detection layer [40] was introduced into the neck of YOLOv11. This layer operates at a down-sampling rate of 1/4 of the input image, generating a feature map of 160×160 pixels, which is four times the spatial resolution of the original P3 layer (80×80). This ensures that the edge textures and other key features of small targets are accurately retained. By integrating the high-resolution features from the shallow layers of the backbone network with the up-sampled semantic features from the deep layers, and enhancing feature consistency through a small convolution block, the detection head outputs a feature map of 160×160 pixels. This structural modification improves the recognition of small targets in the feature grid, significantly enhancing the detection accuracy of small targets and effectively addressing the issue of feature information loss due to the large receptive field of deep features.

4. Experiments and discussions

4.1. Datasets

To ensure the reliability and generalizability of the experimental results, this study employed two publicly available datasets for validation: VisDrone 2019 and AI-TOD. VisDrone 2019 [48] served as the primary dataset for comprehensive

comparative experiments and ablation studies, while AI-TOD [49] was used to further evaluate the generalization ability. The VisDrone 2019 dataset, collected and released by the Machine Learning and Data Mining Laboratory at Tianjin University, contains 8,629 images split into a training set of 6,471 images, a validation set of 548 images, and a test set of 1,610 images. It features 10 categories of common objects in everyday scenes, including pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning-tricycles, buses, and motorcycles. A major challenge posed by this dataset is the significant class imbalance and the presence of numerous small-scale targets. In contrast, the AI-TOD dataset is a specialized remote sensing dataset focused on tiny targets, comprising 28,036 images with 700,621 annotated objects across 8 categories: bridges, ships, vehicles, storage tanks, people, swimming pools, windmills, and airplanes. The core difficulty of AI-TOD lies in the extremely small average target size—approximately 12.8 pixels—which is much smaller than those in other aerial remote sensing datasets, thus greatly increasing detection complexity. This dataset is divided into training, validation, and test sets in a ratio of 6:1:3.

4.2. Experimental setup

All experiments were executed on an NVIDIA GeForce RTX 4070 GPU (8GB RAM) within a Python 3.9.19 environment using PyTorch 2.0.1, CUDA 11.8, and the Ultralytics YOLOv11 framework. To prevent GPU memory overflow, training employed a batch size of 4 with input images uniformly resized to 640×640 pixels. Optimization utilized Stochastic Gradient Descent (SGD) with initial learning rate 0.01, momentum 0.937, and cosine annealing for dynamic learning rate adjustment. Mosaic augmentation enhanced dataset diversity during training. To ensure complete model convergence and comparative fairness, all implementations trained from scratch without pretrained weights for 200 epochs.

4.3. Metrics of evaluation

Model accuracy validation employs comprehensive evaluation metrics: Precision (P) quantifies the accuracy of the positive prediction, while recall (R) measures the coverage of the true target, as shown in Equation (1) and (2).

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

The F1-score harmonizes these metrics to balance detection reliability and completeness, as shown in Equation (3). For robustness to localization, mAP50 evaluates the standard localization capability and mAP50-95 assesses the stability under stringent localization thresholds as shown in Equation (4) and (5).

$$F1 = \frac{2PR}{P + R} \quad (3)$$

$$AP = \int_0^1 p(r) dr \quad (4)$$

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \quad (5)$$

Parameter count directly determines computational complexity, governing inference speed and energy consumption; model size (weight file storage) dictates hardware loading efficiency. Collectively, these metrics define critical deployment constraints, where reduced parameter counts and minimized storage footprints signify superior edge computing adaptability.

4.4. Ablation experiments

To systematically evaluate the effectiveness of the proposed modules, ablation experiments were conducted on an aerial target detection dataset using YOLOv11s as the baseline model. The results are shown in Table 1.

The introduction of the Multi-scale Hierarchical Attentional Fusion (MHAF) module significantly improved the detection accuracy of the model. The experimental results show that the F1 score in the VisDrone dataset increased by 1.7 percentage points to 44.2%, the average precision (mAP50) improved by 2.2 percentage points to 40.7%, and mAP50-95 increased by 1.6 percentage points to 24.4%. This improvement is primarily attributed to the hierarchical feature decoupling design of the MHAF module, which employs parallel branches with convolution kernels of different sizes to enhance the ability to capture edge details of small targets in aerial images, thereby effectively boosting small target detection performance.

On that basis, the integration of the Orientation Decoupled Feature Enhancer (ODFE) module further improved the detection accuracy. The experiments indicate gains of 1.6 percentage points in F1 score, 1.8 percentage points in mAP50, and 1.2 percentage points in mAP50-95. These improvements stem from the dual path structure of the module, where one path focuses on extracting and improving information about horizontal features and the other on vertical features. This directional decoupling effectively strengthens the ability to represent targets with varying orientations in aerial images, enhancing adaptability and recognition accuracy.

To address the challenges posed by significant scale variations in aerial images, a Hierarchical Convolution Pooling Fusion (HCPF) module was incorporated into the model combining MHAF and ODFE. As a result, the overall detection performance improved markedly: the F1 score increased by 1.3 percentage points to 47.2%, mAP50 rose by 0.9 percentage points to 43.4%, and mAP50-95 increased by 0.7 percentage points to 26.3%. This advancement is largely due to the innovative multi-level dilated convolution and pooling collaborative architecture of the HCPF module, which efficiently integrates contextual information on all scales, enhancing the model's ability to capture and discriminate multi-scale target features in aerial scenes, thus mitigating detection difficulties caused by scale variations.

Finally, after integrating features from the MHAF, ODFE, and HCPF modules into a feature fusion layer, a high-resolution P2 detection layer was introduced. This addition significantly improved the performance of the model in detecting small targets, with the F1 score increasing by 1.4 percentage points to

Table 1: MDFR-Net module ablation experiments

MHAF	ODFE	HCPF	P2	mAP50(%)	mAP@50-95(%)	F1(%)
×	×	×	×	38.5	22.8	42.5
✓	×	×	×	40.6	24.4	44.2
✓	✓	×	×	42.5	25.6	45.8
✓	✓	✓	×	43.4	26.3	47.2
✓	✓	✓	✓	46.4	28.4	48.6

48.6%, mAP50 increasing by 3 percentage points to 46.4%, and mAP50-95 improving by 2.1 percentage points to 28.4%. The key to this enhancement lies in the P2 detection layer’s use of the 160×160 pixel shallow feature map output from the second stage of the backbone network. By directly leveraging features containing fine spatial details of small targets, this layer effectively addresses the problem of feature degradation and loss of small targets in deeper network layers, substantially improving both the representation and detection accuracy of small targets.

4.5. Comparative experiments

Based on the comparative analysis of the two precision-recall curves, a significant difference in detection performance between the baseline model and MDFR-Net is evident, depicted in Fig. 5. For categories with scarce samples, such as awning-tricycles, MDFR-Net maintains a stable detection precision of 20.5%, while for typical small-scale targets like bicycles, the precision remains at 19.0%, demonstrating MDFR-Net’s advantage in detecting smaller sample sizes. Furthermore, pedestrian detection precision reaches 54.6%, and motorcycle detection precision is 55.6%, both exceeding the baseline model by more than 10%, indicating the strong generalizability of MDFR-Net. The overall precision achieves 46.4%, reflecting an improvement of nearly 8% across all categories compared to the baseline model. This shows that the model consistently delivers reliable performance across various target types, significantly outperforming the baseline. These quantitative results illustrate that MDFR-Net not only maintains high detection accuracy but also handles complex samples effectively, thereby validating the effectiveness.

In the comparison on the VisDrone dataset, the proposed MDFR-Net model demonstrates comprehensive performance advantages. The results are shown in Table 2. Its mAP@0.5 reaches 46.4%, surpassing the current best model RT-DETR-R18 (41.9%) by 4.5 percentage points and the baseline YOLOv11s model (38.5%) by 7.9 percentage points. For small object detection, the APs metric achieves 19.4%, outperforming RT-DETR-R18 (15.1%) by 4.1 percentage points and the lightweight representative Gold-YOLO (12.9%) by 6.5 percentage points. In terms of model complexity, MDFR-Net achieves this improvement with 15.9 million parameters, reducing the parameter count by 4.1M compared to RT-DETR-R18 (20.0M). Its model size of 31.0 MB is also significantly smaller than YOLOXs (50.4 MB), saving 19.4 MB in storage. These experimental results fully validate that the model’s multi-dimensional feature optimization mechanism enables a substantial improvement in detection accuracy while increasing the pa-

rameter count by only 6.5M compared to the baseline, striking a favorable balance between model performance and complexity.

4.6. Visual analytics

The Fig. 6 illustrates the trade-off between accuracy and complexity for different models. The horizontal axis (x-axis) represents the number of parameters, reflecting computational complexity, while the vertical axis (y-axis) indicates the detection accuracy (mAP@50-95), measuring the performance of target recognition. Through its unique architectural design, MDFR-Net achieves a breakthrough in accuracy while controlling parameter growth, successfully positioning itself as a high-efficiency model in the balance between accuracy and complexity.

Through a comparative analysis of the normalized confusion matrices of the MDFR-Net model and the YOLOv11 model shown in Fig. 7, the proposed model shows a significant performance improvement in object detection tasks. Its core advantages lie in the optimization of key category recognition accuracy and effective suppression of background false positives. Within the vehicle category, the recall rate for cars increased from 0.75 to 0.80, indicating enhanced robustness in detecting high-frequency targets; the misclassification rate of vans as background decreased from 0.26 to 0.21, highlighting an improvement in feature representation capability. The background false positive rate showed a systematic reduction; for example, the false positive rate for pedestrians dropped substantially from 0.61 to 0.48, significantly alleviating the false alarm problem caused by background interference in the baseline model. Particularly notable is the improved discrimination ability for special vehicles: the recall rate for buses surged from 0.44 to 0.51, and for motorcycles from 0.38 to 0.48, demonstrating stronger discriminative power for targets with complex shapes. Overall, the model enhances robust detection in complex traffic scenarios by strengthening key target discrimination and suppressing background noise, providing a more effective solution.

In the analysis of typical complex scenarios in the VisDrone validation set, MDFR-Net shows a significant performance improvement compared to the baseline model. As shown in Fig. 8, the detection confidence of the MDFR-Net model is significantly better than that of the baseline model under multiple scenarios and challenging conditions.

In high-density target scenarios, the baseline model exhibits noticeable missed detections and false positives, particularly in the central regions of aerial images, whereas MDFR-Net achieves a more balanced spatial distribution of predictions, thereby enhancing the completeness of recognition for densely

Table 2: Comparison results of different models on the VisDrone dataset

Models	mAP@50(%)	APs(%)	Model size(MB)	Params(M)
YOLOv5s	32.1	10.1	13.7	7.1
YOLOv6s[50]	31.2	9.4	38.7	18.5
YOLOXs[51]	33.5	10.3	50.4	8.9
YOLOv7-tiny[13]	36.8	11.8	11.5	6.0
YOLOv8s	38.7	12.4	21.4	11.1
YOLOv9s[14]	37.9	12.1	15.8	7.3
YOLOv10s[15]	37.6	12.0	16.6	8.1
YOLOv11s[16]	38.5	12.3	18.3	9.4
RT-DETR-R18[52]	41.9	15.1	38.5	20.0
ASF-YOLO[53]	40.6	14.8	22.8	11.3
Gold-YOLO[54]	39.9	12.9	26.3	13.4
MDFR-Net	46.4	19.4	31.0	15.9

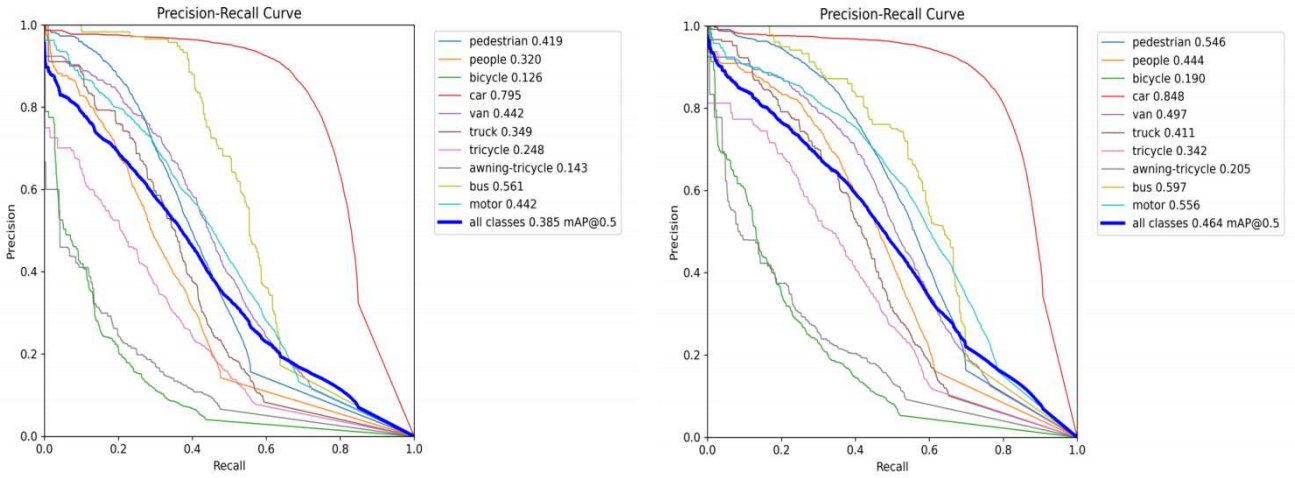


Figure 5: Comparison of precision-recall curves

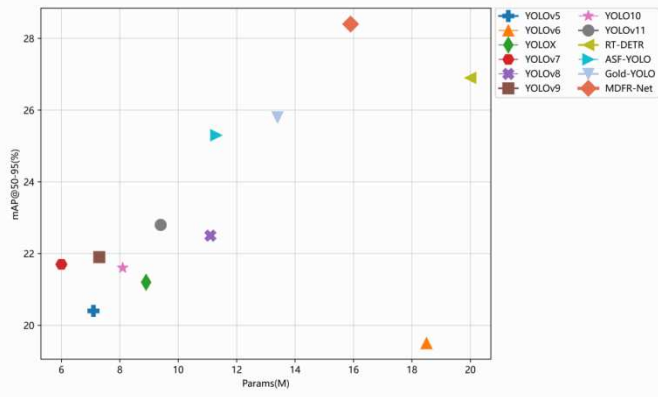


Figure 6: Relationships between the mAP@50-95 and model parameter count for different models

packed targets. In multi-scale detection scenarios, the baseline model suffers from significant cross-scale false detections due to feature confusion, while MDFR-Net substantially improves detection confidence across different scales. In partially occluded scenes, such as vehicles obscured by tree branches, the

baseline model bounding boxes show marked localization deviations, while MDFR-Net maintains accurate target positioning and effectively suppresses erroneous detections caused by background interference. Under low-light nighttime conditions, the baseline model produces responses to non-target objects in regions with low signal-to-noise ratios, but MDFR-Net considerably reduces such noise responses while preserving stable detection performance for targets under weak illumination. These visualization results collectively demonstrate that MDFR-Net improves the generalization and detection capabilities of the model in complex environments characterized by high target density, partial visual information loss, and poor lighting.

Comparative heatmap visualizations using Grad-CAM on the VisDrone dataset vividly illustrate the enhanced feature representation achieved by the improved model, as shown in Fig. 9. Relative to the baseline, MDFR-Net produces heatmaps characterized by significantly more concentrated and intense activations that are precisely localized within critical target regions, such as individuals in dense crowds or distant, small-scale vehicles set against complex backgrounds. These heightened activations are tightly confined within the spatial boundaries of

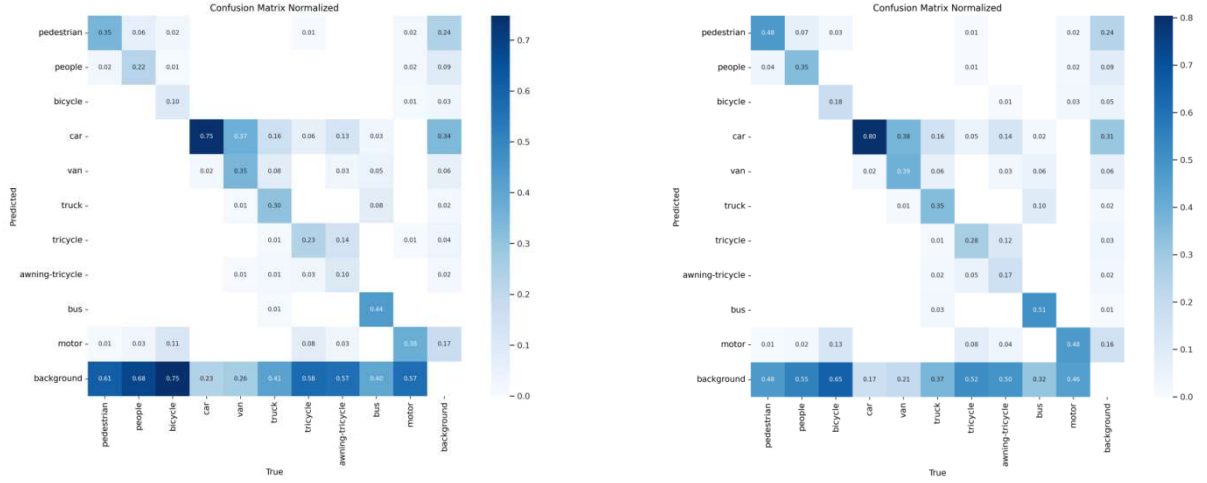


Figure 7: Comparison of confusion matrices



Figure 8: Comparison of YOLOv11(mid) and MDR-Net(right) on the VisDrone dataset

the target objects, reflecting the model's improved focus. At the same time, MDR-Net effectively suppresses responses in non-target areas, particularly in regions with intricate urban

textures, vegetation shadows, or distractor objects that resemble target shapes. This robust background suppression results in a pronounced intensity contrast between the high-response

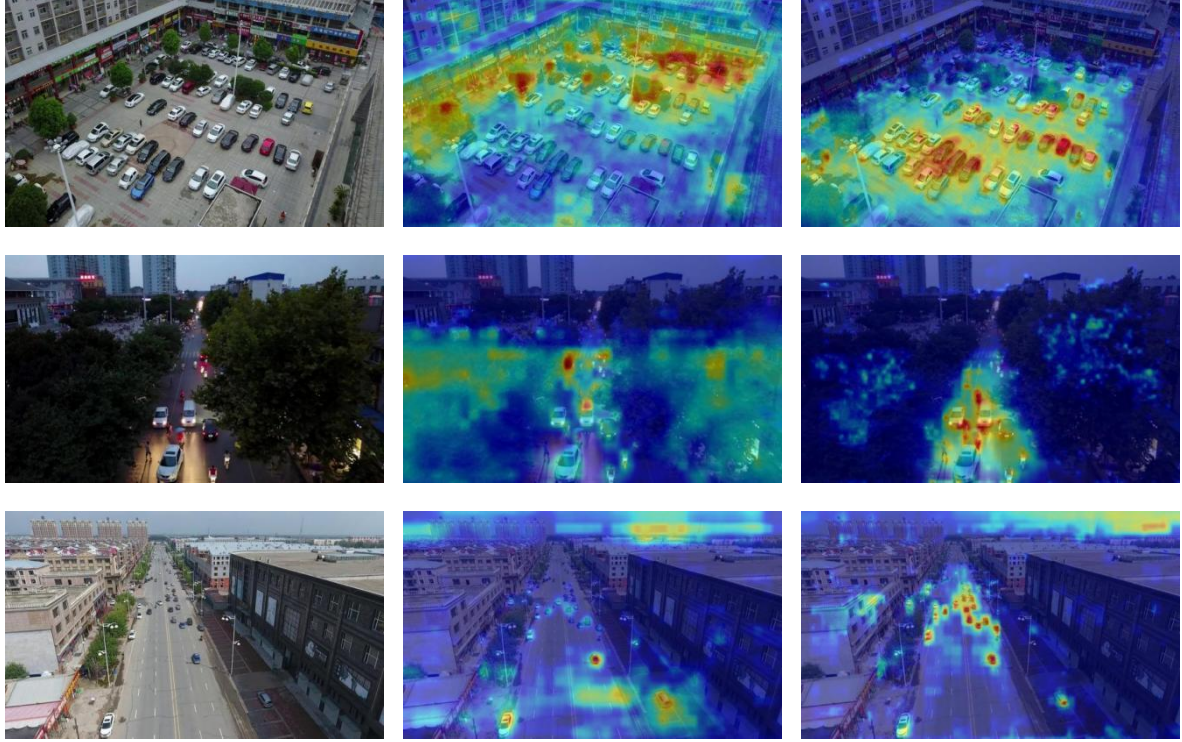


Figure 9: Comparison of the heatmaps of YOLOv11 (center) and MDFR-Net (right) on the VisDrone dataset

Table 3: Comparative experiments on the AI-TOD dataset

Models	mAP@50(%)	mAP@50-95(%)
YOLOv5s	42.4	18.9
YOLOv6s[50]	39.6	16.9
YOLOXs[51]	41.9	18.6
YOLOv7-tiny[13]	40.1	17.5
YOLOv8s	43.8	19.2
YOLOv9s[14]	43.2	18.7
YOLOv10s[15]	42.4	18.3
YOLOv11s[16]	42.9	18.6
RT-DETR-R18[52]	44.3	19.5
ASF-YOLO[53]	41.9	18.5
Gold-YOLO[54]	42.2	18.6
MDFR-Net	47.1	21.7

target zones and the low-response background regions in the heatmaps. As a result, MDFR-Net more distinctly highlights the discriminative features of detection targets, while minimizing irrelevant background interference in complex aerial scenes. This ultimately contributes to greater stability and reliability in detection performance.

4.7. Generalization experiments

The Table 3 compares the performance of MDFR-Net with several other advanced models on the AI-TOD dataset, focusing on the metrics of mAP@50 and mAP@50-95. The results highlight that MDFR-Net consistently outperforms competing algorithms in detection accuracy, achieving improvements of 4.2%

and 2.9% over the baseline model in mAP@50 and mAP@50-95, respectively. This clearly demonstrates MDFR-Net’s strong generalization capability in aerial detection tasks.

Further comparison of detection results on AI-TOD images, as shown in Fig. 10, reveals that MDFR-Net significantly surpasses the baseline model by effectively identifying a greater number of small targets, particularly in challenging conditions such as occluded scenes or low-resolution areas where the baseline often misses very small buildings or vehicles. In addition, the frequent false positives and false negatives commonly seen in the output of the baseline model, especially near the edges of the image or in regions with complex textures, are substantially reduced in the MDFR-Net results, producing clearer and more reliable detections. Additionally, MDFR-Net’s bounding boxes align more precisely with the edges of small targets, effectively mitigating localization bias. Collectively, these comparative findings demonstrate that the improved model enhances perception, robustness, and localization accuracy when confronting the difficulties presented by the AI-TOD dataset.

5. Conclusions

This paper proposes the MDFR-Net model, designed to address the core challenge of accurately detecting small targets in aerial images. The key innovations of MDFR-Net include the MHAF module, which significantly enhances focus on critical features, particularly edge details, through an attention mechanism while effectively suppressing background interference; the ODFE module, which innovatively employs a direction decoupling path to strengthen the model’s ability to

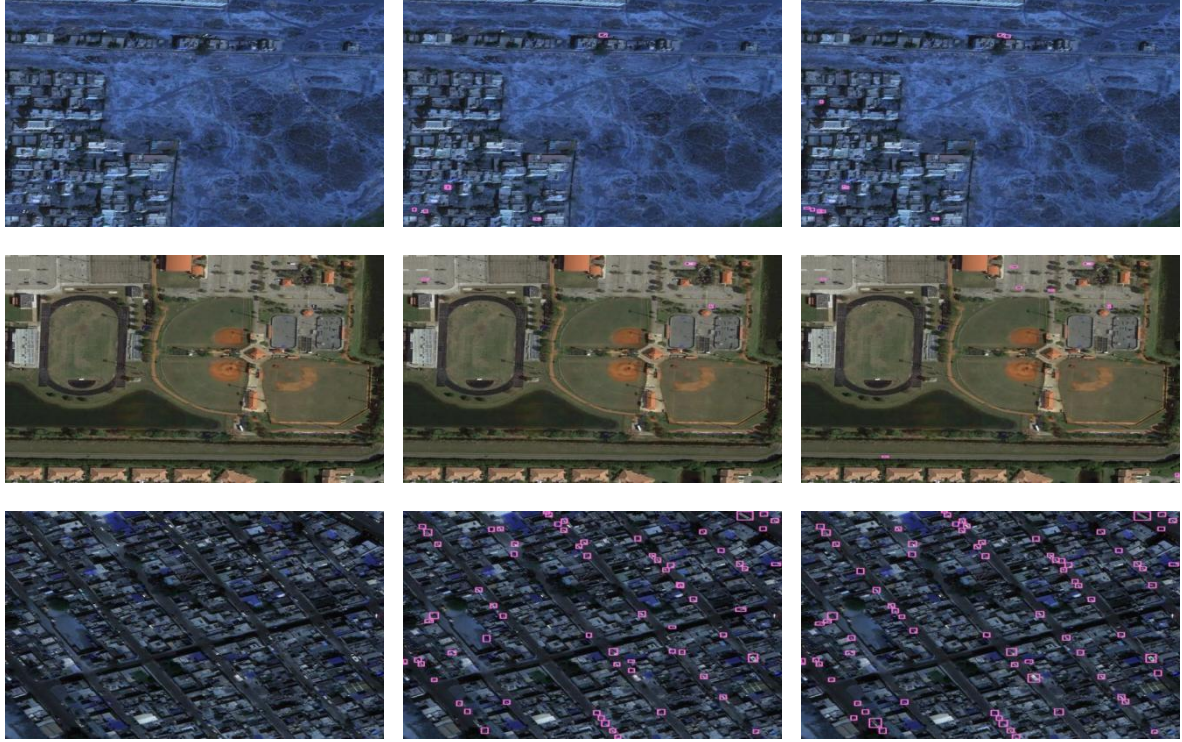


Figure 10: Comparison of the heat maps of YOLOv11 (center) and MDR-Net (right) on the AI-TOD dataset

capture directional features of targets; and the HCPF module, which utilizes a multi-scale fusion strategy to improve detection across varying target scales. Furthermore, the P2 detection layer leverages high-resolution feature maps to effectively mitigate the loss of spatial information of tiny targets that typically occurs in deeper network layers. Comprehensive experiments on the publicly available aerial datasets VisDrone and AI-TOD demonstrate that MDR-Net achieves significant improvements in small target detection accuracy compared to the baseline model YOLOv11 and other state-of-the-art methods, thus validating the effectiveness of its focused architecture and offering a practical high-performance solution for small target detection in aerial imagery.

CRedit authorship contribution statement

Jing Wang: Writing – review & editing, Writing – original draft, Visualization, Methodology. Jia Su: Resources, Project administration, Investigation, Formal analysis. Yanli Hou: Validation, Software, Investigation, Data curation. Boning Hu: Supervision, Funding acquisition, Conceptualization.

Data Availability Statement

The data supporting the findings of this study are publicly available in the official repositories of VisDrone and AI-TOD with the persistent identifiers <https://doi.org/10.21227/7pmg-wx16> (for VisDrone dataset) and <https://doi.org/10.1109/ICPR48806.2021.9413340> (for AI-TOD dataset).

Conflict of Interest Statement

The authors declare that no potential competing interests exist in this work.

Funding Information

This work was supported by the Hebei Province Major Science and Technology Support Program International Science and Technology Cooperation Project [grant numbers 25290102D]. The funding body had no role in the design of the study, collection, analysis, and interpretation of data, or in writing the manuscript. The authors would like to express their gratitude for the financial support provided.

References

- [1] J. Su, Y. Qin, Z. Jia, B. Liang, Mpe-yolo: enhanced small target detection in aerial imaging, *Scientific Reports* 14 (1) (2024) 17799.
- [2] T. Wang, Z. Ma, T. Yang, S. Zou, Petnet: A yolo-based prior enhanced transformer network for aerial image detection, *Neurocomputing* 547 (2023) 126384.
- [3] L. Zheng, Y. Yang, Q. Tian, Sift meets cnn: A decade survey of instance retrieval, *IEEE transactions on pattern analysis and machine intelligence* 40 (5) (2017) 1224–1244.
- [4] S. Suthaharan, Support vector machine, in: *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, Springer, 2016, pp. 207–235.
- [5] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2014, pp. 580–587.
- [6] R. Girshick, Fast r-cnn, in: *Proceedings of the IEEE international conference on computer vision*, IEEE, 2015, pp. 1440–1448.
- [7] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems* 39 (6) (2015) 1137–1149.
- [8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Vol. 9905, Springer, 2016, pp. 21–37.
- [9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2016, pp. 779–788.
- [10] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2017, pp. 7263–7271.
- [11] J. Redmon, A. Farhadi, Yolo3: An incremental improvement, 2018. [arXiv:1804.02767](https://arxiv.org/abs/1804.02767).
- [12] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolo4: Optimal speed and accuracy of object detection, 2020. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934).
- [13] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, Yolo7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, IEEE, 2023, pp. 7464–7475.
- [14] C.-Y. Wang, I.-H. Yeh, H.-Y. Mark Liao, Yolo9: Learning what you want to learn using programmable gradient information, in: *European conference on computer vision*, Vol. 15089, Springer, 2024, pp. 1–21.
- [15] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, Yolo10: Real-time end-to-end object detection, 2024. [arXiv:2405.14458](https://arxiv.org/abs/2405.14458).
- [16] R. Khanam, M. Hussain, Yolo11: An overview of the key architectural enhancements, 2024. [arXiv:2410.17725](https://arxiv.org/abs/2410.17725).
- [17] X. Wang, H. Chen, X. Chu, P. Wang, Aodet: Aerial object detection using transformers for foreground regions, *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024) 1–11.
- [18] D. Chen, D. Miao, X. Zhao, Hynet: Hybrid network transformer for multiple computer vision tasks, *IEEE Transactions on Industrial Informatics* 20 (6) (2024) 8773–8785.
- [19] X. Li, W. Diao, Y. Mao, P. Gao, X. Mao, X. Li, X. Sun, Ogm: Occlusion-guided multi-task network for object detection in uav images, *ISPRS Journal of Photogrammetry and Remote Sensing* 199 (2023) 242–257.
- [20] Y. Chen, Z. Ye, H. Sun, T. Gong, S. Xiong, X. Lu, Global-local fusion with semantic information-guidance for accurate small object detection in uav aerial images, *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025) 1–15.
- [21] Z. Li, S. Lian, D. Pan, Y. Wang, W. Liu, Ad-det: Boosting object detection in uav images with focused small objects and balanced tail classes, *Remote Sensing* 17 (9) (2025) 1556.
- [22] M. Muzammul, A. Algarni, Y. Y. Ghadi, M. Assam, Enhancing uav aerial image analysis: Integrating advanced sahi techniques with real-time detection models on the visdrone dataset, *IEEE Access* 12 (2024) 21621–21633.
- [23] Z. Zhou, Y. Zhu, Kldet: Detecting tiny objects in remote sensing images via kullback-leibler divergence, *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024) 1–16.
- [24] Y. Xiao, T. Xu, Y. Xin, J. Li, Fbrt-yolo: Faster and better for real-time aerial image detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, pp. 8673–8681.
- [25] X. Min, W. Zhou, R. Hu, Y. Wu, Y. Pang, J. Yi, Lwuavdet: A lightweight uav object detection network on edge devices, *IEEE Internet of Things Journal* 11 (13) (2024) 24013–24023.
- [26] X. Duan, Z. Li, J. Zhang, S. Li, J. Lei, L. Zhan, Weighted learnable re-cursive aggregation network for visible remote sensing image detection, *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025) 1–17.
- [27] Y. Xiao, T. Xu, X. Yu, Y. Fang, J. Li, A lightweight fusion strategy with enhanced inter-layer feature correlation for small object detection, *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024) 1–11.
- [28] N. Yin, C. Liu, R. Tian, X. Qian, Sdpdet: Learning scale-separated dynamic proposals for end-to-end drone-view detection, *IEEE Transactions on Multimedia* 26 (2024) 7812–7822.
- [29] G. Mao, H. Liang, Y. Yao, L. Wang, N. Zhang, Espnet: An efficient progressive spatial pyramid pooling network for real-time traffic object detection, *IEEE Transactions on Automation Science and Engineering* 22 (2025) 14048–14061.
- [30] L. Jiang, B. Yuan, J. Du, B. Chen, H. Xie, J. Tian, Z. Yuan, Mffsod-net: Multi-scale feature fusion small object detection network for uav aerial images, *IEEE Transactions on Instrumentation and Measurement* 73 (2024) 1–14.
- [31] Y. Ma, L. Chai, L. Jin, Scale decoupled pyramid for object detection in aerial images, *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023) 1–14.
- [32] J. Song, M. Zhou, J. Luo, H. Pu, Y. Feng, X. Wei, W. Jia, Boundary-aware feature fusion with dual-stream attention for remote sensing small object detection, *IEEE Transactions on Geoscience and Remote Sensing* 63 (2024) 1–13.
- [33] S. Huang, C. Lin, X. Jiang, Z. Qu, Brstd: Bio-inspired remote sensing tiny object detection, *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024) 1–15.
- [34] Q. Ge, W. Da, M. Wang, Marfpnet: Multi-attention and adaptive reparam-eterized feature pyramid network for small target detection on water sur-faces, *IEEE Transactions on Instrumentation and Measurement* 73 (2024) 1–17.
- [35] H. Lin, J. Wang, J. Li, Pfrnet: A small object detection method based on parallel feature extraction and attention mechanism, *IEEE Access* 13 (2025) 26727–26738.
- [36] X. Ding, R. Zhang, Q. Liu, Y. Yang, Real-time small object detection using adaptive weighted fusion of efficient positional features, *Pattern Recognition* 167 (2025) 111717.
- [37] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2017, pp. 1800–1807.
- [38] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, *Advances in neural information processing systems* 33 (2020) 21002–21012.
- [39] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, IEEE, 2019, pp. 658–666.
- [40] J. Ye, Z. Yuan, C. Qian, X. Li, Caa-yolo: Combined-attention-augmented yolo for infrared ocean ships detection, *Sensors* 22 (10) (2022) 3782.
- [41] H. Ren, M. Xia, L. Weng, K. Hu, H. Lin, Dual-attention-guided multi-scale feature aggregation network for remote sensing image change detection, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024) 4899–4916.
- [42] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2018, pp. 7132–7141.
- [43] Q. Gu, Z. Han, S. Kong, H. Huang, Y. Li, Q. Fan, R. Wu, Deyolo: Dual negative weighting label assignment and cross-layer decouple head for yolo in remote sensing images, *Expert Systems with Applications* 281 (2025) 127595.

- [44] P. Li, J. Chen, Q. Chen, L. Huang, Z. Jiang, W. Hua, Y. Li, Detection and picking point localization of grape bunches and stems based on oriented bounding box, *Computers and Electronics in Agriculture* 233 (2025) 110168.
- [45] X. Xia, C. Xu, B. Nan, Inception-v3 for flower classification, in: 2017 2nd international conference on image, vision and computing (ICIVC), IEEE, 2017, pp. 783–787.
- [46] Z. Li, S.-H. Wang, R.-R. Fan, G. Cao, Y.-D. Zhang, T. Guo, Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling, *International Journal of Imaging Systems and Technology* 29 (4) (2019) 577–583.
- [47] J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in: *International workshop on artificial neural networks*, Vol. 930, Springer, 2005, pp. 195–201.
- [48] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. 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*, IEEE, 2019, pp. 213–226.
- [49] J. Wang, W. Yang, H. Guo, R. Zhang, G.-S. Xia, Tiny object detection in aerial images, in: 2020 25th international conference on pattern recognition (ICPR), IEEE, 2021, pp. 3791–3798.
- [50] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., Yolov6: A single-stage object detection framework for industrial applications, 2022. [arXiv:2209.02976](https://arxiv.org/abs/2209.02976).
- [51] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, 2021. [arXiv:2107.08430](https://arxiv.org/abs/2107.08430).
- [52] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Dets beat yolos on real-time object detection, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, IEEE, 2024, pp. 16965–16974.
- [53] M. Kang, C.-M. Ting, F. F. Ting, R. C.-W. Phan, Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation, *Image and Vision Computing* 147 (2024) 105057.
- [54] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, Y. Wang, K. Han, Gold-yolo: Efficient object detector via gather-and-distribute mechanism, *Advances in Neural Information Processing Systems* 36 (2023) 51094–51112.