remote sensing

MDPI

Article

HD-BSNet: A Plug-and-Play Dual-Mechanism Synergistic Enhancement Framework for Small Object Detection

Jianwei Wen [1,2,3], Xiangyue Zheng [1,2,3], Nian Pan [1,2,3], Dan Jia [1,2,3], Haiying Wu [4], Tao Chen [5] and Jin Zhou [1,2,3,*]

[1] National Laboratory on Adaptive Optics, Chengdu 610209, China; wenjianwei241@mails.ucas.ac.cn (J.W.); zhengxiangyue23@mails.ucas.ac.cn (X.Z.); pannian@ioe.ac.cn (N.P.); jiadan0904@163.com (D.J.)
[2] University of Chinese Academy of Sciences, Beijing 101408, China
[3] Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[4] Huayin Ordance Test Center, Huayin 714200, China; xiner1983@163.com
[5] Satellite Launch Center, Xichang 615000, China; starryseact@163.com
* Correspondence: zhoujin@ioe.ac.cn

Highlights

What are the main findings?

- A plug-and-play framework HD-BSNet for small object detection is proposed, which integrates three core modules: high-frequency differential perception, background semantic modeling, and parallel multi-scale focusing, to address the inherent limitations of small object detection in a targeted manner.
- Experimental results on three mainstream datasets (AI-TOD, VisDrone, and DUT Anti-UAV) demonstrate that the framework can significantly improve the detection accuracy of small objects in complex background environments, while effectively reducing the false detection rate.

What are the implications of the main findings?

- It precisely addresses the core pain points of existing small object detection methods—including insufficient detection accuracy for targets smaller than 20 pixels, high false detection rates of GAN-based enhancement methods, and poor framework adaptability—thus providing reliable technical support for high-precision detection in fields such as remote sensing and low-altitude security.
- The plug-and-play design reduces the costs of technical implementation. Its dual-mechanism collaborative enhancement and multi-scale focusing strategy provides a new paradigm for research on feature extraction and background suppression in the field of small object detection, facilitating the transformation and deployment of related technologies in practical scenarios.

Abstract

In remote sensing and low-altitude unmanned aerial vehicle(UAV) detection scenarios, small target detection is extremely challenging due to the low pixel proportion, sparse features, and complex backgrounds of targets. The reliability of low-altitude security, in particular, is directly dependent on the accuracy of small target detection. However, current methods still face three major limitations: insufficient detection accuracy for targets smaller than 20 pixels; artifacts and false textures introduced by Generative Adversarial Network-based enhancement, which lead to increased false detection rates; and the reliance of existing approaches on specialized architectures, resulting in weak generalization capability and difficulty in adapting to multi-scenario deployment requirements. To address these issues, this paper proposes a plug-and-play dual-mechanism collaborative enhancement framework named HD-BSNet. Firstly, a High-Frequency Differential Perception mechanism

Check for updates

Academic Editor: Salah Bourennane

Received: 23 December 2025
Revised: 15 January 2026
Accepted: 23 January 2026
Published: 28 January 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) license.

is designed to enhance the detailed feature representation of small targets. Secondly, a Background Semantic Modeling mechanism is introduced to learn key features that distinguish targets from the background. Additionally, a Parallel Multi-Scale Focus Module is constructed to further reinforce target features. Extensive experiments on three small target datasets demonstrate that the proposed method effectively improves the accuracy and generalization ability of small target detection.

Keywords: background semantic modeling; high-frequency differential perception; plug-and-play; remote sensing; small object detection; UAV detection

1. Introduction

Investigations into the precise detection of small objects carry considerable importance for both the remote sensing imaging domain and the low-altitude UAV detection sector. Limited by the long imaging distance and sensor resolution, targets in remote sensing images are generally small in size. Meanwhile, low-altitude UAV early warning requires the effective identification of long-distance targets at 1 km or farther, where UAVs only occupy 15–40 pixels in images. Targets in both scenarios face difficulty in extracting effective features due to their low pixel proportion.

In addition, the background of remote sensing images usually contains complex terrain and texture information, while low-altitude UAV images face interference from complex ground objects such as mountains, buildings, and trees. These background characteristics can severely obscure the features of small objects and markedly impact detection accuracy, as shown in Figure 1. Thus, carrying out research on the precise detection of small objects for the aforementioned scenarios holds substantial significance.



Figure 1. Small objects in remote sensing and low-altitude UAV images typically exhibit the following characteristics: blurred features and complex backgrounds.

Faced with the distinct difficulties of detecting small objects in remote sensing and low-altitude UAV scenarios, a host of research efforts have delivered notable advancements to push forward this domain. Feature enhancement techniques are employed in studies [1–8] to alleviate the problem of insufficient representational capacity of small objects induced by downsampling, rendering the features of these small targets more susceptible to being represented and captured. Studies [9,10] focus on optimizing label assignment strategies to prevent performance degradation arising from improper label assignment. Moreover, to grasp the associations between small objects and the entities around them, the investigations documented in [11–16] seek to enhance model performance by incorporating multi-scale feature components.

However, mainstream existing methods still have non-negligible limitations. First, existing feature enhancement and multi-scale fusion methods still exhibit limited detection performance for small objects smaller than 20 pixels. Based on the experimental findings of this paper, the mean precision of existing methods for these targets typically falls below 25%. Second, some existing feature enhancement methods for small objects [1–3,6] often rely on image super-resolution technology and restore the structure of small objects via generative adversarial networks. Yet this approach tends to introduce false textures and artifacts, leading to an increase in the false detection rate of small objects [17] and failing to meet the requirements of accurate detection precision. Third, most existing methods are dedicated models that require backbone network reconstruction during integration, making it difficult to meet the practical deployment needs of multi-model and multi-scenario.

To address the aforementioned limitations, specifically including accuracy gaps for small objects, elevated false detection rates, and poor adaptability, this paper proposes HD-BSNet, a plug-and-play dual-mechanism collaborative enhancement framework. This framework supports seamless integration with various detection models without the need to reconstruct the backbone network, and can be adapted to remote sensing and low-altitude UAV application scenarios. Based on the strategies of high-frequency differential perception and background semantic modeling, HD-BSNet can not only directly and efficiently enhance the detailed features of small targets, but also effectively mine the discriminative salient features that distinguish target and background regions. Ultimately, this achieves a significant improvement in small target detection accuracy and effective reduction in the false detection rate under complex scenarios.

The key contributions of this article are outlined below.

(1) We introduce HD-BSNet, a plug-and-play dual-mechanism collaborative object detection framework. This framework can be directly embedded not only into both one-stage and two-stage detection models but also is suitable for remote sensing and low-altitude UAV scenarios, and it can effectively address the issues of accuracy gaps, elevated false detection rates, and poor framework adaptability for small objects.
(2) A dual-mechanism collaborative feature enhancement scheme for small objects is designed: it directly locates high-frequency information loss regions and captures features via differential operations to strengthen detailed representation; meanwhile, it integrates background semantic modeling to learn the dependencies between targets and backgrounds, thereby extracting salient features of small objects.
(3) A parallel multi-scale focusing module is proposed. Through multi-scale window partitioning and normalized focusing, the module enhances the features of small objects, indirectly suppresses interference from complex backgrounds, and improves detection performance.

The rest of this paper is structured in the subsequent manner. Section 2 offers a summary of relevant prior studies. Section 3 outlines the key attributes of the proposed framework. Section 4 delivers in-depth assessments and analyses of how different methods perform experimentally across various datasets. Section 5 consolidates the findings of this research and outlines potential directions for subsequent investigations.

2. Related Work

2.1. Generic Object Detection

Deep learning-based general object detection can be divided into two categories. Specifically, two-stage detectors [18–20] operate in a step-by-step manner: first, they identify candidate regions of interest (ROIs) where targets may exist in input images based on region proposals; subsequently, these candidate regions are fed into convolutional neural networks for target classification and bounding box regression. Early methods were computationally

expensive and relied on low-level image features, but subsequent improvements [21,22] introduced region proposal networks to achieve joint optimization, which significantly improved detection accuracy and speed. In addition, one-stage detectors [23–25] utilize an end-to-end architecture, which can directly predict target classes and bounding boxes based on deep features extracted from input images, eliminating the step of generating candidate regions. For this reason, such methods offer faster detection and localization speeds and are often applied in real-time object detection scenarios. Recently, anchor-free detection methods [26–28] have gained increasing popularity. Unlike traditional methods that rely on predefined anchor boxes and fixed matching criteria, these approaches can dynamically generate predictions according to the actual sizes and shapes of targets in images, thereby achieving more adaptive detection.

Nonetheless, general object detection models continue to encounter limitations when processing small targets. This issue stems primarily from two inherent traits of these targets: an extremely low pixel occupancy rate, and their susceptibility to disruption from complex backgrounds—both factors make it challenging for general detectors to precisely recognize and localize these targets. Thus, additional investigations are required to enhance the detection capability of models for such small targets.

2.2. Small Object Detection

The key challenge associated with small object detection stems from the severe lack of information within target regions. One mainstream solution is feature enhancement based on image reconstruction. Ren et al. [1] achieved the coordinated optimization of edge enhancement and spatial consistency by fusing multi-scale contextual semantic features with edge features for cross-domain interaction, combined with edge loss constraints. Zhu et al. [2] incorporated a multi-scale texture enhancement module into the generative adversarial network (GAN) framework, leveraging the attention mechanism to strengthen feature representation and realizing multi-scale enhancement of edge features. Zhang et al. [3], meanwhile, enhanced the expression of edge features and cross-scale information fusion through edge prior enhancement and a dynamic fusion attention mechanism. Furthermore, Yi et al. [4] proposed a novel attention module named CloAttention, which incorporates both global and local attention mechanisms into the backbone network. This design enhances the model's feature extraction capability and enables it to capture detailed information across global and local scales simultaneously. In addition, the EFEM feature enhancement module designed by Zheng et al. [5] achieves efficient fusion of local and global features by virtue of a multi-branch convolution structure and multi-scale receptive fields. Both of these methods take the exploration of local detailed features as their core entry point to achieve the goal of feature enhancement.

Furthermore, multi-scale feature merging is designed to separate and combine hierarchical features within the network. By fusing feature maps from different depths, it can alleviate the representation bottleneck of the model in cross-scale object detection. Lin et al. [11] introduced the Feature Pyramid Network (FPN). This framework sequentially combines high-resolution shallow features with deep features through a top-down pathway, enabling effective multi-scale feature merging. Subsequently, numerous improved methods based on Feature Pyramid Network (FPN) have been further proposed. The Path Aggregation Network (PANet) put forward by Liu et al. [12] effectively shortens the information transmission path between low-level and high-level features through bottom-up path enhancement. Ge et al. [13] addressed the feature misalignment issue occurring in direct upsampling-based feature aggregation by introducing a cross-attention mechanism. In addition, the MRFAFEM module proposed by Liu et al. [14] can dynamically adjust features with different receptive fields to enhance the detection capability for tiny

objects. Moreover, Qiao et al. [15] substantially advanced the detection performance for small objects through the incorporation of a recursive feature pyramid and a switchable atrous convolution. Zhou et al. [29] proposed a joint framework that represents bounding boxes via Gaussian distribution and implements adaptive online sample mining, aiming to localize targets more accurately and optimize the training process.

However, current methods based on feature enhancement tend to introduce additional noise [17] and incur substantial computational overhead [30] when tackling the problem of information loss. In addition, the aforementioned multi-scale learning methods overlook the importance of high-frequency information such as edge details. Specifically, when multi-scale features are directly fused, the fine-structured components of small targets are prone to being overshadowed by coarse-grained deep-layer semantics, resulting in boundary blurring and feature degradation, which in turn exert a negative impact on detection accuracy.

3. Method

3.1. Overview

To mitigate the drawbacks of limited detection precision, elevated false detection rates, and weak generalization performance in small object detection for remote sensing and low-altitude UAV scenarios, this paper innovatively proposes HD-BSNet, a dual-mechanism collaborative enhancement framework. Its core lies in a collaborative working mechanism of high-frequency differential perception and background semantic modeling, which can effectively enhance the high-frequency features of targets and extract highly discriminative feature representations. The structure of this framework is depicted in Figure 2.

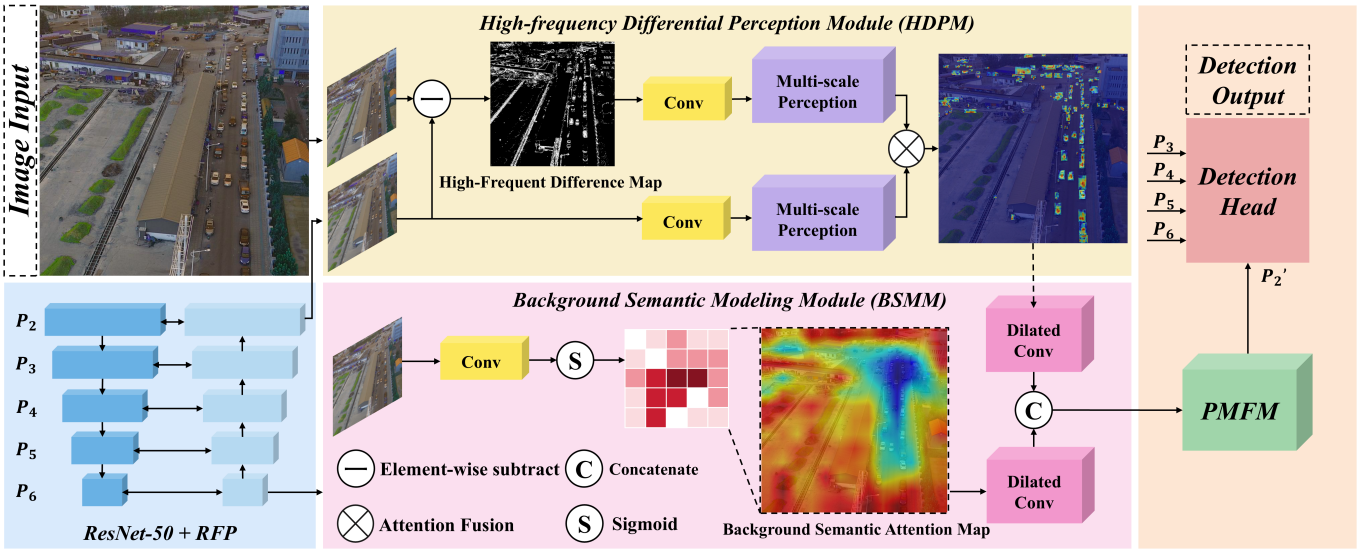


Figure 2. Structure of the proposed HD-BSNet framework. HDPM captures high-frequency detailed features via the differential operation between the original image and shallow feature maps, thereby enhancing the foreground representation capability. BSMM learns background semantic representations based on deep features, and improves the discriminative ability between small objects and complex backgrounds through separate modeling and feature fusion of foreground and background. Additionally, PMFM further strengthens the small object representation in shallow features via multi-scale window division and normalized focusing. P_2 to P_6 are multi-scale feature maps output by the backbone network. Specifically, P_2 is a shallow feature map; after being enhanced via HDPM/BSMM/PMFM, it is finally transformed into P_2'. Subsequently, P_3 to P_6 and the enhanced P_2' are jointly fed into the detection head to perform classification and regression tasks.

First, the input image is processed by the ResNet-50 [31] backbone network and the RFP [15] neck network to produce feature maps of varying scales, denoted as P_i ($i = 2, 3, 4, 5, 6$). Subsequently, the shallow P_2 feature map which contains abundant spatial feature information, is enhanced primarily through three proposed modules: the High-Frequency Differential Perception Module (HDPM), the Background Semantic Modeling Module (BSMM), and the Parallel Multi-scale Focusing Module (PMFM).

3.2. High-Frequency Differential Perception Module

(1) High-Frequency Feature Map Extraction: During the downsampling process of the backbone network, target information loss is inevitable, and such loss is particularly severe for small objects. In this case, the information of small objects is prone to being completely erased, making it impossible to perform predictions relying on these low-representational-capability features. We aim to re-inject high-frequency information related to small objects, which is rich in edge and texture details, into the downsampled feature maps.

It should be clarified that the essence of image high-frequency information lies in regions with rapidly changing grayscale values [7], which mathematically corresponds to the first-order gradient of the image signal. During the convolutional downsampling process of the backbone network, although the shallow feature map P_2 retains the richest spatial details, the smoothing effect of convolutional operations leads to the loss of some high-frequency components. In contrast, the element-wise difference between the downsampled original image and P_2 essentially approximates the solution to the first-order gradient, which can directly restore the high-frequency details smoothed by convolution.

The core reason for selecting P_2 as the reference for the difference operation is that it preserves more abundant spatial location information compared with deep feature maps, which ensures that the extracted high-frequency information is highly correlated with small targets. Compared with GAN-based methods, the difference operation extracts high-frequency information directly from the original image, thereby avoiding artifacts and false texture noise that are prone to be introduced during the generation process. In contrast to traditional filtering methods, the preliminary semantic constraints embedded in P_2 can guide the difference operation to accurately focus on target-related high-frequency details, preventing the blind enhancement of background noise. Given the input image $I \in \mathbb{R}^{B\times 3\times H\times W}$ and the P_2 feature map $F_{p2} \in \mathbb{R}^{B\times C\times H_{p2}\times W_{p2}}$, the high-frequency feature F_{HF} is obtained via element-wise differentiation:

$$F_{HF} = \mathrm{Conv}_{\mathrm{down}}(\mathrm{Interpolate}(I)) - F_{p2} \tag{1}$$

where $\mathrm{Interpolate}(\cdot)$ denotes the bilinear interpolation downsampling operation, with an output size of (H_{p2}, W_{p2}). The kernel size of $\mathrm{Conv}_{\mathrm{down}}$ is $C \times 3 \times 1 \times 1$; it maps 3 channels to C channels via 1 × 1 convolution, ensuring the normal operation of the difference operation.

(2) High-Frequency Feature Perception Guidance: After obtaining the high-frequency feature map via differentiation, a key challenge lies in effectively leveraging the prior information contained within it to enhance the ability to detect small objects. As illustrated in Figure 2, the high-frequency feature map needs to be combined with the P_2 feature map subsequently to reinforce the feature information of small objects on the P_2 feature map. However, prior to the combination, it is also necessary to undergo processing of feature alignment and multi-scale perception. Since the P_2 feature and the high-frequency feature originate from different processing paths, they exhibit significant differences in three aspects: channel dimension, value distribution, and semantic space. The P_2 feature undergoes multi-layer processing by deep convolutional networks, resulting in a relatively stable feature distribution and encoding high-level semantic information. In contrast, the high-frequency feature is directly obtained through the differentiation between the original

image and P_2, with an unnormalized value distribution and mainly encoding geometric detail information. This heterogeneity makes it difficult to directly fuse the two types of features.

To address this issue, our method achieves alignment at three levels by designing a "1 × 1 convolution + BatchNorm + ReLU" combination: (1) Channel alignment: projecting features into a unified channel space via 1 × 1 convolution; (2) Distribution alignment: normalizing features from different sources to a similar value scale through BatchNorm, ensuring numerical stability for subsequent fusion operations; (3) Semantic alignment: the 1 × 1 convolution essentially learns a mapping from the geometric feature space to the semantic feature space, enabling geometric information such as edges and textures in the high-frequency feature to be converted into semantically compatible representations with the P_2 feature.

Feature alignment not only ensures the stability of numerical calculations but also, more importantly, establishes semantic correlations between the two types of heterogeneous features, laying a foundation for subsequent multi-scale perception and attention fusion. The feature alignment of the P_2 feature and the high-frequency feature can be expressed as follows:

$$F_{p2}^{\text{align}} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(F_{p2}))) \tag{2}$$

$$F_{HF}^{\text{align}} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(F_{HF}))) \tag{3}$$

where ReLU denotes the Rectified Linear Unit; BN denotes Batch Normalization; and $\text{Conv}_{1\times1}$ denotes the convolution operation with a kernel size of $C \times C \times 1 \times 1$.

After completing the feature alignment operation for both, this paper identifies a key issue of non-selectivity in high-frequency feature maps. Specifically, performing differential operations captures detailed information in both target regions and background regions simultaneously, yet lacks the ability to autonomously distinguish the importance of such detailed information. Therefore, it is necessary to further address two challenges: the determination of spatial positions for target enhancement, and the quantitative definition of the scope of action of information at each position.

Although the feature alignment operation has initially established the mapping relationship between feature spaces from different sources, its feature receptive field is still limited to the local range of 1 × 1. To achieve effective distinction between target edges and background noise, high-frequency information needs to possess continuous contextual correlation information. To address the non-selectivity challenge of high-frequency feature maps, this paper introduces a multi-scale perception module to realize selective activation of features, and its specific architecture is illustrated in Figure 3.

The multi-scale perception module expands the feature receptive field to 13 × 13 through progressive convolutions of 3 × 3 → 5 × 5 → 7 × 7, and establishes three hierarchical discriminative criteria: the local layer (3 × 3) detects the presence of edges, the structural layer (5 × 5) judges the integrity of contours, and the contextual layer (7 × 7) confirms the independence of objects. Based on this hierarchical contextual understanding, the module can achieve precise distinction between target and background regions. Meanwhile, to maintain dimensional consistency and improve the quality of modulated features, we also perform multi-scale perception enhancement on the P_2 branch. The process of multi-scale perception can be mathematically expressed as follows:

$$\text{Pzconv}(X) = (\text{Conv}_{7\times7}^{\text{DW}}(\beta(\cdots\beta(\text{Conv}_{3\times3}^{\text{DW}}(X))\cdots))) + X \tag{4}$$

$$F_{p2}^{\text{multi}} = \text{Pzconv}(F_{p2}^{\text{align}}) \tag{5}$$

$$F_{HF}^{\text{multi}} = \text{Pzconv}(F_{HF}^{\text{align}}) \tag{6}$$

where $\text{Conv}_{k\times k}^{\text{DW}}$ denotes the depthwise separable convolution with a kernel size of k, and β denotes the pointwise convolution. The P_2 branch and the high-frequency feature branch extract semantic and geometric information from the multi-scale perception module respectively, achieving fine-grained correspondence and selective enhancement through the attention mechanism. The mathematical formula of this process is expressed as follows:

$$A = \sigma(\text{Conv}_{1\times 1}(F_{HF}^{\text{multi}})) \tag{7}$$

$$F_{\text{enhanced}} = F_{HF}^{\text{multi}} \odot A + F_{p2}^{\text{multi}} \tag{8}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function; $\odot$ denotes element-wise multiplication; and $A \in [0,1]^{B\times 1\times H\times W}$ denotes the attention weight map.

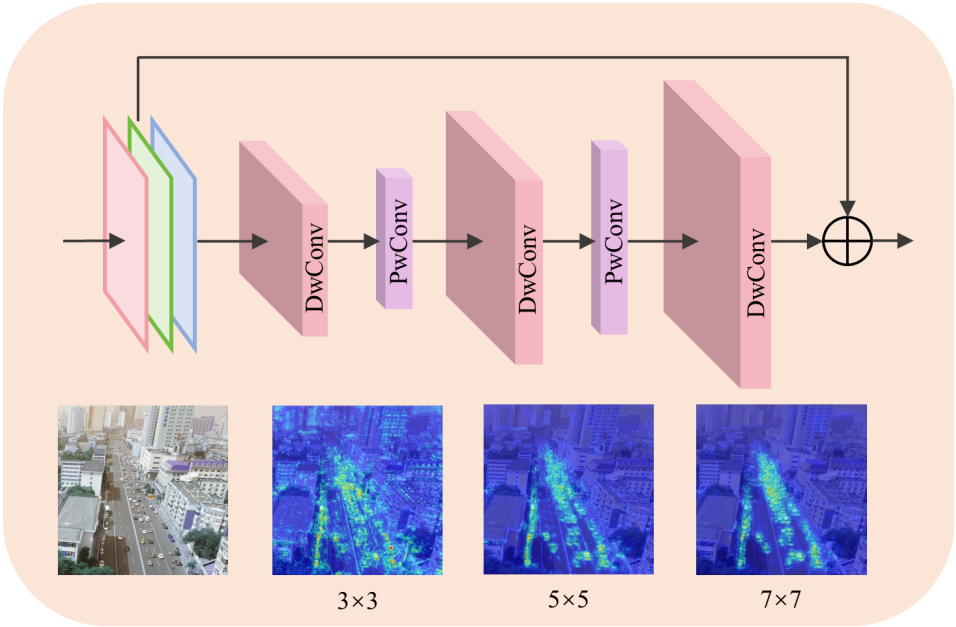


Figure 3. This figure details the Multi-scale Perception Module, a core sub-module of the HDPM in Figure 2. The upper part illustrates the network structure of this module, and the lower part shows the attention heatmap visualization under different convolution scales, intuitively reflecting the module's ability to gradually focus on target regions. Specifically, the feature map with a 3 × 3 convolution kernel can capture detailed information of both targets and backgrounds; the feature map with a 5 × 5 convolution kernel significantly reduces attention to background regions due to the expanded receptive field; and the feature map with a 7 × 7 convolution kernel further focuses its attention on target regions.

3.3. Background Semantic Modeling Module

The background semantic modeling module (BSMM) is designed to acquire clues for distinguishing between targets and backgrounds, thereby helping the model obtain more discriminative features, as illustrated in Figure 2. Although the top-level feature map P_6 lacks clear spatial texture information, it possesses the largest receptive field and the richest global semantic information, enabling it to effectively capture the overall layout of the scene and the semantic features of background regions. Furthermore, for small object detection tasks, after multiple downsampling operations, small objects exhibit weak feature responses at the P_6 scale, which causes the P_6 feature map to naturally focus on the semantic representation of background regions. Leveraging this characteristic, the model can learn accurate and pure background semantic representations from P_6 features through a specially designed background extraction network.

Specifically, we first process the P_6 feature via convolutional layers and BatchNorm layers to extract background-related semantic information, then generate a background semantic attention map with values ranging from $[0, 1]$ through the Sigmoid activation function. This process is mathematically expressed as follows:

$$M_{bg} = \sigma(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F_{p6}))) \tag{9}$$

where $M_{bg} \in [0, 1]$ denotes the background semantic attention map; F_{p6} denotes the top-level P_6 feature map. After obtaining the background semantic map and the high-frequency enhanced P_2 feature, we do not directly perform foreground-background separation and fusion. Instead, we introduce a multi-scale semantic enhancement module based on dilated convolution, as illustrated in Figure 4. This design is motivated by two considerations: first, simple weighted separation only achieves spatial division of features and fails to fully capture the respective multi-scale semantic information of foreground targets and background regions; second, in small object detection tasks, there are significant scale differences between foreground and background, where foreground targets often occupy only a few pixels while background regions may span the entire image, making it difficult for single-scale feature representations to accommodate both simultaneously.

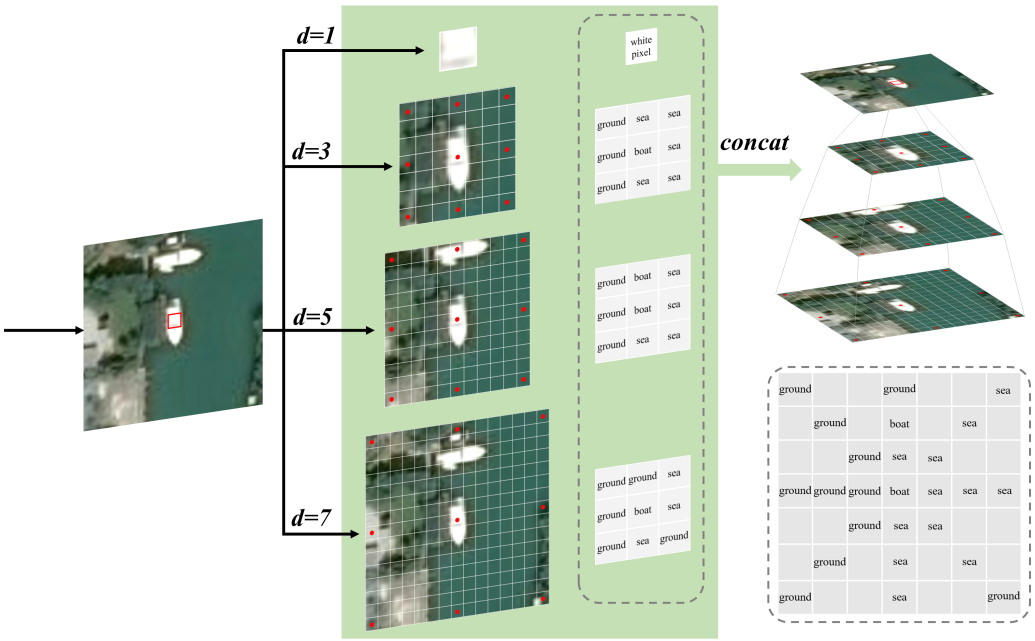


Figure 4. This figure details the structure of the Dilated Convolution Module, a core sub-module of the BSMM in Figure 2. "*d*" denotes the dilation rate.

Specifically, through dilated convolutions with dilation rates of 1, 3, 5, and 7, the receptive field is expanded while maintaining feature resolution. This enables the foreground branch to understand target features at multiple levels, from local precise localization to global contextual correlation, and the background branch to model scene information at multiple scales, from texture details to semantic categories. After the multi-scale features of the two branches are concatenated in the channel dimension, they are fused and dimensionally reduced via the prediction layer to generate highly discriminative and semantically rich feature representations, which effectively improves the detection accuracy of small objects and reduces the false detection rate. The mathematical process is expressed as follows:

$$T_k^r(X) = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{k\times k}^{d=r}(X))) \tag{10}$$

$$\mathrm{ASPP}(X) = \mathrm{Concat}[T_1^1(X), T_3^3(X), T_3^5(X), T_3^7(X)] \tag{11}$$

where T_k^r denotes the $k \times k$ dilated convolution with a dilation rate of r; $\mathrm{Concat}(\cdot)$ denotes the tensor concatenation operation.

3.4. Parallel Multi-Scale Focusing Module

Feature maps post feature fusion are able to acquire varied features across numerous channels and produce reactions in distinct spatial areas. Nevertheless, certain channels may produce extensive high responses over large spatial regions, and complex backgrounds with rich textures are also prone to triggering diffused activations. These excessively dispersed activations may lead to feature ambiguity in multi-target contexts, a consequence that is harmful to the precise extraction of small object features. To address this issue, this module is designed to strengthen the feature representation of small objects by focusing on local high-response regions and balancing response intensities across multiple regions, while indirectly mitigating the interference of background noise.

Inspired by the multi-scale window partitioning idea of the Local Attention Pyramid Module (LAPM) [32], we propose a Parallel Multi-scale Aggregation Module (PMFM), whose structure is illustrated in Figure 5. Unlike the serial recursive structure of LAPM, PMFM adopts a brand-new parallel architecture. Its core process is as follows: First, it parallelly divides the feature map into image patches of different granularities using multi-scale windows; subsequently, it independently performs instance normalization on the image patches of each scale to balance response differences between regions and highlight local signals; then, it generates attention weight maps for each scale via the Sigmoid function to enhance target features and weaken the background; finally, it introduces a set of learnable weights to adaptively fuse the attention maps generated at each scale into a complete channel attention map.

This parallelized design offers the following advantages: First, processing at all levels is performed independently based on the original input features, avoiding the issue of error accumulation across layers caused by recursive computation in LAPM. Second, each level can obtain complete original feature information, ensuring that even in deep-level processing, subtle original details can be retained and overcoming the defect of potential loss of low-level information in serial structures. Third, by fusing multi-scale attention via a set of learnable weights, the module can adaptively learn and balance the importance of different scale levels, achieving more flexible multi-scale feature focusing. Overall, PMFM adopts a lightweight design and maintains favorable plug-and-play properties, ultimately realizing more accurate and robust multi-scale enhancement of small target features.

Assuming the feature map F has a size of $H \times W$ ($H = W$), the module processes the feature through multiple levels, denoted as $p = n, n-1, \ldots, 1, 0$. The patch size in the input feature map F_p at each level is $h \times w$, which can be mathematically expressed as follows:

$$h = \frac{H}{2p} \tag{12}$$

$$w = \frac{W}{2p} \tag{13}$$

The attention map generated via normalization and the Sigmoid activation function at each level can be denoted as A_p. Subsequently, at each level, the attention map is used to enhance the local high-response regions of the original feature F_a while suppressing background noise. These are calculated by the following formula:

$$A_p = \sigma(\mathrm{PatchIn}(F_p, p)) \tag{14}$$

$$F_p^{\text{enhanced}} = \theta \cdot F_a + (1 - \theta) \cdot A_p \odot F_a \tag{15}$$

where $A_p \in [0, 1]$ represents the retention (or suppression) weight for each position, PatchIn$(\cdot)$ denotes instance normalization applied to image patches, and $\theta \in [0, 1]$ is a balance factor that ensures enhancement does not completely overwrite the original feature, set to 0.5 herein. Subsequently, the results of each level are fused using learnable weights to obtain the final output of the module, denoted as F_{output}:

$$\alpha = \text{Softmax}(w_p) \tag{16}$$

$$F_{\text{output}} = \sum_{p=0}^{p} \alpha_p \cdot F_p^{\text{enhanced}} \tag{17}$$

where $w_p \in \mathbb{R}^{p+1}$ denotes the learnable hierarchical weight parameters, $\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_p]$ represents the Softmax-normalized weights, which satisfy the constraint $\sum_{p=0}^{p} \alpha_p = 1$.

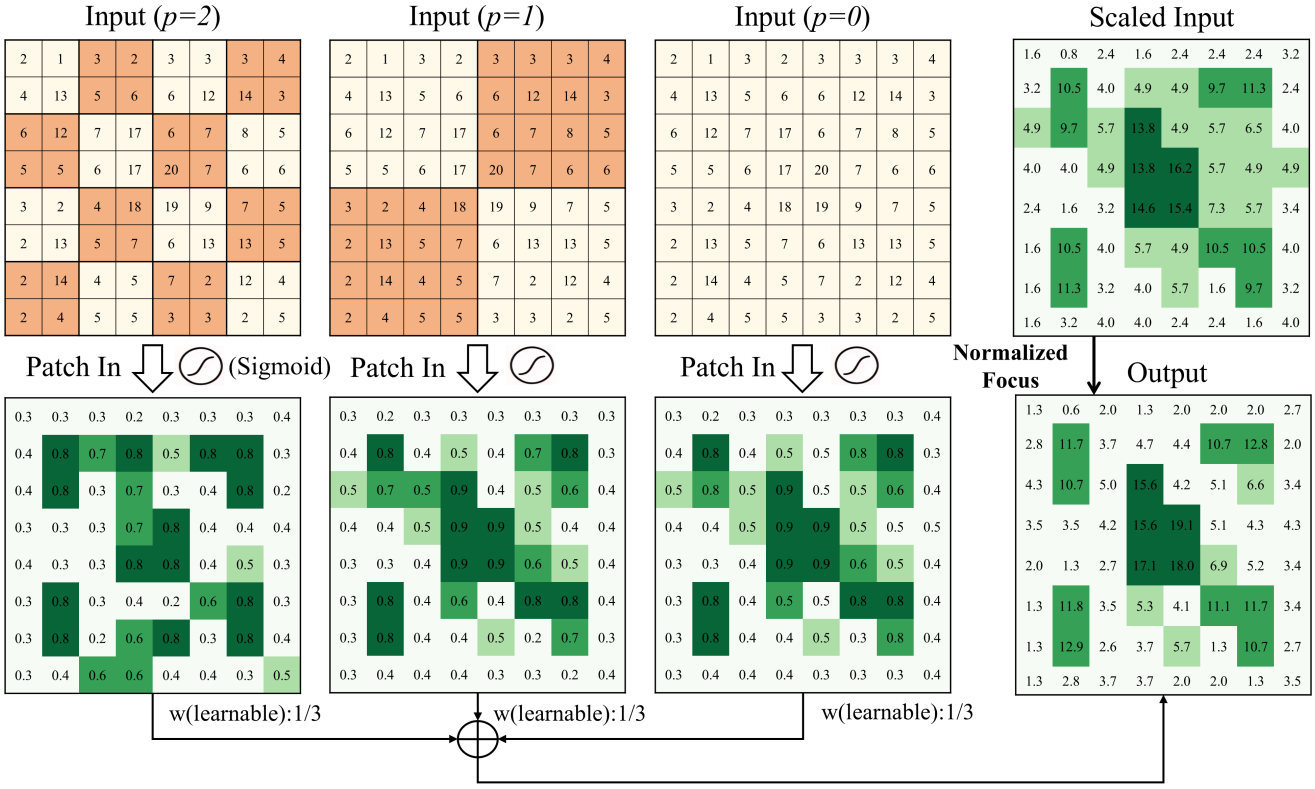


Figure 5. Parallel Multi-Scale Focusing Module. We scaled the "Input" to the total value of the "Result", enabling a clear comparison between the two.

4. Experiments

4.1. Experimental Setup

(1) Datasets: To comprehensively evaluate the performance of the framework, three datasets, namely AI-TOD [33], VisDrone [34], and DUT Anti-UAV [35], are selected for the experiments. These datasets not only contain a large number of small target samples smaller than 32 pixels but also accurately reflect the complexity of remote sensing and low-altitude UAV scenarios.

AI-TOD is a dataset dedicated to small object detection tasks in optical remote sensing images. This dataset contains 28,036 images with a resolution of 800 × 800 pixels, covering 8 categories. Its most prominent feature is the extremely small target size, with an average geometric size of just 12.8 pixels—far smaller than that of mainstream datasets

like DOTA [36] (55.2 pixels), DIOR [37] (74.2 pixels), and MS COCO [38] (99.5 pixels). The dataset is divided into a training set of 11,214 images, a validation set of 2804 images, and a test set of 14,018 images.

VisDrone2019 is an unmanned aerial vehicle (UAV) aerial image dataset designed for small object detection tasks. This dataset contains 8599 images, which are divided into a training set of 6471 images, a validation set of 548 images, and a test set of 1580 images. Its annotations cover 11 categories of common ground targets such as pedestrians, vehicles, and bicycles, while the background environments cover diverse weather conditions including sunny, cloudy and foggy days and various scenarios including urban streets, squares and suburban areas. Beyond containing a large volume of small objects with pixel dimensions under 32 × 32, this dataset also includes a substantial number of such targets, acting as a key benchmark for assessing the generalization performance of small object detection algorithms.

DUT Anti-UAV is a visible light anti-UAV dataset released by the team from Dalian University of Technology. It comprises 10,000 images in total, split into a training subset of 5200 images, a validation subset of 2600 images, and a test subset of 2200 images. UAV targets in this dataset are dominated by small targets with an average target pixel size of 34 × 23, accounting for 0.1% of the image size, while the dataset also features rich background types and diverse interference factors.

(2) Implementation Details: The code we used is built upon the open-source PyTorch (Version 1.12.0) framework [39] and MMDetection toolkit [40], with experiments run on an A100 GPU equipped with 24G memory. Based on the differences in dataset scale and model convergence characteristics, we set 36 training epochs for AI-TOD and 12 training epochs for VisDrone. We selected the mainstream and moderate-performance models [15,20,21] as baseline models, and the pre-trained ResNet-50 [31] was adopted as the backbone network. The batch size was set to 2, and the optimization process used the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.005, a momentum of 0.9, and a weight decay factor of 0.0001; all remaining parameters followed MMDetection's default settings. During the inference stage, a confidence threshold of 0.05 was configured to exclude background regions. Non-Maximum Suppression (NMS) with an IoU threshold of 0.5 was applied, and the top 3000 detection results were kept. For DUT Anti-UAV, the model was trained for 200 epochs, with YOLOv11 selected as the representative baseline model, and all models on this dataset were trained from scratch. The batch size was set to 16, and the patience value was 60.

(3) Evaluation Metrics: Our experiments follow the evaluation metrics in MS COCO [38], using Precision (P), Recall (R) and Average Precision (AP) to quantitatively evaluate the detection performance on the datasets. AP is defined as follows:

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$AP = \int_0^1 P(r)\, dr \tag{20}$$

Here, TP represents the count of correctly matched samples across varying IoU thresholds; FP stands for the count of incorrectly matched samples; and FN refers to the count of missed samples. For each category, AP refers to Average Precision, while AP50 denotes the AP calculated at an IoU threshold of 0.5.

On the AI-TOD dataset, small-sized objects are further categorized into "very tiny", "tiny", "small", and "medium", with respective size ranges of $[2^2, 8^2]$, $[8^2, 16^2]$, $[16^2, 32^2]$, and

$[32^2, 1 \times 10^5]$, respectively. These metrics are detailed in [33]. Therefore, on this dataset, the model performance will be further evaluated in accordance with this subdivision criteria.

4.2. Comparative Experiments

To ensure the targeted and fair evaluation of HD-BSNet, we selected comparative algorithms tailored to the core challenges of AI-TOD, VisDrone, and DUT Anti-UAV, respectively, based on their distinct scenario characteristics. These algorithms include state-of-the-art (SOTA) methods published in recent years, which ensures that the experiments can comprehensively and objectively validate the effectiveness of the proposed framework.

(1) Experiment on AI-TOD: Table 1 displays the test set experimental results of our proposed method alongside other SOTA detectors, covering methods [41–44] that have achieved SOTA performance in remote sensing small object detection. As shown in Table 1, our method highly competitive results across all metrics, with the AP, AP50, and AP75 metrics reach 29.8%, 61.7%, and 26.9% respectively. All these values outperform other SOTA small object detection algorithms, demonstrating the method's advantages in detection accuracy. Importantly, while our method does not attain the best performance for very tiny target detection, it achieves a 5.2% improvement in the APt metric over the SOTA performance for tiny targets smaller than 20 pixels. This demonstrates that our method performs exceptionally well when tackling the challenges of small object detection in remote sensing contexts.

Furthermore, visual comparison results of different approaches are presented in Figures 6 and 7. In Figure 6, we visualized the global attention maps of the DetectoRS model with and without the integration of HD-BSNet to enable in-depth analysis and understanding of the advantages of the proposed method. Specifically, relative to the baseline approach, our method significantly reduces its attention to background regions and focuses on small objects, indicating that it can effectively avoid interference caused by complex backgrounds. Further observation of Figure 6 clearly shows that HD-BSNet exhibits broad adaptability in small object detection. It can accurately identify and locate small objects that are often missed or difficult to detect by other methods under various scenarios such as sparse blurring, dense distribution, and dark occlusion. In addition, the regions marked by orange dashed lines also verify the superior performance of HD-BSNet in detecting occluded targets. In these regions, targets are obscured by clouds and fog, and the dark environment further increases the detection difficulty. However, through high-frequency information enhancement and background information modeling, our method successfully captures these targets under dark and occluded conditions. Such results not only confirm the validity of our method in carrying out small object detection tasks but also illustrate its capacity to address complex situations and challenging targets.

Table 1. Evaluation results of different models on the AI-TOD dataset. Red and blue indicate the best and second-best results, respectively.

Method	Venue	Backbone	AP (%)	AP50 (%)	AP75 (%)	AP_{vt} (%)	AP_{t} (%)	AP_{s} (%)	AP_{m} (%)
SSD-512 [23]	ECCV2016	VGG-16	6.5	21.0	2.4	1.0	5.3	11.9	21.2
Faster R-CNN [21]	TPAMI2017	ResNet-50	9.7	22.1	6.8	0.0	5.2	21.7	32.5
RetinaNet [24]	ICCV2017	ResNet-50	2.4	8.4	0.7	0.8	2.7	4.0	8.8
Mask R-CNN [19]	ICCV2017	ResNet-50	10.1	22.7	7.6	0.0	5.8	22.1	32.6
PANet [12]	CVPR2018	ResNet-50	9.7	22.2	6.8	0.0	5.4	21.6	32.5
Cascade R-CNN [20]	CVPR2018	ResNet-50	11.8	25.6	9.5	0.1	7.9	23.0	34.8
YOLOv3 [45]	arXiv2018	DarkNet-53	12.3	25.3	5.6	3.2	12.0	18.5	24.2
TridentNet [46]	ICCV2019	ResNet-50	4.9	13.1	2.7	0.0	2.8	10.8	19.2
CenterNet [26]	ICCV2019	ResNet-18	8.2	24.3	3.4	1.3	6.9	12.9	22.3

Table 1. *Cont.*

Method	Venue	Backbone	AP (%)	AP50 (%)	AP75 (%)	AP_{vt} (%)	AP_t (%)	AP_s (%)	AP_m (%)
AutoAssign [47]	arXiv2020	ResNet-50	13.4	36.2	6.9	3.5	13.5	17.0	24.9
ATSS [48]	CVPR2020	ResNet-50	8.9	22.5	5.3	2.1	10.1	11.4	15.0
TOOD [49]	ICCV2021	ResNet-50	14.9	34.7	10.6	3.3	12.8	21.7	33.1
M-CenterNet [33]	ICPR2021	DLA-34	14.5	40.7	6.4	6.1	15.0	19.4	20.4
NWD [50]	arXiv2021	ResNet-50	20.8	49.3	14.3	6.4	19.7	29.6	38.3
RFLA [41]	ECCV2022	ResNet-50	24.8	55.2	18.5	9.3	24.8	30.3	38.2
NWD-RKA [10]	ISPRS2022	ResNet-50	23.4	53.5	16.8	8.7	23.8	28.5	36.0
HANet [51]	TCSVT2023	ResNet-50	22.1	53.7	14.4	10.9	22.2	27.3	36.8
DINO [52]	ICLR2023	ResNet-50	23.2	56.6	15.4	9.9	23.1	29.3	37.6
Dotd [9]	TPAMI2024	ResNet-50	16.1	39.2	10.6	8.3	17.6	18.1	22.1
MENet [53]	TGRS2024	Swin-T	23.2	56.2	15.0	9.7	23.9	25.3	34.4
ORFENet [14]	TGRS2024	ResNet-50	16.6	38.3	11.6	5.9	17.6	29.9	23.0
DNTR [43]	TGRS2024	ResNet-50	24.9	54.3	18.4	11.5	25.0	29.0	35.3
HS-FPN [42]	AAAI2025	ResNet-50	25.1	55.7	19.1	12.1	25.3	29.9	36.9
CFENet [44]	TGRS2025	CSPDarkNet-s	26.6	59.2	19.6	9.0	27.5	32.3	38.5
Faster R-CNN w/HD-BSNet		ResNet-50	26.8	58.6	23.3	11.0	30.5	32.9	36.9
Cascade R-CNN w/HD-BSNet		ResNet-50	28.2	59.1	25.2	11.9	31.5	34.5	40.3
DetectoRS w/HD-BSNet		ResNet-50	29.8	61.7	26.9	11.5	32.7	36.9	42.7

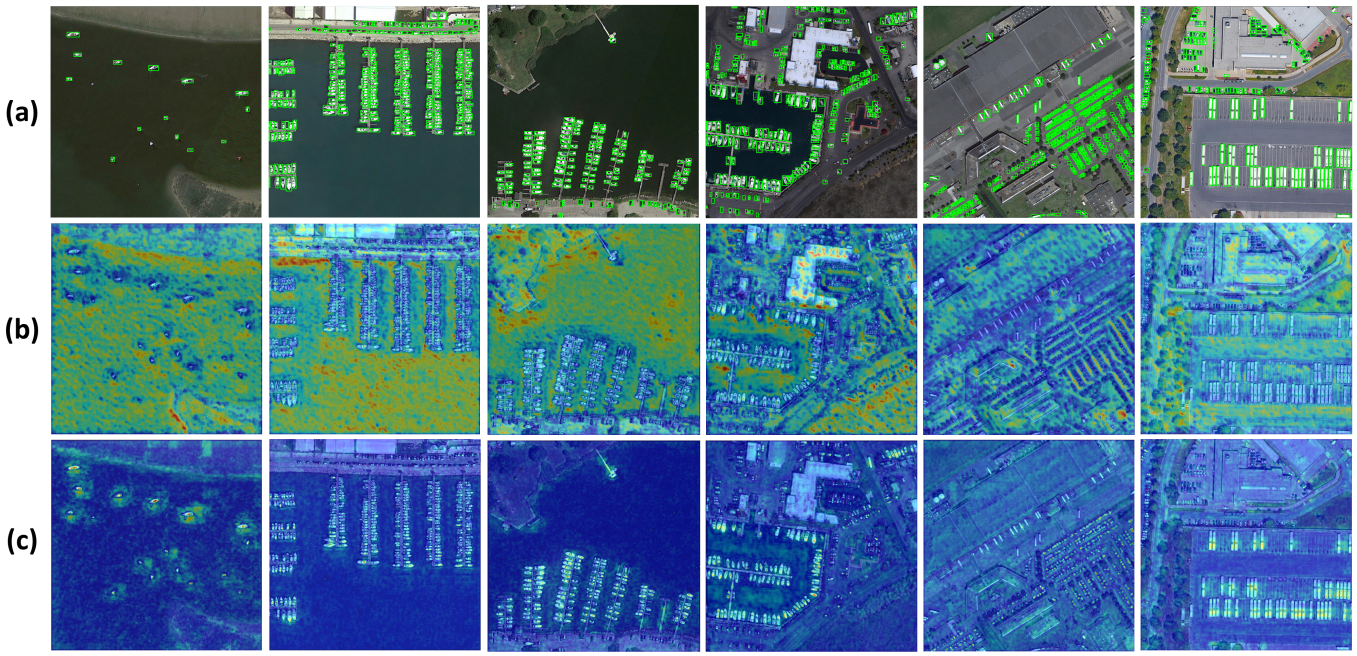


Figure 6. Qualitative comparison map of global attention distribution on the AI-TOD dataset. (**a**) Ground Truth (GT). (**b**) Visualization on DetectoRS. (**c**) Visualization on DetectoRS w/HD-BSNet. Visualization results indicate that the significant improvement in HD-BSNet's detection accuracy is attributed to its ability to guide the model's attention to focus more on small targets while greatly reducing the attention allocated to background regions.

(2) hlExperiment on VisDrone: As shown in Table 2, we conductv ed rigorous evaluations of our method against other SOTA methods on the VisDrone dataset to further verify universality and generalization. The VisDrone dataset features a large target size span covering targets of all scales. From Table 2, HD-BSNet outperforms all other methods across all metrics and can effectively improve detection performance for targets of every size, demonstrating the favorable generalization ability of the proposed method. Notably,

the AP_s of HD-BSNet reaches 20.2%, which surpasses that of DetectoRS by 2.5%, further showcasing the method's advantages in detecting small targets.

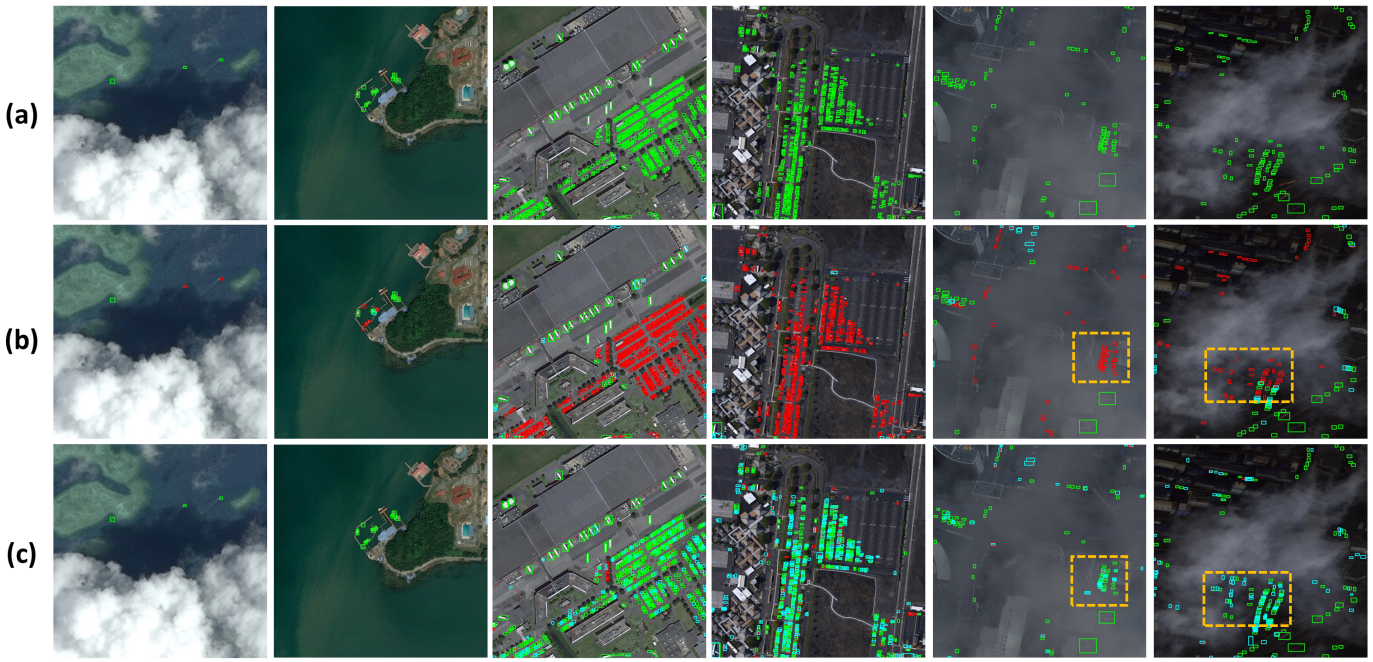


Figure 7. Qualitative comparison map of object detection on the AI-TOD dataset. (**a**) Ground Truth (GT). (**b**) Visualization on DetectoRS. (**c**) Visualization on DetectoRS w/HD-BSNet. Note that green, cyan, and red boxes denote TP, FP, and FN in the detection results, respectively. Comparison results show that HD-BSNet can effectively optimize the small target detection performance of the model in scenarios such as sparse and blurred targets, dense target distribution, and shadow occlusion.

Table 2. Evaluation results of different models on the VisDrone dataset. Red and blue indicate the best and second-best results, respectively.

Method	AP (%)	AP50 (%)	AP75 (%)	AP_s (%)	AP_m (%)	AP_l (%)
YOLOF [54]	15.1	26.3	15.4	6.1	25.2	32.4
RT-DETR [55]	15.9	36.5	10.7	13.8	28.4	23.1
EF-DETR [56]	13.2	24.2	12.3	8.1	18.4	35.7
RetinaNet [24]	17.6	29.6	18.1	8.1	29.4	36.8
AutoAssign [47]	23.2	43.5	21.8	15.2	33.3	40.9
CenterNet [26]	21.4	36.1	21.8	12.6	31.7	38.3
PANet [12]	23.0	38.1	24.2	13.9	34.6	39.0
DINO [52]	26.8	44.2	28.9	17.5	37.3	41.3
Faster R-CNN [21]	24.5	42.5	24.6	16.1	35.9	36.5
Cascade R-CNN [20]	25.6	43.1	26.2	16.4	37.4	41.4
DetectoRS [15]	27.0	45.1	27.5	17.7	39.2	44.5
Faster R-CNN w/HD-BSNet	27.0	47.9	27.1	18.4	37.9	42.2
Cascade R-CNN w/HD-BSNet	28.6	48.6	29.0	19.8	39.9	45.4
DetectoRS w/HD-BSNet	29.2	49.6	29.7	20.2	40.8	49.7

(3) Experiments on DUT Anti-UAV: As shown in Table 3, the proposed method achieves the best performance across overall metrics, with its precision and recall accuracy all outperforming those of other models. Compared with the baseline model YOLOv11, the improved model sees increases of 0.9%, 9.6%, 5.4%, and 6.7% in P, R, AP50, and AP50–95 metrics, respectively. This result not only verifies the excellent cross-model adaptability

of HD-BSNet. This characteristic is specifically reflected in its applicability not only to two-stage detectors but also to one-stage models, where it can still significantly improve the detection accuracy of small objects. It also demonstrates the method's strong cross-scenario generalization capability, as it can effectively enhance detection performance in both remote sensing and low-altitude UAV scenarios.To fully confirm the validity of our approach for detecting small UAVs in intricate background environments, Figure 8 intuitively presents the results of different methods in complex background scenarios. Specifically, compared with other baseline models, the proposed method demonstrates more prominent capability in identifying small UAV targets under complex backgrounds. As can be seen from the comparison of the visualization results in Columns 1–3, by guiding the model to focus its attention on small targets and weakening the focus on background regions, the method effectively improves detection accuracy while significantly reducing the missed detection of small targets; further analysis of the results in Column 4 reveals that this method can also effectively avoid false detections caused by complex background interference.

Table 3. Evaluation results on the DUT Anti-UAV dataset. Red and blue indicate the best and second-best results, respectively.

Method	P	R	AP50	AP50–95	GFLOPs	Params (M)
YOLOv5 [57]	93.4	79.6	88.1	57.1	7.1	2.5
YOLOv6 [58]	91.9	82.5	89.8	59.1	11.7	4.2
YOLOv8 [59]	94.4	82.3	90.1	60.0	8.1	3.0
YOLOv9 [60]	92.8	83.1	89.7	60.1	7.6	1.9
YOLOv10 [61]	93.8	82.6	90.5	59.7	6.5	2.2
YOLOv11 [62]	93.6	82.2	89.7	59.1	6.3	2.5
YOLOv11 w/HD-BSNet	94.5	91.8	95.1	65.8	14.6	2.7

4.3. Ablation Experiments

This section conducts an in-depth discussion on ablation experiments, with the goal of assessing the influence of three novel modules: High-Frequency Differential Perception, Background Semantic Modeling, and Parallel Background Suppression, on the proposed method. All experiments in this part were performed on the AI-TOD test set, with DetectoRS selected as the representative baseline model.

(1) Effectiveness of the High-Frequency Differential Perception Module: We first evaluate the effectiveness of the proposed High-Frequency Differential Perception Module, which can effectively capture the high-frequency information of small objects that is lost during downsampling. As shown in the Table 4, after introducing HDPM into the baseline model, the detection AP increases by 10.6% (from 16.6% to 27.2%) and AP50 rises by 25.2% (from 33.3% to 58.5%), with the most significant performance improvements observed in AP_{vt} and AP_s for small object detection. This finding indicates that the loss of key detailed information of small objects during feature extraction and fusion degrades detection performance, and this lost information is mainly located in high-frequency components. Capturing the high-frequency information lost during downsampling through differential operations and re-enhancing the details of shallow features can effectively enhance the target representation capability.

(2) Effectiveness of the Background Semantic Modeling Module: After introducing BSMM into the baseline model, the AP and AP50 on the AI-TOD dataset increase by 11.8% and 25.6%, respectively, as shown in Table 4. Notably, the AP_{vt} and AP_t for small objects rise by 10.8% and 19.3%, respectively, fully demonstrating the effectiveness of background semantic modeling in improving small object detection accuracy. By injecting semantic cues that distinguish foreground from background into the shallow P_2 features, the model

can learn more discriminative target features, thereby enhancing the accuracy of target localization and recognition.

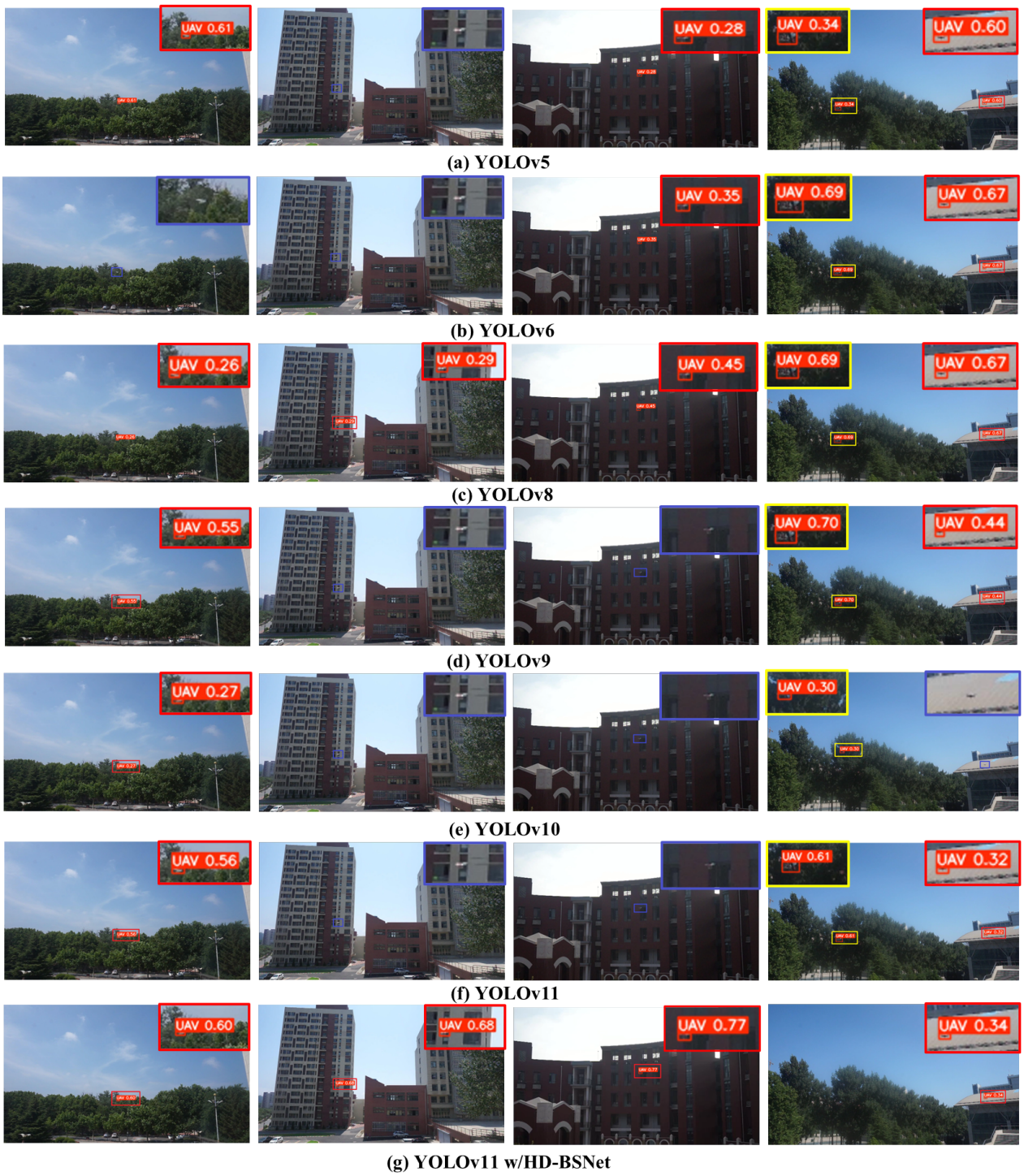


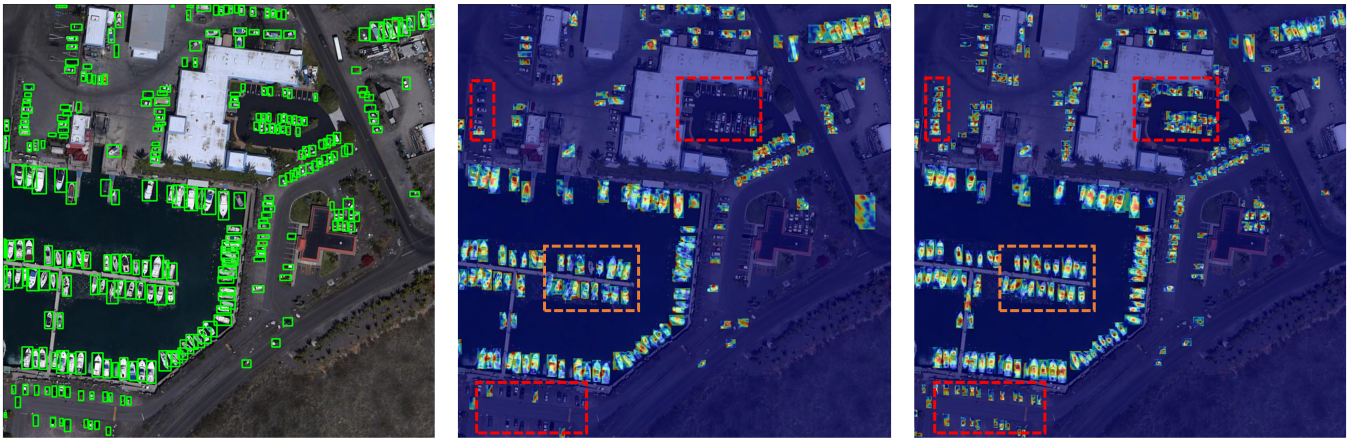
Figure 8. Comparison diagram of detection results of different models on the DUT Anti-UAV dataset. From left to right are four different scenarios: complex backgrounds, similar backgrounds, dark backgrounds, and confusing backgrounds. Note that red, yellow, and purple bounding boxes denote TP, FP, and FN in the detection results, respectively.

(3) Effectiveness of the Parallel Multi-Scale Focusing Module: We visualized the local attention maps of the DetectoRS model with and without PMFM, as shown in Figure 9. In the regions marked by orange dashed lines in the figure, it can be clearly observed

that after adding PMFM, the model's attention can focus more accurately on the target center region. This is attributed to its parallel multi-scale aggregation and normalization operations, which effectively enhance target features and suppress background interference. It can be further seen from the parts marked by red dashed lines that PMFM also helps the model capture small targets that were originally missed, improving the recall capability.

Table 4. Ablation experiments on the AI-TOD test set. HDPM denotes the high-frequency differential perception module. BSMM denotes the background semantic modeling module. PMFM denotes the parallel multi-scale focusing module. Red and blue indicate the best and second-best results, respectively.

HDPM	BSMM	PMFM	AP (%)	AP50 (%)	AP75 (%)	AP_{vt} (%)	AP_t (%)	AP_s (%)
			16.6	33.3	14.5	0.1	12.1	31.3
✓			27.2	58.5	23.6	10.3	29.8	34.1
	✓		28.4	58.9	26.0	10.9	31.4	35.6
		✓	26.6	57.1	23.2	8.7	28.8	33.8
✓	✓		28.9	60.2	25.9	11.1	31.4	36.4
✓		✓	27.3	58.3	24.1	10.7	29.9	34.4
	✓	✓	29.0	60.3	26.0	11.0	31.8	36.8
✓	✓	✓	29.8	61.7	26.9	11.5	32.7	36.9



(a) GT **(b) DetectoRS** **(c) DetectoRS w/PMFM**

Figure 9. Qualitative comparison diagram of local attention distribution. Within the area marked by the orange dashed line, the model's attention focuses more precisely on the center region of small targets after integrating PMFM—this verifies its role in local high-response focusing. Within the area marked by the red dashed line, the attention successfully captures small targets that were previously missed, confirming its effect in reducing the miss detection rate.

As shown in Table 4, after introducing PMFM into the baseline model, the detection AP is improved by 10% and APvt by 8.6%. In addition, we further compared PMFM with its inspiration source LAPM [32]. As shown in Table 5, PMFM achieves a higher value in the AP metric compared with LAPM, while the computational overhead remains at the same level. Taken together, these results fully verify the effectiveness of PMFM in improving the accuracy of small object detection.

Table 5. Comparison between PMFM and LAPM on AI-TOD.

PMFM	LAPM	AP (%)	AP50 (%)	AP75 (%)	AP_t (%)	AP_s (%)	GFLOPs
✓		26.6	57.1	23.2	28.8	33.8	152
	✓	26.3	56.8	23.1	28.6	33.7	152

5. Discussion

Given that the functions and mechanisms of each module have been elaborated in the ablation experiments, this section will focus on an in-depth analysis of the core limitations of this study and attribute their causes based on the methodological principles.

5.1. Limitations and Causes of Detection Performance for Verytiny Targets

The experimental results in Table 1 demonstrate that HD-BSNet yields relatively limited performance improvements in the detection of ultra-tiny targets (sized [2^2, 8^2] pixels). Taking the AI-TOD dataset as the test benchmark, the model with DetectoRS as the backbone network achieves a mere 11.5% average precision for verytiny targets (AP_{vt}). This limitation arises from the combined effects of the inherent feature defects of veryvtiny targets and the design logic of the core modules in the framework.

From the perspective of the working mechanism of the HDPM: it extracts high-frequency details of targets through differential operations between the original image and the shallow feature map P_2. However, verytiny targets occupy an extremely low spatial proportion in the P_2 feature map, and the intensity of their inherent high-frequency signals is far weaker than that of background noise in complex scenarios. This renders differential operations unable to effectively separate target signals from background interference, thus failing to achieve targeted feature enhancement.

On the other hand, the window partitioning granularity of the PMFM is constrained by the resolution of the feature map. Its minimum window size is still larger than the pixel range of verytiny targets, making it impossible to accurately focus on such targets. As a result, the enhanced features remain susceptible to background information interference, which further compromises detection accuracy.

In addition, the feature dimensions of verytiny targets are extremely sparse and their semantic information is inherently insufficient—a common challenge in the field of small object detection. Although the existing dual-mechanism enhancement strategy can strengthen the feature representation of conventional small targets, it cannot compensate for the lack of semantic information in verytiny targets, which further limits the model's ability to recognize such targets.

5.2. Conflict Between Computational Complexity and Real-Time Deployment and Its Causes

HD-BSNet incurs high computational costs, which presents a significant gap with the requirements of real-time deployment. The experimental data in Table 3 indicate that integrating this framework into the YOLOv11 model increases the computational complexity (GFLOPs) from 6.3 of the baseline model to 14.6, representing a computational overhead increase of approximately 132%. This makes it difficult to deploy the model on resource-constrained embedded devices or airborne systems. The core cause of this problem lies in the convolutional operation design of the two core modules.

First, the HDPM adopts a progressive standard convolution structure with kernel sizes of 3 × 3→5 × 5→7 × 7, instead of leveraging the sparse sampling characteristics of dilated convolution. To achieve full coverage of high-frequency information regions of targets, dense computations are required for feature extraction, which directly leads to a substantial increase in computational complexity.

Second, the BSMM employs a parallel design of multi-scale dilated convolutions with dilation rates of 1, 3, 5, and 7 to expand the receptive field and capture multi-scale background semantic information. Although the number of kernel parameters of a single 3 × 3 dilated convolution is fixed, the parallel computation of multi-scale branches accumulates the overall computational overhead. This ultimately reduces the inference

speed of the model, making it unable to meet the application requirements of real-time detection scenarios.

6. Conclusions

To address the limitations in small object detection, including the accuracy gap, high false detection rate, and poor framework adaptability, this paper proposes a plug-and-play HD-BSNet framework. First, HDPM captures high-frequency details of small targets via differential operations, enhancing foreground representation without introducing additional noise. Second, BSMM further improves the discriminability between small targets and the background through separate modeling and fusion of foreground and background information. In addition, PMFM leverages normalization-based focused attention to further strengthen target representation. Evaluations on the AI-TOD, VisDrone, and DUT Anti-UAV datasets demonstrate that HD-BSNet is compatible with both one-stage and two-stage object detectors, and achieves significant detection accuracy improvements in two typical scenarios, remote sensing and low-altitude UAV. These results fully verify the effectiveness and generalization capability of the proposed framework. It should be noted that, due to the integration of multi-scale perception modules, the framework has a moderate number of parameters but relatively high computational complexity. Future research will focus on optimizing small object detection in low-altitude UAV scenarios, with an emphasis on reducing computational overhead through lightweight design to further improve the framework's deployment feasibility in practical applications.

Author Contributions: Conceptualization, J.W. and T.C.; data curation, J.W. and D.J.; investigation, N.P. and J.Z.; methodology, J.W. and X.Z.; software, J.W.; writing—original draft J.W.; writing—review and editing, J.W., X.Z., N.P., J.Z. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Youth Innovation Promotion Association of Chinese Academy of Sciences No. 2021376.

Data Availability Statement: The original contributions presented in the study are included in the article. The code is available at: https://github.com/Winslow666/HD-BSNet (accessed on 18 December 2025) Further inquiries can be directed to the corresponding author.

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

References

1. Ren, Z.; He, L.; Lu, J. Context aware edge-enhanced GAN for remote sensing image super-resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1363–1376. [CrossRef]
2. Zhu, Y.; Huang, Y.; Yang, M.; Mao, D.; Zhang, Y.; Jiao, L.; Zhang, Y.; Yang, J. SAR Image Super-resolution based on Multi-scale Edge Texture-oriented GAN Approach. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 20359–20374. [CrossRef]
3. Zhang, Y.; Wei, S.; Sun, Y.; Shen, J.; Yang, Z.; Yan, J. EESAGAN: Edge-Enhanced and Structure-Aware GAN for Remote Sensing Image Super-Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 24947–24962. [CrossRef]
4. Yi, H.; Liu, B.; Zhao, B.; Liu, E. Small object detection algorithm based on improved YOLOv8 for remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1734–1747. [CrossRef]
5. Zheng, X.; Qiu, Y.; Zhang, G.; Lei, T.; Jiang, P. ESL-YOLO: Small object detection with effective feature enhancement and spatial-context-guided fusion network for remote sensing. *Remote Sens.* **2024**, *16*, 4374. [CrossRef]
6. Fu, C.; Yuan, H.; Shen, L.; Hamzaoui, R.; Zhang, H. 3DAttGAN: A 3D attention-based generative adversarial network for joint space-time video super-resolution. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 3117–3128. [CrossRef]
7. Cao, B.; Yao, H.; Zhu, P.; Hu, Q. Visible and clear: Finding tiny objects in difference map. In *Computer Vision–ECCV 2024. ECCV 2024. Lecture Notes in Computer Science*; Springer Nature: Cham, Switzerland, 2024; pp. 1–18.
8. Li, Y.; Luo, J.; Zhang, Y.; Tan, Y.; Yu, J.-G.; Bai, S. Learning to holistically detect bridges from large-size VHR remote sensing imagery. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11507–11523. [CrossRef]

9. Xu, C.; Wang, J.; Yang, W.; Yu, L. Dot distance for tiny object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021; pp. 1192–1201.
10. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.-S. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [CrossRef]
11. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
12. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
13. Ge, L.; Wang, G.; Zhang, T.; Zhuang, Y.; Chen, H.; Dong, H.; Chen, L. Regression-guided refocusing learning with feature alignment for remote sensing tiny object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
14. Liu, D.; Zhang, J.; Qi, Y.; Wu, Y.; Zhang, Y. Tiny object detection in remote sensing images based on object reconstruction and multiple receptive field adaptive feature enhancement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. [CrossRef]
15. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.
16. Li, Y.; Wang, L.; Wang, T.; Yang, X.; Luo, J.; Wang, Q.; Deng, Y.; Wang, W.; Sun, X.; Li, H.; et al. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1832–1849. [CrossRef]
17. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
18. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
20. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
21. 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]
22. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
24. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
26. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
27. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
28. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
29. Zhou, Z.; Zhu, Y. KLDet: Detecting tiny objects in remote sensing images via Kullback–Leibler divergence. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
30. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Shim, S.-H.; Hyun, S.; Bae, D.; Heo, J.-P. Local attention pyramid for scene image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7774–7782.
33. Wang, J.; Yang, W.; Guo, H.; Zhang, R.; Xia, G.-S. Tiny object detection in aerial images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3791–3798.

34. 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 IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
35. Zhao, J.; Zhang, J.; Li, D.; Wang, D. Vision-based anti-uav detection and tracking. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25323–25334. [CrossRef]
36. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
37. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
38. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019. Available online: https://dl.acm.org/doi/10.5555/3454287.3455008 (accessed on 22 January 2026).
40. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155. [CrossRef]
41. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.-S. RFLA: Gaussian receptive field based label assignment for tiny object detection. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–29 October 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 526–543.
42. Shi, Z.; Hu, J.; Ren, J.; Ye, H.; Yuan, X.; Ouyang, Y.; He, J.; Ji, B.; Guo, J. HS-FPN: High frequency and spatial perception FPN for tiny object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–28 February 2025; Volume 39, pp. 6896–6904.
43. Liu, H.-I.; Tseng, Y.-W.; Chang, K.-C.; Wang, P.-J.; Shuai, H.-H.; Cheng, W.-H. A denoising fpn with transformer r-cnn for tiny object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4704415. [CrossRef]
44. Hu, H.; Chen, S.B.; Tang, J. CFENet: Contextual Feature Enhancement Network for Tiny Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703113. [CrossRef]
45. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
46. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6054–6063.
47. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. Autoassign: Differentiable label assignment for dense object detection. *arXiv* **2020**, arXiv:2007.03496. [CrossRef]
48. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
49. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Washington, DC, USA 2021; pp. 3490–3499.
50. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
51. Guo, G.; Chen, P.; Yu, X.; Han, Z.; Ye, Q.; Gao, S. Save the tiny, save the all: Hierarchical activation network for tiny object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 221–234. [CrossRef]
52. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
53. Zhang, T.; Zhang, X.; Zhu, X.; Wang, G.; Han, X.; Tang, X.; Jiao, L. Multistage enhancement network for tiny object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611512. [CrossRef]
54. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13039–13048.
55. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–20 June 2024; pp. 16965–16974.
56. Cheng, S.; Song, J.; Zhou, M.; Wei, X.; Pu, H.; Luo, J.; Jia, W. Ef-detr: A lightweight transformer-based object detector with an encoder-free neck. *IEEE Trans. Ind. Inform.* **2024**, *20*, 12994–13002. [CrossRef]

57. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Fang, J.; Michael, K.; Montes, D.; Nadar, J.; Skalski, P.; et al. ultralytics/yolov5: V6. 1-Tensorrt, Tensorflow Edge Tpu and Openvino Export and Inference. Zenodo, 2022. Available online: https://zenodo.org/records/6222936 (accessed on 22 January 2026). [CrossRef]
58. 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]
59. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO, Ver. 8.0.0, January 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 22 January 2026).
60. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–29 October 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 1–21.
61. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
62. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [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.