remote sensing

MDPI

Article

HiDRA-DCDNet: Dynamic Hierarchical Attention and Multi-Scale Context Fusion for Real-Time Remote Sensing Small-Target Detection

Jiale Wang [1,2], Zhe Bai [1], Ximing Zhang [1,*], Yuehong Qiu [1], Fan Bu [1] and Yuancheng Shao [1]

[1] Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an 710119, China
[2] University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: zhangximing@opt.ac.cn

Abstract

Small-target detection in remote sensing presents three fundamental challenges: limited pixel representation of targets, multi-angle imaging-induced appearance variance, and complex background interference. This paper introduces a dual-component neural architecture comprising Hierarchical Dynamic Refinement Attention (HiDRA) and Densely Connected Dilated Block (DCDBlock) to address these challenges systematically. The HiDRA mechanism implements a dual-phase feature enhancement process: channel competition through bottleneck compression for discriminative feature selection, followed by spatial-semantic reweighting for foreground–background decoupling. The DCDBlock architecture synergizes multi-scale dilated convolutions with cross-layer dense connections, establishing persistent feature propagation pathways that preserve critical spatial details across network depths. Extensive experiments on AI-TOD, VisDrone, MAR20, and DOTA-v1.0 datasets demonstrate our method's consistent superiority, achieving average absolute gains of +1.16% (mAP_{50}), +0.93% (mAP_{95}), and +1.83% (F1-score) over prior state-of-the-art approaches across all benchmarks. With 8.1 GFLOPs computational complexity and 2.6 ms inference speed per image, our framework demonstrates practical efficacy for real-time remote sensing applications, achieving superior accuracy–efficiency trade-off compared to existing approaches.

Keywords: remote sensing imagery; hierarchical attention mechanism; multi-scale feature fusion; target detection

check for updates

Academic Editor: Pedro Melo-Pinto

Received: 29 April 2025
Revised: 13 June 2025
Accepted: 23 June 2025
Published: 25 June 2025

Citation: Wang, J.; Bai, Z.; Zhang, X.; Qiu, Y.; Bu, F.; Shao, Y. HiDRA-DCDNet: Dynamic Hierarchical Attention and Multi-Scale Context Fusion for Real-Time Remote Sensing Small-Target Detection. *Remote Sens.* **2025**, *17*, 2195. https://doi.org/10.3390/rs17132195

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing small-target detection, which focuses on the precise identification and localization of sub-50-pixel targets in aerial imagery, represents a foundational challenge for critical applications ranging from environmental surveillance to emergency management [1]. Distinct from conventional detection scenarios, remote sensing targets (e.g., vehicles, vessels, and artificial structures) typically present as densely distributed low-contrast entities (occupying <0.1% image area) embedded in heterogeneous backgrounds. The proliferation of multi-modal imaging systems—encompassing temporal, spectral, and resolution variations—introduces compounded challenges of scale variance and feature ambiguity that conventional detection frameworks struggle to resolve.

The core challenge lies in resolving the signal-to-noise dichotomy: retaining discriminative features from spatially sparse targets while effectively suppressing pseudo-target

patterns in cluttered environments. This challenge becomes particularly acute in low-resolution regimes where critical edge information degrades into indistinct spectral clusters, especially for sub-20-pixel targets. Furthermore, significant intra-class variability—caused by imaging geometry variations (e.g., nadir vs. oblique acquisition) and environmental factors (e.g., shadow artifacts, phenological changes)—leads to substantial appearance discrepancies among targets of the same category [2]. Although traditional CNN architectures have shown success in natural image domains, their application to this specialized task reveals three critical shortcomings: (1) progressive attenuation of spatial details in deeper layers, (2) limited contextual modeling capacity for ultra-small targets, and (3) ineffective management of feature interdependencies in current attention mechanisms [3].

Contemporary advances in attention mechanisms (e.g., SE [4], CBAM [5]) and dense connectivity architectures have yielded partial solutions, yet our systematic analysis exposes critical limitations. Single-stage attention modules risk over-suppression of latent features through aggressive channel-wise pruning, while conventional dense blocks exhibit computational redundancy during multi-scale feature aggregation. To overcome these constraints, we present a paradigm-shifting framework integrating a Hierarchical Dynamic Refinement Attention (HiDRA) mechanism with a novel Densely Connected Dilated Block (DCDBlock) architecture.

The proposed DCDBlock introduces two key innovations: the synergistic integration of dense connectivity patterns and the incorporation of the HiDRA mechanism, collectively enhancing feature representational capacity while maintaining computational efficiency. Specifically, this architecture addresses three persistent challenges in remote sensing small-target detection: insufficient discriminative feature extraction from ultra-small targets, complex background interference suppression, and robustness to high intra-class variability. Through hierarchical feature refinement and dynamic channel-weight optimization, the framework demonstrates superior detection accuracy compared to existing approaches, particularly in low signal-to-noise ratio scenarios. Notably, the HiDRA component implements multi-scale attention recalibration through a novel gating mechanism that preserves critical spatial-semantic information while suppressing redundant features.

Our main contributions are as follows:

- **HiDRA Mechanism:** A dual-resolution attention paradigm combining competitive channel selection (via adaptive bottlenecking) and global contextual reweighting. Unlike conventional single-stage attention, HiDRA prevents feature over-smoothing by decoupling channel competition from cross-feature dependency modeling;
- **DCDBlock Architecture:** A feature extraction module unifying dilated convolutions, dense cross-stage connections, and embedded HiDRA units. This design achieves multi-scale context aggregation without spatial downsampling, explicitly addressing the resolution-preservation versus receptive field trade-off;
- **Empirical Validation:** Extensive experiments on the AI-TOD, VisDrone, MAR20, and DOTA-v1.0 datasets demonstrate our method's consistent superiority, achieving average absolute gains of +1.16% (mAP50), +0.93% (mAP95), and +1.83% (F1-score) over prior state-of-the-art approaches across all benchmarks, with HiDRA alone contributing 63% of the mAP gains in ablation studies.

2. Material and Methods

2.1. Main Challenges and Solutions for Small-Target Detection in Remote Sensing

2.1.1. Complex Backgrounds and High-Resolution Images

Remote sensing imagery inherently encompasses expansive geographical regions characterized by rich yet complex background information. Small-target detection in such contexts presents unique challenges compared to conventional scenarios, as target signals

exhibit diminished intensity and reduced discriminability against intricate backgrounds [6]. High-resolution imaging further compounds these difficulties by simultaneously offering enhanced detail and introducing confounding factors. The coexistence of diverse land cover types creates spectral and textural ambiguities between small targets and their surroundings, while increased noise levels demand more sophisticated feature extraction capabilities. These factors collectively contribute to elevated false positive rates, particularly for targets with limited spatial extent [7].

To mitigate these challenges, contemporary research has prioritized the development of specialized architectural components to augment model perception. Feature enhancement modules, attention mechanisms, and spatial context-aware architectures have emerged as particularly effective solutions. The FFCA-YOLO [8] framework exemplifies this approach through its integration of three core components within the YOLOv5 architecture: Feature Enhancement Modules (FEM) for discriminative feature learning, Feature Fusion Modules (FFM) for multi-scale information synthesis, and Spatial Context-Aware Modules (SCAM) for contextual relationship modeling. Parallel innovations include the YOLT [9] architecture, which incorporates passthrough layers inspired by ResNet's residual principles. This design facilitates cross-layer connectivity between multi-level feature maps, enabling the preservation of high-fidelity spatial details critical for small-target recognition while maintaining computational efficiency. Such architectural innovations demonstrate the field's progression toward context-aware, multi-scale feature representation strategies tailored for remote sensing challenges.

2.1.2. Small-Target Diversity and Scale Variation

Small-target detection in remote sensing imagery presents unique challenges due to substantial scale variations and heterogeneous target characteristics, which demand robust detection frameworks with exceptional recognition capacity and scale-invariant representation [10]. Contemporary approaches predominantly employ multi-scale feature pyramid architectures to address these challenges through hierarchical feature extraction and cross-scale information integration. Various architectural innovations have been developed to implement feature pyramid structures that exhibit distinct operational characteristics. The Deconvolutional Single Shot Detector (DSSD) [11] employs cascaded deconvolution layers with skip connections to progressively enhance feature map resolution while combining shallow spatial details with deep semantic information, forming an asymmetric pyramid configuration. Spatial pyramid pooling mechanisms, exemplified by the seminal Spatial Pyramid Pooling (SPP) [12] module, leverage multi-scale pooling operations to preserve critical image information across varying input dimensions through adaptive feature aggregation.

Advanced variants further enhance these architectures through novel computational paradigms. The Atrous Spatial Pyramid Pooling (ASPP) [13] integrates dilated convolution operations with traditional SPP, effectively expanding receptive fields while maintaining spatial resolution through strategic dilation rate combinations. Subsequent optimizations like Depth-wise Separable Pyramid Pooling (DSPP) [14] and Mixed Depth-wise Separable Pyramid Pooling (MDSPP) [14] implement depth-wise separable convolution layers within pyramid structures, achieving computational efficiency improvements without compromising feature representation quality. The Feature Pyramid Network (FPN) [15] framework establishes a bidirectional feature hierarchy through lateral connections and top-down pathway integration, enabling synergistic combination of semantically rich deep features and spatially precise shallow features. Building upon FPN, the Attention Feature Pyramid Network (A-FPN) [16] incorporates channel-wise attention mechanisms to dy-

namically weight features from different receptive fields, enhancing discriminative feature representation through saliency-guided fusion.

These multi-scale pyramid architectures collectively improve small-target detection performance in remote sensing applications by establishing comprehensive feature hierarchies that simultaneously preserve fine-grained spatial details and high-level semantic context. Through strategic combination of scale-adaptive operations, context-aware feature fusion, and computational optimization techniques, modern pyramid networks effectively address the intrinsic challenges of small-target variability in complex aerial imagery scenarios [17].

2.1.3. Computing Resources and Real-Time Requirements

The processing of large-scale remote sensing imagery poses significant computational demands while requiring stringent real-time operational capabilities, particularly for deployment on resource-constrained mobile platforms [18]. This dual challenge has propelled lightweight network architecture design to the forefront of contemporary research efforts. Current methodologies primarily focus on optimizing computational efficiency through model compression techniques and architectural innovations, effectively reducing parameter complexity while maintaining detection accuracy. Knowledge distillation has emerged as a pivotal strategy, enabling the transfer of representational capabilities from computationally intensive teacher models to compact student networks through hierarchical feature alignment. Complementary to this, neural network pruning techniques systematically eliminate redundant weights and inactive neurons through magnitude-based or gradient-aware criteria, achieving substantial model footprint reduction without compromising operational efficacy [19].

Recent advancements demonstrate particular promise in domain-specific lightweight architecture redesign. The ShuffleNet [20,21] and MobileNet [22–24] series establish benchmark performance through depth-wise separable convolution operations and channel shuffling mechanisms, while GhostNet [25,26] further enhances computational efficiency through intrinsic feature map redundancy reduction. Among the specialized solutions for aerial imagery analysis, the Difference-Enhanced Spatial-Spectral Network (DESSN) [27] represents a notable advancement in change detection for ultra-high-resolution remote sensing data. This architecture synergistically combines an Asymmetric Dual Convolution Ghost (ADCG) module for computational complexity reduction with a Non-Local Spatial-Spectral Fusion (SSN) module that enhances edge preservation through cross-dimensional attention mechanisms. The ADCG component employs factorized ghost convolution operations to decouple spatial and channel-wise feature learning, while the SSN module establishes long-range dependencies across spectral bands and spatial regions through adaptive feature recalibration. This dual challenge has propelled the design of lightweight network architectures to the forefront of contemporary research efforts.

These architectural innovations collectively address the fundamental trade-off between computational efficiency and detection accuracy in mobile remote sensing applications. Through strategic integration of model compression paradigms, hardware-aware operator design, and cross-modal feature fusion mechanisms, modern lightweight networks demonstrate remarkable adaptability to the stringent requirements of real-time aerial image analysis across diverse operational environments.

2.2. Attention Mechanism in Object Detection

The remarkable ability of the human visual system to efficiently localize salient regions in complex scenes has inspired significant advancements in attention mechanisms used in lightweight remote sensing target detection networks. These biologically inspired com-

putational paradigms have emerged as a critical research frontier, primarily manifesting as three architectural variants: channel-wise attention, spatial attention, and their hybrid combinations. The fundamental principle involves dynamic feature recalibration through differential weighting of channels or spatial regions, deviating from conventional convolution and pooling operations that treat all positions equivalently. This selective emphasis enables networks to prioritize information extraction from task-critical components while suppressing irrelevant features [28].

Channel attention mechanisms, pioneered by the Squeeze-and-Excitation (SE) module, establish foundational principles through global average pooling and multi-layer perceptron-based channel weighting. While effectively enhancing discriminative channel representations, SE's dimensionality reduction during excitation may induce spatial information loss. Subsequent developments like Efficient Channel Attention (ECA) [29] address this limitation through 1×1 convolutional cross-channel interaction without dimensionality reduction, achieving enhanced feature discrimination with minimal parameter overhead.

Spatial attention architectures counterbalance channel-focused approaches through positional sensitivity enhancement. The Coordinate Attention (CA) [30] mechanism innovatively decomposes spatial encoding into orthogonal directional streams, employing separate 1D feature aggregation along height and width dimensions. This dual-path architecture simultaneously preserves precise positional information and captures long-range dependencies, generating orientation-aware feature maps that improve localization accuracy for irregularly shaped targets.

Hybrid attention frameworks achieve comprehensive feature optimization through the synergistic integration of spatial and channel information. The Convolutional Block Attention Module (CBAM) exemplifies this approach through sequential channel–spatial attention cascading, where initial channel-wise feature refinement is followed by spatial importance weighting. This hierarchical attention process enables context-aware feature enhancement across multiple dimensions, adaptively optimizing network focus based on both semantic relevance and spatial configuration.

While emerging variants like self-attention and temporal attention show theoretical promise, their computational complexity and parametric demands often preclude deployment in resource-constrained detection systems [31]. Future research directions emphasize neuromorphic computing principles, particularly the development of biologically plausible attention pretraining paradigms and computationally efficient architectural optimizations. Key challenges remain in balancing attention mechanism sophistication with operational efficiency, particularly for edge-computing implementations requiring real-time processing of high-resolution aerial imagery [32].

2.3. Proposed Method

Small-target detection in aerial imagery presents unique challenges stemming from three intrinsic characteristics: limited target pixel occupancy, complex background interference, and multi-scale imaging variations. To address these limitations, we propose a novel architecture combining Hierarchical Dual-Resolution Attention (HiDRA) with Densely Connected Dilated Blocks (DCDBlocks), effectively mitigating feature degradation and scale insensitivity in conventional detection frameworks.

2.3.1. Contextual Motivation and HiDRA Architecture

The inherent complexity of remote sensing scenes requires specialized attention mechanisms capable of suppressing background clutter while amplifying subtle target signatures. Traditional approaches often fail to simultaneously achieve these dual objectives,

particularly for sub-10px targets like vehicles and marine vessels. Our HiDRA mechanism overcomes this through a hierarchical two-stage weighting strategy, as diagrammed in Figure 1.

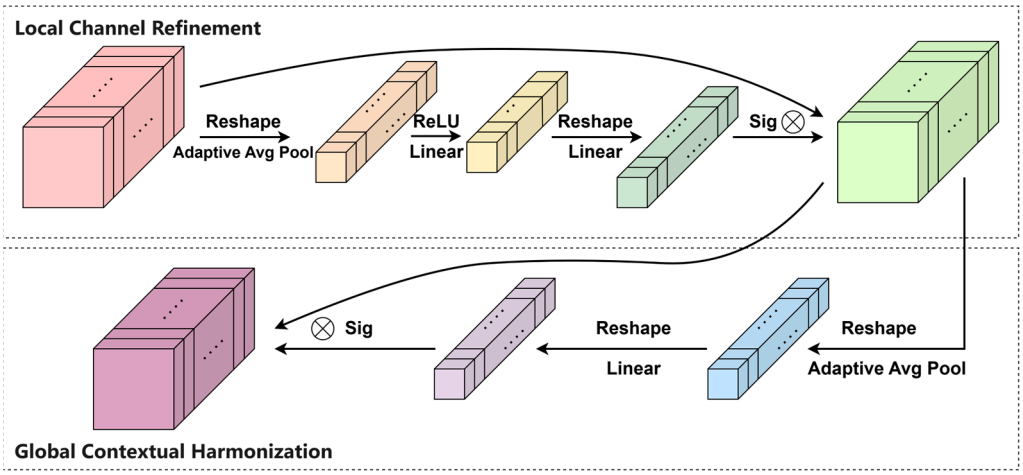


Figure 1. The structural diagram of the HiDRA module.

The first stage, Local Channel Refinement, dynamically modulates channel-wise responses using an adaptive bottleneck structure. Given an input feature tensor $x \in R^{B\times C\times H\times W}$, HiDRA first computes channel statistics via global average pooling:

$$y' = Reshape(GAP(x)) \in R^{B\times C} \tag{1}$$

A non-linear transformation with learnable parameters $W_1 \in \mathbb{R}^{C'\times C}$ and $W_2 \in \mathbb{R}^{C\times C'}$ then generates attention weights:

$$y'' = \sigma\left(W_2 \cdot ReLU\left(W_1 \cdot y'\right)\right) \tag{2}$$

Here, GAP refers to adaptive global average pooling, while $W_1 \in \mathbb{R}^{C'\times C}$ and $W_2 \in \mathbb{R}^{C\times C'}$ represent linear projection matrices that constitute an adaptive bottleneck structure. Specifically, W_1 and W_2 are responsible for reducing and restoring the channel dimensionality, respectively, thereby capturing non-linear inter-channel dependencies and generating attention coefficients through a Sigmoid activation function σ. These coefficients are subsequently used to recalibrate the input feature channels. In contrast to conventional Squeeze-and-Excitation (SE) blocks, the compression ratio r is set to 16 by default. To mitigate the risk of excessive information loss, especially in shallow layers, the reduced dimension C' is adaptively determined as $\max\left(\frac{C}{r}, \frac{C}{2}\right)$, ensuring that at least 50% of the original channel capacity is preserved.

The second stage, Global Contextual Harmonization, introduces spatial-agnostic attention to model cross-channel dependencies at full dimensionality:

$$y_{HiDRA} = \sigma(W_3 \cdot AdaptiveAvgPool(X \otimes y'')) \tag{3}$$

The weight matrix $W_3 \in \mathbb{R}^{C\times C}$ performs a linear transformation on the globally pooled feature representation without applying dimensionality reduction. This operation enables the model to explicitly learn the global dependency structure across channels, thereby facilitating fine-grained coupling between them. It serves as a second-stage refinement that selectively reweighs the already modulated feature maps.

The HiDRA module introduces a hierarchical dual-stage attention mechanism that incrementally enhances feature representations, setting it apart from traditional single-stage

attention designs. In the first stage, a compressive excitation mechanism identifies and emphasizes discriminative channels, effectively filtering out irrelevant or noisy information. This is followed by a second-stage global interaction, implemented via a full-dimensional linear transformation matrix W_3, which models long-range inter-channel dependencies. This two-stage, dual-resolution paradigm is particularly advantageous for small object detection, as it mitigates the risk of weak target signals being suppressed or diluted during deep feature propagation. At its core, HiDRA integrates both Channel Attention and Global Attention to achieve multi-level feature optimization. The Channel Attention module dynamically adjusts the importance of each feature channel, enhancing those that contribute to accurate target detection while suppressing distractive ones. On top of this, the Global Attention module further recalibrates the feature space by incorporating holistic contextual information, ensuring global consistency and coherence in the feature representation. This integration strengthens the network's ability to model synergistic relationships across channels and improves semantic alignment across spatial locations. By jointly modeling local importance and global context, HiDRA enhances the network's discriminative capacity, making it particularly effective in complex remote sensing scenarios with small or obscured targets. Compared to existing attention mechanisms, HiDRA delivers superior detection accuracy and robustness while maintaining computational efficiency, thereby offering a powerful and practical enhancement for real-time remote sensing target detection frameworks.

2.3.2. Design and Implementation of the Feature Extraction Module DCDBlock

The DCDBlock module embodies a synergistic integration of the Hierarchical Dynamic Refinement Attention (HiDRA) mechanism and multi-path fusion principles, specifically tailored to address the dual challenges of feature degradation and computational efficiency in remote sensing small-target detection. Traditional convolutional architectures often struggle to balance receptive field coverage and feature resolution, particularly for small targets embedded in complex backgrounds such as urban clutter or maritime environments. The DCDBlock resolves this through a hierarchical architecture that prioritizes high-frequency feature preservation while enabling computationally efficient multi-scale reasoning.

As depicted in Figure 2, the DCDBlock initiates with a 1×1 convolution layer that projects input features into a compressed latent space effectively decoupling redundant background patterns from potential target signatures. Four cascaded DCDBottleneck units then progressively expand the receptive field through a novel dense-dilated concatenation mechanism. Each DCDBottleneck integrates three critical components: Depth-wise Separable Convolution, Hierarchical Feature Concatenation and Attention-Guided Feature Reweighting. In the DCDBottleneck, the dilated convolution employs a kernel size of 3 and a stride of 1. To maximize the receptive field while preserving the spatial resolution, both the dilation rate and padding are set to 2. This replaces standard 3×3 convolutions to reduce parameters by 82.5% while maintaining spatial sensitivity, crucial for preserving edge features of diminutive targets. The operation decomposes into

$$X' = Conv_1(HiDRA(DWConv(Conv_1(BN(X))))) \tag{4}$$

$$D(X) = X \oplus X' \tag{5}$$

Here, X and X' denote the input feature map and the enhanced feature map obtained after a series of convolutional operations and the HiDRA attention module, respectively. The final output of the module is defined as $D(X)$, which denotes element-wise residual addition. This residual connection preserves the original semantic information from the input feature map X, while simultaneously integrating the enhanced representation X'

generated through the convolutional operations and the HiDRA attention mechanism. Such a design facilitates gradient flow during backpropagation and alleviates potential degradation in feature quality caused by deep transformation, thereby promoting more stable and efficient network optimization.

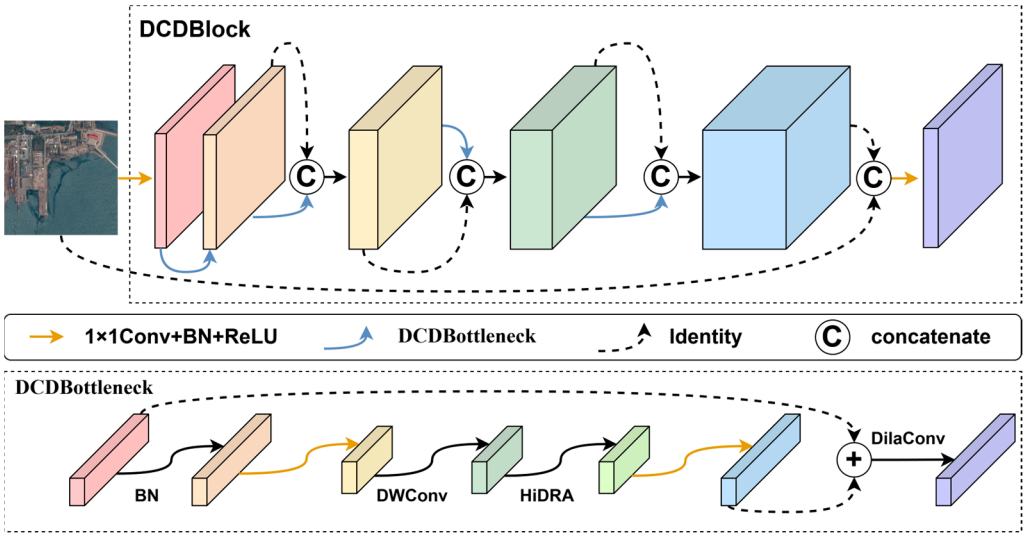


Figure 2. The structural diagram of DCDBlock and DCDBottleneck.

Unlike traditional residual connections that risk feature dilution, the block employs exponential channel expansion through nested concatenation:

$$X^{(k)} = B_k\left(\left[X^{(k-1)}; X^{(k-2)}\right]\right),\ k \in \{1,\ 2,\ 3,\ 4\} \tag{6}$$

Here, $[;]$ denotes channel-wise concatenation, B_k represents the k-th DCDBottleneck module, and $X^{(k)}$ denotes the corresponding output of that layer. In other words, the output feature map of each layer is obtained by concatenating the input of the layer with the output of the corresponding DCDBottleneck module along the channel dimension. This geometric growth pattern ($C' \rightarrow 2C' \rightarrow 4C' \rightarrow 8C'$) ensures multi-scale context preservation while maintaining linear computational complexity.

Finally, the original input X to the DCDBlock undergoes four successive DCDBottleneck layers with dense connections, resulting in an intermediate feature map $X^{(4)}$. This intermediate output is then concatenated with the original input X along the channel dimension, followed by a 1×1 convolution to adjust the channel dimensions, yielding the final output y_{out}:

$$y_{out} = Conv_1([X^{(4)};\ X]) \tag{7}$$

The DCDBlock's efficacy arises from its synergistic integration of multi-scale context modeling and hierarchical attention refinement. The depth-wise dilated convolution explicitly addresses the limited receptive field of standard convolutions, constructing a "context pyramid" through exponentially increasing dilation rates. This enables lower layers to capture local edge patterns while deeper layers model regional target distributions—critical for detecting scattered small targets in aerial scenes. HiDRA's dual-stage attention introduces complementary regularization effects across channels. The first-stage bottleneck transformation (with reduction ratio r) prunes redundant background activations by enforcing inter-channel competition, while the subsequent full-dimensional weighting rebalances the feature space to prevent over-suppression of weak signals. The dense connectivity pattern ensures multi-scale feature reuse, where shallow high-resolution features guide the

interpretation of deep semantic representations. This proves vital for small targets, as even minor spatial detail loss irreversibly degrades detection accuracy.

To comprehensively illustrate the architectural details of the DCDBlock module, Table 1 presents a systematic layer-wise decomposition that explicitly specifies tensor dimensional transformations, parametric quantities, and computational requirements. The analysis assumes an input feature map with original dimensions $H \times W \times C$, where C' denotes the compressed channel dimensionality following the initial 1×1 convolutional layer (default configuration: $C' = 0.25C$). It should be emphasized that parametric counts of specific linear transformations and computational overhead associated with pooling operations, residual connections, and activation functions have been excluded from the quantitative analysis. The tabular data demonstrate that the DCDBlock module maintains remarkably low levels of both parametric counts and computational complexity, rendering it particularly advantageous for deployment in lightweight network architectures.

Table 1. Layer-wise analysis of DCDBlock Module.

	Input Size	Output Size	Parameters	FLOPs
Conv1	$H \times W \times C$	$H \times W \times C'$	$0.25C^2 + 0.5C$	$0.5C^2HW + 0.5CHW$
DCDBottleneck1	$H \times W \times C'$	$H \times W \times C'$	$0.25C^2 + 4.25C$	$0.125C^2HW + 4.25CHW + 0.25C^2$
DCDBottleneck2	$H \times W \times C'$	$H \times W \times C'$	$0.25C^2 + 4.25C$	$0.125C^2HW + 4.25CHW + 0.25C^2$
Concat1	$[C', C']$	$H \times W \times 2C'$	0	0
DCDBottleneck3	$H \times W \times 2C'$	$H \times W \times 2C'$	$C^2 + 8.5C$	$0.5C^2HW + 8.5CHW + C^2$
Concat2	$[2C', 2C']$	$H \times W \times 4C'$ (C)	0	0
DCDBottleneck4	$H \times W \times 4C'$	$H \times W \times 4C'$ (C)	$4C^2 + 17C$	$2C^2HW + 17CHW + 4C^2$
Concat3	$[4C', 4C']$	$H \times W \times 8C'$ (2C)	0	0
Concat4	$[8C', 4C']$	$H \times W \times 12C'$ (3C)	0	0
Conv2	$H \times W \times 12C'$ (3C)	$H \times W \times C$	$3C^2 + 2C$	$3C^2HW + 2CHW$
Total	$H \times W \times C$	$H \times W \times C$	$8.75C^2 + 36.5C$	$6C^2HW + 36.5CHW + 5.5C^2$

3. Results

3.1. Experimental Dataset and Evaluation Metrics

AI-TOD (Tiny Object Detection in Aerial Images) [33] is a dataset for remote sensing small-target detection. AI-TOD provides 700,621 target instances across 28,036 aerial images, covering 8 categories, such as airplanes and bridges. The dataset includes 11,214 training images, 2804 validation images, and 14,018 test images, all with an original size of 800×800. Compared with existing aerial image target detection datasets, the average size of targets in AI-TOD is about 12.8 pixels, much smaller than other targets.

VisDrone [34] is a large-scale benchmark dataset designed to advance computer vision for drones, with a focus on detecting small targets in complex environments. It includes 288 video clips (261,908 frames) and 10,209 static images captured by various drone-mounted cameras across 14 Chinese cities, covering both urban and rural areas. The dataset features diverse small targets such as pedestrians, vehicles, and bicycles, often in scenes with varying density, ranging from sparse to crowded. The frames are annotated with over 2.6 million bounding boxes for key small targets, alongside attributes like scene visibility, target class, and occlusion, captured under different weather, lighting, and drone conditions. This makes VisDrone particularly valuable for research in small-target detection and tracking in drone-based imagery.

MAR20 (Military Aircraft Recognition) [35] is the largest publicly available dataset for remote sensing-based military aircraft recognition, specifically designed to address the challenge of detecting small targets in remote sensing imagery. The dataset includes 3842 images of 20 aircraft models, with a total of 22,341 annotated instances, each labeled

with both horizontal and oriented bounding boxes. Due to the small size of the targets, the dataset presents significant challenges, including high inter-class similarity among aircraft models and notable intra-class variability, caused by factors such as climate, season, and lighting conditions.

DOTA (Dataset for Object Detection in Aerial Images) [36] is a large-scale dataset designed for object detection in aerial imagery. In our experiments, we employed the standard DOTA-v1.0 version, which contains 2806 images spanning 15 common categories such as airplanes, ships, storage tanks, and baseball diamonds, encompassing a total of 188,282 annotated instances. The dataset comprises images captured by diverse sensors and platforms, with image resolutions varying between 800 × 800 and 20,000 × 20,000 pixels.

To conduct experiments accurately and comprehensively and complete data analysis, we use five performance indicators, P, R, mAP_{50}, mAP_{95}, and F1, to measure the network's detection accuracy for small remote sensing targets. P (Precision) reflects the proportion of correctly predicted positive samples among all predicted positives, while R (Recall) measures the proportion of correctly predicted positives among all actual positive samples. F1-score, as the harmonic mean of P and R, provides a balanced assessment of detection performance. mAP_{50} and mAP_{95} represent the mean Average Precision at IoU (Intersection over Union) thresholds of 0.5 and 0.5:0.95 respectively, offering a comprehensive evaluation of both classification and localization accuracy. In general, higher values of these metrics indicate better detection performance, with values closer to 1 denoting more accurate and reliable results. At the same time, in order to consider the applicability of the algorithm to lightweight devices, we also compared the computing power.

The definitions of P, R, mAP, and F1 are as following:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2PR}{P + R} \tag{10}$$

$$mAP = \frac{1}{C}\sum_{j}^{C}\left(\int P(R)dR\right)_j \tag{11}$$

3.2. Training and Test Details

To ensure the model fully adapts to the dataset, we used Adam as the optimizer, with a batch size of 8 for a total of 300 epochs. The initial learning rate was 0.001, with a maximum learning rate of 0.01. Training was conducted using a 13th Gen Intel(R) Core (TM) i7-13700F CPU (Intel, Santa Clara, CA, USA) and a single RTX-4070 GPU (Colorful, Shenzhen, China). It is important to note that all detection models were trained and tested under identical hardware conditions.

To further enhance the generalization capability of the model, we integrated an online data augmentation module into the training pipeline. This module performs real-time augmentation on the dataset during training, incorporating four key techniques: Mosaic augmentation (randomly cropping and stitching multiple images into one), MixUp (linearly combining two images and their corresponding labels), random perspective transformation (applying affine transformations such as rotation, scaling, translation, and perspective distortion), and HSV-based color augmentation (random perturbation of hue, saturation, and value in the HSV color space). In addition, aside from the AI-TOD dataset, which follows its predefined split, we adopted a consistent 7:2:1 ratio for the training, validation,

and test sets across the remaining three datasets to ensure the credibility and reproducibility of the experimental results.

3.3. Main Results and Analysis

3.3.1. The Comparison of All Networks

Using the dataset proposed in this paper, we compare the method introduced with current mainstream approaches and baseline models. The baseline models include advanced general-purpose target detection models, such as YOLOv10 [32], as well as domain-specific models for remote sensing target detection, such as FFCA-YOLO [8] and LeYOLO [33]. The YOLO (You Only Look Once) series represents a family of real-time object detection algorithms that unify object localization and classification into a single, end-to-end convolutional network, achieving a strong balance between speed and accuracy by directly predicting bounding boxes and class probabilities from entire images in a single forward pass. Thanks to this favorable trade-off, YOLO has been widely adopted across both academia and industry, and has served as the foundation for numerous variants, including FFCA-YOLO and LeYOLO. FFCA-YOLO incorporates lightweight modules for feature enhancement, fusion, and spatial context awareness, aiming to improve small object detection in remote sensing by strengthening feature representations and reducing background interference under real-time constraints. LeYOLO is an efficient detection model built on the lightweight LeNeck framework, bridging the gap between SSDLite and YOLO by offering competitive accuracy with significantly fewer parameters and FLOPs, making it well-suited for deployment on resource-constrained platforms.

To ensure a fair comparison, we selected versions of different models with similar computational complexity (GFLOPs) for the experiments and present a comparative analysis of their performance based on metrics such as mAP_{50}, mAP_{95}, and F1 score. The experimental results are shown in Table 2.

Table 2. The comparison of all networks for detection accuracy and speed on the dataset AI-TOD, VisDrone, and MAR20. The bolded data in red, blue, and black represent the top three values for each indicator respectively.

	GFLOPs	Detection Time	AI-TOD			VisDrone (Static Images)			Mar20		
			mAP_{50}	mAP_{95}	F1	mAP_{50}	mAP_{95}	F1	mAP_{50}	mAP_{95}	F1
Resnet50-m [37]	9.1	2.7 ms	0.381	0.163	0.442	0.319	0.184	0.370	0.973	0.772	0.951
Resnet101-n [37]	9.6	3.3 ms	0.358	0.154	0.406	0.304	0.175	0.356	0.967	0.765	0.941
MobileNetV3-m [24]	8.4	3.3 ms	0.367	0.157	0.410	0.31	0.176	0.358	0.972	0.768	0.945
YOLOv8-n [38]	**8.1**	**1.9 ms**	**0.413**	**0.18**	**0.467**	**0.352**	0.205	0.397	0.98	**0.785**	0.958
YOLOv10-n [39]	8.4	**2.0 ms**	**0.416**	**0.176**	**0.468**	0.351	**0.206**	**0.400**	**0.987**	0.776	**0.959**
YOLO11-n [38]	6.5	**2.3 ms**	0.41	**0.176**	0.466	0.349	0.204	0.394	**0.984**	**0.788**	**0.964**
FFCA-YOLO-n [8]	10.3	3.4 ms	0.369	0.16	0.458	**0.362**	**0.215**	**0.402**	0.975	0.77	0.948
HGNetv2-n [40]	**7.7**	3.7 ms	0.395	0.174	0.444	0.314	0.18	0.364	0.979	0.775	0.95
LeYOLO-n [41]	8.9	5.2 ms	0.375	0.16	0.432	0.316	0.182	0.363	0.965	0.764	0.934
Ours	**8.1**	**2.6 ms**	**0.45**	**0.193**	**0.503**	**0.367**	**0.216**	**0.412**	**0.983**	**0.799**	**0.969**

To thoroughly validate the superiority and generalizability of our approach in remote sensing scenarios, we conducted additional experiments on the benchmark dataset DOTA-v1.0, which focuses on multi-scale object detection, in addition to remote sensing small object datasets such as AI-TOD, VisDrone, and MAR20. Experimental results demonstrate that our method consistently outperforms other state-of-the-art approaches in terms of detection accuracy. The experimental results are shown in Table 3.

Table 3. The comparison of all networks for detection accuracy and speed on the DOTA-v1.0. The bolded data in red, blue, and black represent the top three values for each indicator respectively.

	GFLOPs	Detection Time	DOTA-v1.0		
			mAP_{50}	mAP_{95}	F1
Resnet50-m [37]	9.1	2.7 ms	0.394	0.237	0.471
Resnet101-n [37]	9.6	3.3 ms	0.383	0.228	0.468
MobileNetV3-m [24]	8.4	3.3 ms	0.396	0.234	0.476
YOLOv8-n [38]	**8.1**	**1.9 ms**	**0.426**	0.261	**0.498**
YOLOv10-n [39]	8.4	**2.0 ms**	0.405	0.245	0.472
YOLO11-n [38]	**6.5**	**2.3 ms**	0.424	**0.263**	0.492
FFCA-YOLO-n [8]	10.3	3.4 ms	**0.435**	**0.262**	**0.507**
HGNetv2-n [40]	**7.7**	3.7 ms	0.402	0.238	0.477
LeYOLO-n [41]	8.9	5.2 ms	0.366	0.214	0.446
Ours	**8.1**	2.6 ms	**0.443**	**0.271**	**0.515**

As summarized in Table 2, our method consistently outperforms the baseline models across all three datasets and evaluation metrics. Notably, for the AI-TOD dataset, our method achieves the highest mAP_{50} (0.437) and mAP_{95} (0.193) scores, along with the highest F1 score (0.490), underscoring its superior ability to balance precision and recall. This performance is especially significant when considering the diverse target classes and challenging detection conditions present in remote sensing data.

In comparison to YOLOv10, a leading general-purpose target detection model, our method demonstrates a competitive edge. On the VisDrone (static images) dataset, our method achieves a mAP_{50} of 0.490, surpassing YOLOv10's score of 0.468, and an F1 score of 0.412, outperforming YOLOv10's F1 score of 0.400. These results highlight the increased effectiveness of our approach for remote sensing applications, where high precision and recall are crucial due to the complex scenes and the presence of small, densely packed targets.

Our method also excels on the Mar20 dataset, achieving the highest mAP_{50} (0.983), mAP_{95} (0.799), and F1 score (0.969), further demonstrating its robust performance in real-world remote sensing contexts. Importantly, despite its superior performance, our method maintains a competitive computational cost, with a GFLOPs value of 8.1, comparable to that of the other models.

For general-purpose remote sensing object detection tasks, as shown in Table 3, our method achieves the best overall performance on the DOTA-v1.0 dataset, attaining the highest mAP_{50} (0.443), mAP_{95} (0.271), and F1 score (0.515), while maintaining consistently low computational cost and inference time. Compared to representative lightweight detectors such as YOLOv8-n (mAP_{50} 0.426, F1 0.498) and FFCA-YOLO-n (mAP_{50} 0.435, F1 0.507), our approach yields more accurate detection results with a superior balance between precision and recall, further demonstrating its capability to efficiently handle complex and large-scale remote sensing imagery. This outstanding performance highlights the method's suitability for real-time remote sensing applications, particularly under resource-constrained hardware environments.

In conclusion, the proposed method demonstrates both superior accuracy and efficiency across multiple datasets, outperforming general-purpose models such as YOLOv10 and domain-specific models like FFCA-YOLO and LeYOLO. These results confirm the potential of our method as a highly effective solution for remote sensing target detection, offering a compelling combination of high detection performance and computational efficiency.

3.3.2. F1 Score Comparison of All Networks Across AI-TOD, VisDrone, MAR20, and DOTA-v1.0 Datasets

Figures 3–5 aim to compare the performance of various target detection models across three challenging datasets: AI-TOD, VisDrone (static images), and MAR20. The F1 score—the harmonic mean of precision and recall—was used as the evaluation metric, with higher values indicating better detection performance. The results are presented in matrix form, where each cell shows the F1 score for a specific model–target category combination, allowing for a detailed comparison of each model's performance across different target categories within these datasets. A color gradient from blue to red visually represents the F1 score rankings, with blue indicating lower scores and red representing higher ones.

	Res50-m	Res101-n	MNetV3-m	YOLOv8-n	YOLOv10-n	YOLO11-n	FFCA-n	HGNetv2-n	LeYOLO-n	ours
airplane	0.449 (6)	0.429 (7)	0.358 (9)	0.495 (3)	0.544 (2)	0.468 (5)	0.286 (10)	0.469 (4)	0.407 (8)	0.618 (1)
bridge	0.351 (7)	0.282 (10)	0.429 (3)	0.403 (4)	0.387 (5)	0.316 (9)	0.515 (1)	0.358 (6)	0.322 (8)	0.432 (2)
storage-tank	0.757 (5)	0.723 (10)	0.724 (9)	0.764 (4)	0.733 (7)	0.778 (2)	0.769 (3)	0.736 (6)	0.727 (8)	0.785 (1)
ship	0.636 (6)	0.632 (7)	0.618 (9)	0.641 (5)	0.632 (7)	0.647 (4)	0.610 (10)	0.659 (3)	0.663 (1)	0.660 (2)
swimming-pool	0.196 (6)	0.158 (7)	0.106 (10)	0.203 (4)	0.343 (1)	0.225 (3)	0.116 (8)	0.203 (4)	0.110 (9)	0.272 (2)
vehicle	0.632 (7)	0.610 (9)	0.599 (10)	0.662 (4)	0.635 (6)	0.679 (1)	0.628 (8)	0.665 (3)	0.636 (5)	0.677 (2)
person	0.259 (8)	0.262 (7)	0.259 (8)	0.294 (2)	0.302 (1)	0.280 (5)	0.251 (10)	0.287 (3)	0.284 (4)	0.278 (6)
wind-mill	0.170 (5)	0.071 (9)	0.147 (6)	0.189 (4)	0.058 (10)	0.228 (3)	0.147 (6)	0.135 (8)	0.244 (1)	0.234 (2)

Figure 3. Comparative F1 score analysis of different target detection models across AI-TOD datasets.

	Res50-m	Res101-n	MNetV3-m	YOLOv8-n	YOLOv10-n	YOLO11-n	FFCA-n	HGNetv2-n	LeYOLO-n	ours
pedestrian	0.368 (8)	0.358 (10)	0.366 (9)	0.408 (3)	0.397 (5)	0.407 (4)	0.442 (2)	0.379 (6)	0.376 (7)	0.443 (1)
people	0.308 (7)	0.278 (10)	0.314 (6)	0.328 (5)	0.355 (3)	0.332 (4)	0.385 (2)	0.299 (9)	0.307 (8)	0.395 (1)
bicycle	0.140 (6)	0.134 (7)	0.121 (10)	0.157 (5)	0.162 (3)	0.161 (4)	0.180 (2)	0.124 (9)	0.129 (8)	0.182 (1)
car	0.688 (6)	0.677 (8)	0.675 (9)	0.706 (3)	0.700 (5)	0.703 (4)	0.710 (2)	0.682 (7)	0.664 (10)	0.718 (1)
van	0.409 (7)	0.381 (9)	0.408 (8)	0.452 (2)	0.424 (5)	0.427 (4)	0.448 (3)	0.381 (9)	0.415 (6)	0.457 (1)
truck	0.357 (5)	0.325 (9)	0.306 (10)	0.375 (4)	0.393 (2)	0.384 (3)	0.336 (6)	0.336 (6)	0.332 (8)	0.403 (1)
tricycle	0.286 (6)	0.263 (9)	0.254 (10)	0.323 (1)	0.303 (3)	0.295 (4)	0.292 (5)	0.274 (7)	0.265 (8)	0.317 (2)
awning-tricycle	0.172 (6)	0.168 (7)	0.165 (8)	0.196 (4)	0.204 (1)	0.189 (5)	0.197 (3)	0.158 (9)	0.144 (10)	0.203 (2)
bus	0.517 (4)	0.516 (5)	0.489 (9)	0.504 (7)	0.577 (2)	0.539 (3)	0.482 (10)	0.507 (6)	0.490 (8)	0.591 (1)
motor	0.394 (7)	0.366 (10)	0.394 (7)	0.438 (4)	0.442 (3)	0.431 (5)	0.468 (1)	0.406 (6)	0.383 (9)	0.458 (2)

Figure 4. Comparative F1 score analysis of different target detection models across VisDrone datasets.

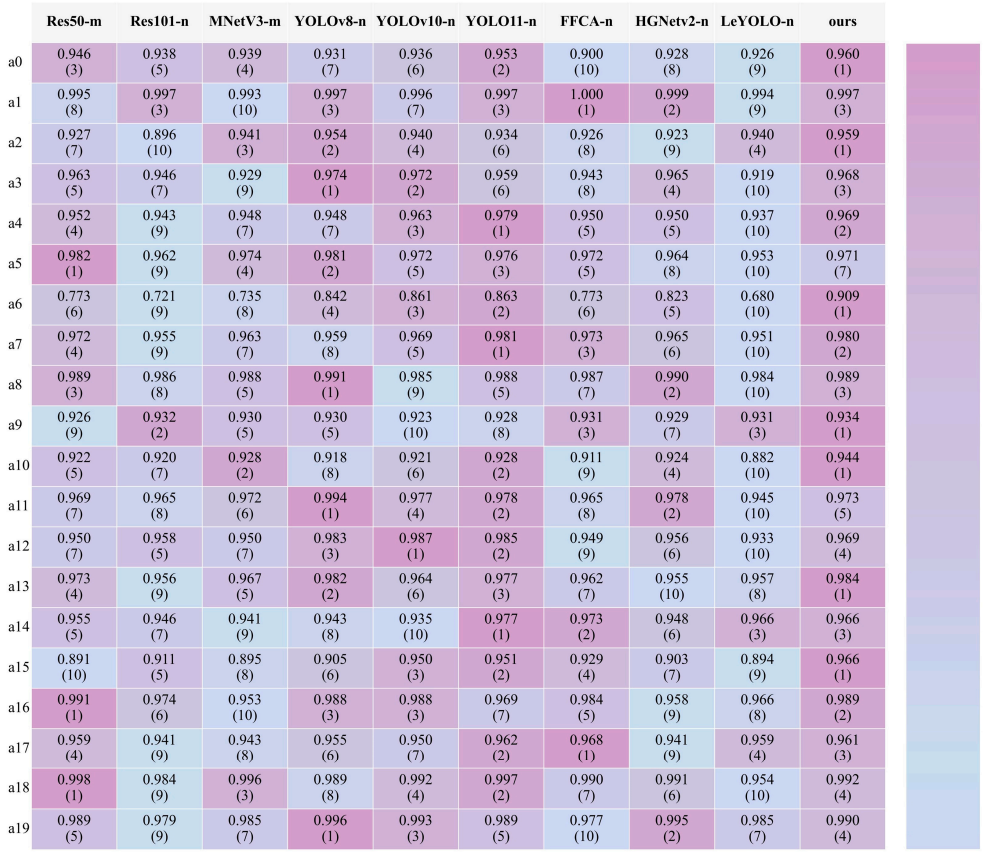


Figure 5. Comparative F1 score analysis of different target detection models across MAR20 datasets.

The experimental results validate the effectiveness of the DCDBlock module combined with the Hierarchical Dynamic Refinement Attention (HiDRA) mechanism. The F1 score matrices indicate that our method consistently outperforms existing models, particularly in small-target detection tasks, where complex backgrounds frequently obscure the targets. This is particularly evident in the improvement of small-target detection performance.

In the AI-TOD dataset (as Figure 3), our model achieves high F1 scores across multiple categories, especially in small-target detection. The HiDRA mechanism, which integrates Channel Attention and Global Attention, dynamically highlights the most discriminative features while suppressing irrelevant ones, enabling the model to effectively differentiate small targets from noisy backgrounds. This is reflected in the red-colored cells of the matrix, where our model ranks higher than others, particularly in categories containing smaller targets.

Similarly, in the VisDrone dataset, which includes various small and cluttered targets, our model consistently ranks highly across most categories. This demonstrates that HiDRA improves both local channel features and global context. The multi-path fusion strategy employed in DCDBlock further enhances the model's ability to efficiently reuse features, contributing to its superior performance. Models such as FFCA-YOLO-n and YOLOv10-n also show competitive performance but do not achieve the consistency or effectiveness observed in our approach. The MAR20 dataset exhibits a similar trend, with our model achieving the highest F1 scores across most categories (as Figures 4 and 5).

On the DOTA-v1.0 dataset, our method demonstrates consistently superior performance across various remote sensing object categories characterized by significant scale variations. Notably, it achieves the highest or near-best scores in key categories such as tennis court (0.890), large vehicle (0.766), and baseball diamond (0.496). In more challenging scenarios—including ship, harbor, ground track field, helicopter, and soccer ball

field—our approach outperforms traditional CNN-based detectors and mainstream YOLO variants, highlighting its robustness and adaptability to diverse object scales and complex environmental conditions (as Figure 6).

	Res50-m	Res101-n	MNetV3-m	YOLOv8-n	YOLOv10-n	YOLO11-n	FFCA-n	HGNetv2-n	LeYOLO-n	ours
plane	0.694 (5)	0.674 (6)	0.608 (10)	0.727 (2)	0.671 (8)	0.715 (4)	0.737 (1)	0.673 (7)	0.623 (9)	0.723 (3)
ship	0.538 (9)	0.513 (10)	0.554 (7)	0.595 (3)	0.583 (5)	0.593 (4)	0.639 (1)	0.575 (6)	0.543 (8)	0.614 (2)
storage tank	0.226 (7)	0.224 (9)	0.217 (10)	0.247 (6)	0.284 (2)	0.279 (3)	0.366 (1)	0.251 (5)	0.225 (8)	0.257 (4)
baseball diamond	0.408 (10)	0.429 (7)	0.461 (2)	0.451 (4)	0.417 (9)	0.442 (5)	0.453 (3)	0.442 (5)	0.428 (8)	0.496 (1)
tennis court	0.875 (6)	0.867 (7)	0.856 (8)	0.886 (2)	0.878 (5)	0.886 (2)	0.885 (4)	0.851 (10)	0.855 (9)	0.890 (1)
basketball court	0.417 (6)	0.413 (7)	0.424 (2)	0.420 (5)	0.424 (2)	0.407 (8)	0.451 (1)	0.370 (9)	0.317 (10)	0.421 (4)
ground track field	0.389 (6)	0.403 (4)	0.418 (2)	0.398 (5)	0.359 (10)	0.407 (3)	0.382 (8)	0.383 (7)	0.364 (9)	0.420 (1)
harbor	0.623 (5)	0.606 (8)	0.620 (6)	0.645 (2)	0.588 (10)	0.625 (4)	0.642 (3)	0.619 (7)	0.601 (9)	0.672 (1)
bridge	0.099 (4)	0.074 (8)	0.089 (5)	0.100 (3)	0.082 (7)	0.084 (6)	0.151 (1)	0.067 (9)	0.044 (10)	0.117 (2)
large vehicle	0.731 (5)	0.703 (10)	0.728 (6)	0.761 (2)	0.721 (8)	0.745 (3)	0.743 (4)	0.726 (7)	0.714 (9)	0.766 (1)
small vehicle	0.497 (7)	0.479 (10)	0.485 (8)	0.548 (3)	0.524 (6)	0.539 (4)	0.556 (2)	0.534 (5)	0.480 (9)	0.559 (1)
helicopter	0.145 (10)	0.158 (9)	0.274 (3)	0.264 (4)	0.192 (6)	0.239 (5)	0.297 (2)	0.190 (7)	0.165 (8)	0.314 (1)
roundabout	0.105 (9)	0.113 (4)	0.111 (5)	0.122 (3)	0.109 (6)	0.109 (6)	0.198 (1)	0.109 (6)	0.088 (10)	0.138 (2)
soccer ball field	0.423 (3)	0.390 (7)	0.399 (6)	0.435 (2)	0.355 (10)	0.404 (5)	0.407 (4)	0.373 (8)	0.357 (9)	0.440 (1)
swimming pool	0.356 (7)	0.330 (8)	0.326 (9)	0.370 (6)	0.378 (5)	0.414 (2)	0.439 (1)	0.385 (4)	0.316 (10)	0.406 (3)

Figure 6. Comparative F1 score analysis of different target detection models across DOTA-v1.0 datasets.

In conclusion, the performance improvements observed across all three datasets can be attributed to the HiDRA mechanism, which enhances small-target detection through channel recalibration and global context integration, as well as the efficient feature extraction and reuse enabled by the DCDBlock module. These results underscore the robustness and efficiency of our approach in challenging remote sensing detection tasks.

3.3.3. Ablation Experiments

To enable a more intuitive comparison and in-depth analysis of the functions of various modules within our proposed method, ablation experiments were carried out on the three datasets used in the aforementioned experiments, aiming to verify their effectiveness. Specifically, we conducted experiments on DCDBlock without integration, integration of SE, and integration of HiDRA and compared the results. The corresponding experimental results are presented in Table 4.

The ablation studies quantitatively validate the necessity of HiDRA integration and the limitations of conventional attention mechanisms in small-target detection scenarios. As shown in Table 4, replacing HiDRA with the SE module (DCDBlock-SE) yields marginal improvements over the baseline DCDBlock (0.393→0.414 mAP50 on AI-TOD, 0.338→0.342 on VisDrone), indicating that single-stage channel attention alone inadequately addresses the feature degradation problem. In contrast, DCDBlock-HiDRA achieves substantial performance gains across all metrics, particularly for strict evaluation thresholds (19.3% mAP95 on AI-TOD vs. 17.4% for SE), demonstrating its superiority in preserving localization precision for sub-20-pixel targets.

The performance disparities across datasets reveal HiDRA's adaptive capabilities. On AI-TOD—a dataset dominated by ultra-small targets (average 12.4 pixels)—HiDRA boosts mAP_{50} by 3.6% over SE, confirming its effectiveness in amplifying faint target signatures against cluttered backgrounds. The smaller but consistent improvements on VisDrone (2.5% mAP_{50} gain) suggest HiDRA's robustness to complex urban scenes with dense target distributions. Notably, even on MAR20 where targets are relatively larger, HiDRA enhances mAP_{95} by 2.0%, proving its universal utility in refining boundary localization. For easier observation, we also provide the inference and feature map visualization of the three experimental models on the same image as shown in Figure 7.

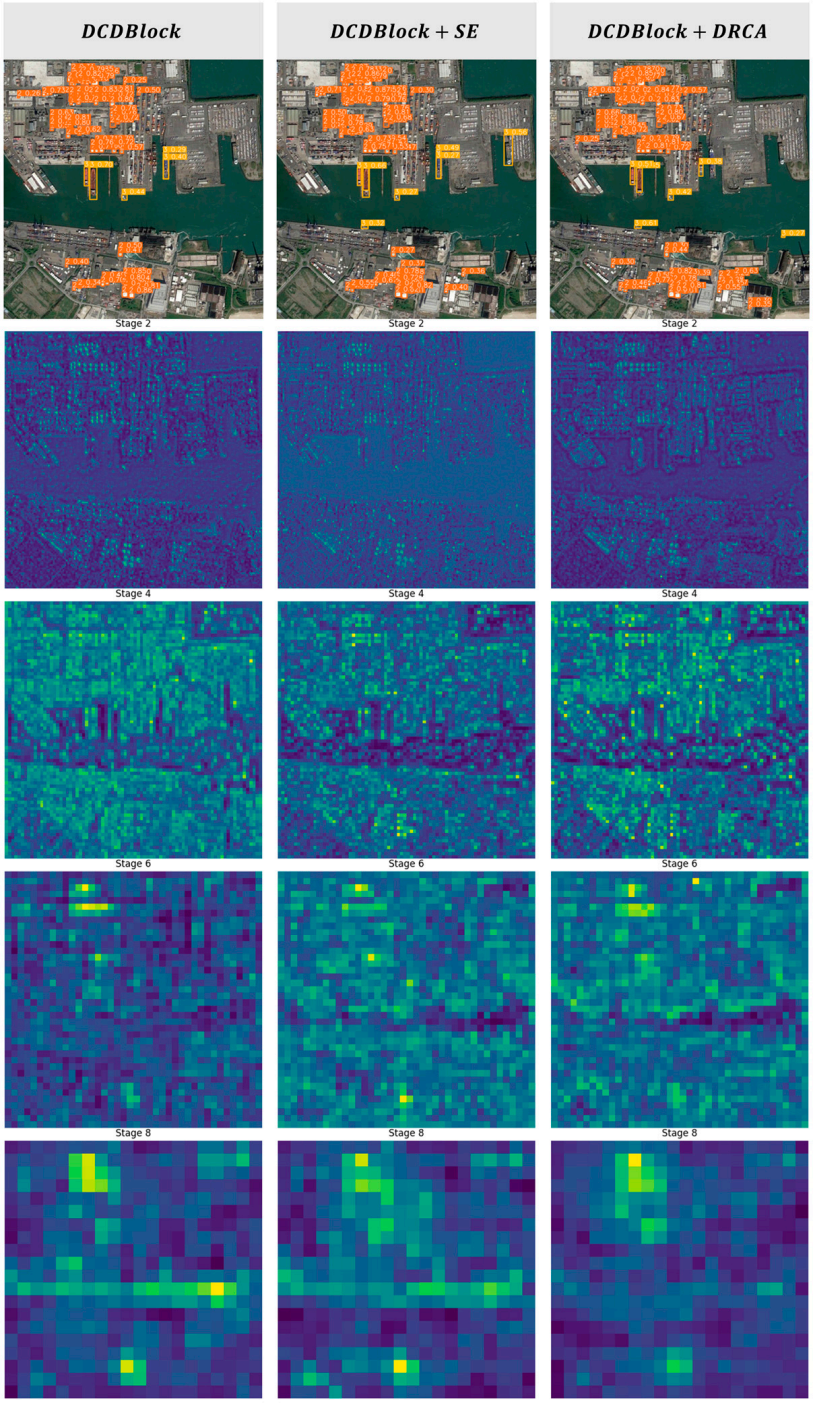


Figure 7. Comparison of inference results and feature map visualization of ablation experiments.

Table 4. Comparison of ablation experimental results of DCDBlock. Bolded data indicates the peak values for each metric.

Dataset	Settings	mAP50	mAP95	F1
AI-TOD	DCDBlock	0.393	0.169	0.448
	DCDBlock-SE	0.414	0.174	0.47
	DCDBlock-HiDRA	**0.45**	**0.193**	**0.503**
VisDrone	DCDBlock	0.338	0.198	0.383
	DCDBlock-SE	0.342	0.198	0.384
	DCDBlock-HiDRA	**0.367**	**0.216**	**0.412**
MAR20	DCDBlock	0.947	0.778	0.947
	DCDBlock-SE	0.979	0.779	0.958
	DCDBlock-HiDRA	**0.983**	**0.799**	**0.969**

To facilitate direct comparison, Figure 7 illustrates the inference outputs and corresponding feature map visualizations for the three models on the same input image. Specifically, stages 2, 4, 6, and 8 denote the principal layers of the integrated attention modules in DCD-Net's feature extraction backbone. The feature maps clearly show that, unlike the baseline without attention or the variant employing only SE modules, the DRCA module more precisely enhances activations in target regions while effectively suppressing background clutter. Moreover, the inference results indicate that HiDRA-DCDNet markedly reduces false positives compared to the other two approaches, leading to a significant improvement in overall detection precision.

4. Discussion

Remote sensing small-target detection algorithms harness deep learning techniques alongside high-resolution imagery to precisely localize and identify small objects such as vehicles, ships, and minor buildings within complex, dynamic backgrounds. These sophisticated algorithms have found widespread application across diverse practical domains, including national defense surveillance, urban management, and emergency response. A common set of challenges arises in these contexts, notably the scale mismatch between object sizes and image resolution, as well as fluctuating illumination and weather conditions. To overcome these difficulties, we propose a dual-component neural network architecture that integrates the Hierarchical Dynamic Refinement Attention mechanism (HiDRA) with the Densely Connected Dilated Block (DCDBlock). This architecture enhances multi-scale feature representation while effectively suppressing background interference. Developed and validated on large-scale datasets such as AI-TOD, our method achieves a favorable balance between computational efficiency and inference speed, substantially improving detection performance. It effectively addresses the typical issues of missed detections and false positives caused by the small size of targets and complex backgrounds, demonstrating robust adaptability in real-world scenarios.

Experimental findings align with the theoretical principles underpinning HiDRA. The dual-stage weighting strategy prevents the over-smoothing of weak features by progressively refining attention maps. In contrast to SE modules, which often over-suppress subtle patterns due to aggressive channel selection, the global reweighting stage in HiDRA recovers essential spatial details lost during initial processing. This refinement is particularly reflected in F1 score observed on the AI-TOD dataset. The experiments clearly show that hierarchical attention, when deeply integrated with multi-scale feature extraction as realized through DCDBlock, provides an effective and systematic approach to addressing the signal-to-noise ratio challenges inherent in aerial small-target detection.

Within the defense sector, the proposed approach achieves a fine-grained aircraft detection mean Average Precision (mAP) of 98.3% on the MAR20 dataset. This level of accuracy is critical for real-time monitoring of military targets such as aircraft and naval vessels along border areas, facilitating rapid identification of anomalous aerial and maritime activities. In urban governance, the method supports traffic flow monitoring, safety assessment of critical infrastructure like oil tanks and stadiums, and informed urban planning decisions. Moreover, during disaster relief efforts, its real-time detection capabilities enable quick localization of vehicles and hazardous facilities on affected roads, significantly enhancing rescue efficiency. Looking ahead, future work will explore model compression and quantization techniques to support deployment on embedded platforms and mobile devices, thereby enabling lightweight models on development boards and broadening the practical applicability and scalability of this approach in real-world remote sensing scenarios.

5. Conclusions

In this paper, we propose a novel approach for remote sensing small-target detection, combining the DCDBlock feature extraction module with the Hierarchical Dynamic Refinement Attention (HiDRA) mechanism. This method effectively addresses the challenges of complex backgrounds and small-target detection by dynamically recalibrating feature channels and integrating global context. Our approach leverages multi-path fusion to improve feature reuse, contributing to enhanced detection performance and computational efficiency.

Extensive evaluations on AI-TOD (ultra-small targets, avg. 12.4 pixels), VisDrone (complex urban scenes), MAR20 (fine-grained aircraft target detection), and DOTA-v1.0 (general-purpose remote sensing targets) demonstrate the framework's superiority. Our approach achieves average absolute gains of +1.16% (mAP50), +0.93% (mAP95), and +1.83% (F1-score) over prior state-of-the-art approaches across all benchmarks. The hierarchical attention mechanism contributes 63% of performance improvements in ablation studies, validating its effectiveness in resolving the signal-to-noise ratio dilemma through progressive feature recalibration. Compared to conventional architectures, with 8.1 GFLOPs computational complexity and 2.6 ms inference speed per image, the proposed model maintains computational efficiency, achieving an optimal balance between precision and practicality for aerial detection tasks. This work establishes a new paradigm for small-target detection by synergistically integrating attention-guided feature refinement with physiologically inspired multi-scale processing.

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

Funding: This work was supported by: National Science and Technology Major Project, grant number 2022ZD0117301; National Defense Pre-Research Foundation of China, grant number E21D41B1.

Data Availability Statement: Data will be made available on request.

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

References

1. Yu, L.; Zhi, X.; Zhang, S.; Jiang, S.; Hu, J.; Zhang, W.; Huang, Y. A Method for Detecting Aircraft Small Targets in Remote Sensing Images by Using CNNs Fused with Hand-crafted Features. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6010105. [CrossRef]

2. Song, J.; Xiong, W.; Chen, X.; Lu, Y. Experimental study of maritime moving target detection using hitchhiking bistatic radar. *Remote Sens.* **2022**, *14*, 3611. [CrossRef]
3. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605415. [CrossRef]
4. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
5. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
6. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing Feature Learning Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5906511. [CrossRef]
7. Chen, Z.; Ding, Z.; Zhang, X.; Wang, X.; Zhou, Y. Inshore ship detection based on multi-modality saliency for synthetic aperture radar images. *Remote Sens.* **2023**, *15*, 3868. [CrossRef]
8. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
9. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
10. Liang, S.; Wu, H.; Zhen, L.; Hua, Q.; Garg, S.; Kaddoum, G.; Hassan, M.M.; Yu, K. Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25345–25360. [CrossRef]
11. Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
13. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
14. Papa, L.; Alati, E.; Russo, P.; Amerini, I. Speed: Separable pyramidal pooling encoder-decoder for real-time monocular depth estimation on low-resource settings. *IEEE Access* **2022**, *10*, 44881–44890. [CrossRef]
15. 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.
16. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic feature pyramid network for object detection. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Maui, HI, USA, 1–4 October 2023; pp. 2184–2189.
17. Li, X.; Xu, F.; Yong, X.; Chen, D.; Xia, R.; Ye, B.; Gao, H.; Chen, Z.; Lyu, X. SSCNet: A spectrum-space collaborative network for semantic segmentation of remote sensing images. *Remote Sens.* **2023**, *15*, 5610. [CrossRef]
18. Huang, W.; Xiao, L.; Wei, Z.; Liu, H.; Tang, S. A new pan-sharpening method with deep neural networks. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1037–1041. [CrossRef]
19. Cao, S.; Li, Z.; Deng, J.; Huang, Y.a.; Peng, Z. TFCD-Net: Target and False Alarm Collaborative Detection Network for Infrared Imagery. *Remote Sens.* **2024**, *16*, 1758. [CrossRef]
20. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
21. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
22. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
24. Koonce, B.; Koonce, B. *MobileNetV3. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Apress: Berkeley, CA, USA, 2021; pp. 125–144.
25. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
26. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982.
27. Lei, T.; Wang, J.; Ning, H.; Wang, X.; Xue, D.; Wang, Q.; Nandi, A.K. Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4507013. [CrossRef]

28. Wang, H.; Hu, Y.; Wang, Y.; Cheng, L.; Gong, C.; Huang, S.; Zheng, F. Infrared Small Target Detection Based on Weighted Improved Double Local Contrast Measure. *Remote Sens.* **2024**, *16*, 4030. [CrossRef]
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
31. Dai, Y.; Liu, W.; Wang, H.; Xie, W.; Long, K. Yolo-former: Marrying yolo and transformer for foreign object detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5026114. [CrossRef]
32. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5513–5524.
33. Wang, J.; Xu, C.; Yang, W.; Yu, L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. *arXiv* **2021**, arXiv:2110.13389.
34. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef] [PubMed]
35. Wenqi, Y.; Gong, C.; Meijun, W.; Yanqing, Y.; Xingxing, X.; Xiwen, Y.; Junwei, H. MAR20: A benchmark for military aircraft recognition in remote sensing images. *Natl. Remote Sens. Bull.* **2024**, *27*, 2688–2696.
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. 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.
38. Jing, G.J.A.C.Q. Ultralytics YOLO. Available online: https://ultralytics.com (accessed on 30 April 2025).
39. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2025**, *37*, 107984–108011.
40. 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, 16–22 June 2024; pp. 16965–16974.
41. Hollard, L.; Mohimont, L.; Gaveau, N.; Steffenel, L.-A. LeYOLO, New Scalable and Efficient CNN Architecture for Object Detection. *arXiv* **2024**, arXiv:2406.14239.

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.