Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 13 June 2026
Revised 25 July 2026
Accepted 08 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (15s): 498–518

KEYWORDS:
Tiny object detection; Aerial image analysis; Feature Pyramid Network; Coordinate Attention; Faster R-CNN; ResNet-18; Remote sensing; Lightweight object detection.

LCANet: A Lightweight Faster R-CNN Framework with Selective P_2 Coordinate Attention for Tiny Object Detection in Aerial Images

Mangamma Choppara[1], Kunjam Nageswara Rao[2]

[1]Department of Computer Science and System Engineering, Andhra University, Vizag, India
mangachoppara@gmail.com

[2]Department of Computer Science and System Engineering, Andhra University, Vizag, India
kunjamnag@gmail.com

ABSTRACT: Accurate detection of tiny objects in aerial images remains a challenging problem in remote sensing due to extreme scale variations, complex backgrounds, dense object distributions, and the limited spatial representation of small targets. Although modern deep learning-based object detectors achieve remarkable detection accuracy, many rely on computationally intensive backbone networks and complex attention mechanisms that restrict their deployment on resource-constrained platforms such as unmanned aerial vehicles (UAVs) and edge devices. Lightweight detection frameworks reduce computational complexity but often suffer from degraded localization performance for tiny objects because high-resolution spatial features are gradually lost during feature extraction.

To address this limitation, this paper proposes a lightweight object detection framework that selectively integrates Coordinate Attention (CA) into the highest-resolution feature pyramid level (P_2) of a Faster R-CNN architecture employing a ResNet-18 backbone and a Feature Pyramid Network (FPN). Unlike conventional approaches that introduce attention modules across multiple feature levels, the proposed method enhances only the P_2 feature map, where the fine spatial detail required for tiny-object localization is best preserved, by factoring channel attention into two 1D feature-encoding vectors along the horizontal and vertical directions. This selective integration strengthens directional and positional feature representation while preserving the computational efficiency of the lightweight detector, and it leaves the Region Proposal Network (RPN), RoIAlign, and Faster R-CNN detection head entirely unchanged.

The proposed framework is evaluated on the DOTA v2.0 aerial-image dataset using a structured patch-based partition of 8,734 training patches and 1,000 validation patches under Task 2 (horizontal bounding box) settings. The baseline ResNet-18-FPN Faster R-CNN model converges smoothly over a 10-epoch schedule, reaching a final training loss of 0.5775 and a validation mAP@50 of 73.71%. Integrating the proposed P_2-selective Coordinate Attention module reduces the final training loss to 0.3782 and raises the validation mAP@50 to 86.20%, an absolute improvement of 12.49 percentage points, while training and validation loss curves demonstrate faster convergence and improved optimization stability throughout training. Qualitative evaluation on high-density and

ultra-dense test scenes further confirms that the proposed model recovers small, densely packed targets missed by the baseline.

These findings indicate that selectively enhancing the highest-resolution feature representation is an effective and computationally efficient strategy for improving tiny-object localization without modifying the Faster R-CNN detection pipeline. The proposed framework therefore provides a practical solution for lightweight aerial object detection in real-world remote sensing and UAV applications.

1. INTRODUCTION

Object detection in aerial imagery has become an essential research area in computer vision and remote sensing due to its wide range of applications, including intelligent transportation systems, urban planning, disaster management, environmental monitoring, military reconnaissance, precision agriculture, maritime surveillance, and autonomous unmanned aerial vehicle (UAV) navigation. Unlike conventional natural-image datasets, aerial images are acquired from high altitudes using satellites, drones, or airborne sensors, resulting in unique imaging characteristics that significantly increase the complexity of object detection.

Among the numerous challenges encountered in aerial image analysis, tiny object detection remains one of the most difficult problems. Objects such as vehicles, ships, storage tanks, and small buildings often occupy only a few pixels within extremely high-resolution images. As convolutional neural networks progressively down sample feature maps through multiple convolutional layers, much of the fine spatial information required to localize these tiny objects is gradually lost. Consequently, lightweight object detectors frequently miss small targets or produce inaccurate localization, particularly in scenes containing dense object distributions and cluttered backgrounds.

In addition to their small size, aerial objects exhibit substantial scale variation, arbitrary orientations, complex background textures, partial occlusions, and uneven spatial distributions. A single aerial image may simultaneously contain large structures such as airports and buildings together with tiny vehicles occupying only a negligible portion of the image. Moreover, objects are frequently surrounded by visually similar background patterns such as roads, parking areas, vegetation, rooftops, or shadows, making discrimination considerably more challenging. These characteristics reduce the effectiveness of conventional object detection algorithms originally developed for natural-image datasets.

Recent advances in deep learning have significantly improved object detection performance through two-stage and one-stage detection frameworks. Two-stage detectors, particularly Faster R-CNN [1], have demonstrated superior localization accuracy because the Region Proposal Network (RPN) effectively identifies candidate object regions before classification and bounding-box regression. Furthermore, the introduction of Feature Pyramid Networks (FPN) [2] has substantially improved multi-scale feature representation by combining low-level spatial information with high-level semantic information across multiple pyramid levels. Consequently, FPN has become one of the fundamental components in modern aerial object detection systems.

Although deeper backbone networks such as ResNet-50, ResNet-101, and ResNeXt achieve excellent detection accuracy, their computational complexity and memory requirements limit their applicability to embedded systems and real-time UAV platforms. Lightweight backbone networks, including ResNet-18 [6], provide an attractive alternative by substantially reducing computational cost. However, the reduction in network depth inevitably weakens feature extraction capability, particularly for tiny objects whose discriminative features are already limited. Therefore, maintaining high detection accuracy while preserving computational efficiency remains a major research challenge.

Attention mechanisms have recently emerged as an effective strategy for improving feature representation in convolutional neural networks. Modules such as Squeeze-and-Excitation (SE) [12], Convolutional Block Attention Module (CBAM) [13], and Coordinate Attention (CA) [14] enable networks to emphasize informative features while suppressing irrelevant background information. Among these approaches, Coordinate Attention is particularly suitable for object localization because it encodes both channel dependencies and long-range positional information through separate horizontal and vertical feature aggregation. This characteristic enables the network to preserve spatial awareness while introducing only a modest computational overhead.

Despite the success of attention mechanisms, most existing studies apply attention modules throughout the backbone network or across multiple feature pyramid levels. Although such strategies generally improve feature representation, they also increase computational complexity and parameter count. More importantly, they do not explicitly exploit the fact that the P_2 feature map within the Feature Pyramid Network preserves the highest spatial resolution, making it particularly suitable for representing tiny objects. The potential benefits of selectively enhancing only the highest-resolution feature level for lightweight aerial object detection remain insufficiently explored.

Motivated by this observation, this paper proposes a lightweight multi-scale object detection framework with Selective P_2 Coordinate Attention. Instead of introducing attention throughout the network, the proposed approach integrates a single Coordinate Attention module exclusively into the P_2 feature map of a lightweight Faster R-CNN detector employing a ResNet-18 backbone and Feature Pyramid Network. Because P_2 retains rich spatial detail while benefiting from multi-scale feature fusion, selectively enhancing this feature level enables more accurate localization of tiny aerial objects without modifying the Region Proposal Network, RoIAlign operation, or detection head. Consequently, the proposed framework maintains the computational efficiency of the lightweight architecture while substantially improving tiny-object detection performance.

The proposed method is evaluated using the challenging DOTA v2.0 aerial-image dataset [15] under identical experimental settings for both the baseline and proposed models. Experimental results demonstrate that selective P_2 Coordinate Attention consistently improves optimization stability and detection accuracy, increasing the validation mAP@50 from 73.71% to 86.20%, while preserving the simplicity of the original Faster R-CNN detection framework.

The major contributions of this work are summarized as follows:

- A lightweight multi-scale aerial object detection framework is developed by combining a ResNet-18 backbone with a Feature Pyramid Network to achieve an effective balance between computational efficiency and detection accuracy.
- A selective Coordinate Attention strategy is proposed by integrating a single Coordinate Attention module exclusively at the highest-resolution P_2 feature map, enabling enhanced spatial feature representation for tiny-object localization without modifying the Faster R-CNN detection pipeline.
- The Region Proposal Network, RoIAlign, and Faster R-CNN detection head are kept entirely unchanged, isolating the observed performance gain to the proposed attention module and strengthening the causal interpretability of the result.
- Comprehensive experiments on the DOTA v2.0 dataset demonstrate that the proposed framework significantly improves tiny-object detection performance, achieving an mAP@50 of 86.20%, an absolute improvement of 12.49 percentage points over the lightweight baseline, while maintaining computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work on aerial object detection, lightweight networks, multi-scale feature learning, and attention mechanisms. Section 3 details the proposed methodology. Section 4 describes the experimental setup. Section 5 presents and discusses the results. Section 6 concludes the paper and outlines directions for future work.

2. RELATED WORK

Recent advances in deep learning have significantly improved object detection performance in aerial imagery. However, accurately detecting tiny objects remains a challenging task because of extreme scale variations, complex backgrounds, arbitrary object orientations, and dense object distributions. To address these challenges, researchers have proposed various detection frameworks involving deep convolutional neural networks, lightweight backbone architectures, multi-scale feature learning strategies, and attention mechanisms. This section reviews representative studies in these research directions and identifies the research gap addressed by the proposed framework.

2.1 Deep Learning-Based Aerial Object Detection

The rapid advancement of deep learning has substantially improved object detection performance in aerial imagery by enabling convolutional neural networks to learn discriminative feature representations directly from large-scale datasets. Nevertheless, aerial object detection remains considerably more challenging than object detection in natural scenes because of large intra-class variation, arbitrary object orientation, dense object distribution, complex backgrounds, and significant scale differences.

Among these challenges, accurately detecting tiny objects has received increasing attention, since small targets typically occupy only a few pixels, resulting in insufficient semantic information and weak feature representation during deep feature extraction [1], [15], [16].

The introduction of the faster r-cnn framework represented a major milestone in object detection by integrating a region proposal network (rpn) into an end-to-end trainable architecture, significantly improving localization accuracy and detection performance compared with earlier proposal-based approaches [1]. Building upon this framework, the feature pyramid network (fpn) further enhanced multi-scale object detection by combining high-level semantic information with low-level spatial features through a top-down feature-fusion architecture [2]. Owing to its excellent localization capability, the faster r-cnn–fpn framework has become one of the most widely adopted baselines for aerial image analysis and remote sensing applications involving vehicle detection, ship detection, aircraft detection, and dense object localization [15], [17], [22].

To further improve detection efficiency, several one-stage detectors, including the yolo and ssd families [4], have been proposed for real-time object detection. Recent versions such as yolov7 have achieved remarkable improvements in inference speed while maintaining competitive detection accuracy [3]. However, despite their computational advantages, one-stage detectors often experience performance degradation when detecting tiny objects, because repeated feature down-sampling reduces the discriminative information available for very small targets. Consequently, numerous comparative studies report that two-stage detectors generally provide superior localization accuracy for tiny-object detection, particularly in high-resolution aerial imagery where precise localization is more critical than inference speed [16], [19], [20].

More recently, transformer-based object detection frameworks have attracted considerable interest because of their ability to model long-range feature dependencies using self-attention mechanisms. Detr and its subsequent variants have demonstrated competitive detection performance by eliminating hand-crafted components such as anchor generation and non-maximum suppression [5]. Nevertheless, transformer-based approaches generally require substantial computational resources, longer training schedules, and larger annotated datasets, thereby limiting their applicability in lightweight aerial surveillance systems and resource-constrained uav platforms [19], [20]. Consequently, convolution-based detection frameworks continue to remain attractive for lightweight aerial object detection because they provide an effective balance between computational efficiency and localization accuracy.

Although remarkable progress has been achieved over the past decade, recent surveys consistently conclude that tiny-object detection in aerial imagery remains an open research problem due to insufficient spatial feature representation, information loss during hierarchical feature extraction, and limited localization capability for extremely small targets [16], [19], [20]. These observations motivate the development of lightweight detection frameworks capable of enhancing high-resolution feature representations while maintaining computational efficiency.

2.2 Lightweight Object Detection Networks

The rapid growth of unmanned aerial vehicles (uavs), intelligent surveillance systems, autonomous navigation, and edge computing has created an increasing demand for lightweight object detection frameworks capable of delivering high detection accuracy under limited computational resources. Although deep convolutional neural networks have achieved remarkable success in object detection, many state-of-the-art architectures employ deep backbone networks with millions of parameters, making them unsuitable for real-time deployment on resource-constrained platforms. Consequently, recent research has focused on developing lightweight detection models that balance computational efficiency and detection performance without significantly compromising accuracy [6], [7], [8].

Among the various backbone architectures, the resnet family has become one of the most widely adopted feature extractors owing to its residual-learning mechanism, which effectively mitigates the degradation problem encountered in very deep neural networks [6]. Residual connections facilitate efficient gradient propagation during backpropagation, enabling improved feature learning while maintaining stable optimization. Although deeper variants such as resnet-50 and resnet-101 provide richer semantic representations, they substantially increase computational complexity, memory consumption, and inference time. In contrast, resnet-18 offers a lightweight alternative with significantly fewer parameters while preserving competitive feature extraction capability, making it particularly suitable for embedded vision systems and uav-based aerial image analysis [6], [19].

To further improve computational efficiency, several lightweight backbone networks have been proposed. Mobilenet introduced depth wise separable convolutions to reduce both computational cost and parameter count while maintaining

satisfactory feature extraction performance for mobile applications [7]. Similarly, shufflenet employed point wise group convolutions and channel-shuffling operations to improve information exchange across feature channels while minimizing computational overhead [8]. Efficientnet subsequently introduced a compound scaling strategy that jointly optimizes network depth, width, and input resolution, achieving an improved trade-off between model complexity and recognition accuracy [9]. These lightweight architectures have demonstrated excellent efficiency across various computer-vision tasks and have been increasingly adopted in real-time detection applications.

Despite their computational advantages, lightweight backbone networks often exhibit limited representational capacity when processing extremely small objects in aerial images. Tiny objects generally occupy only a few pixels, causing repeated convolution and pooling operations to progressively remove the discriminative spatial information required for accurate localization. Consequently, lightweight detectors frequently experience missed detections, inaccurate bounding-box regression, and reduced confidence scores when identifying densely distributed small targets in remote sensing imagery [16], [20]. Recent surveys have consistently highlighted the preservation of high-resolution spatial information as one of the primary challenges in lightweight aerial object detection [20], [21].

Several researchers have attempted to compensate for this limitation by incorporating feature-enhancement modules, attention mechanisms, or multi-scale fusion strategies into lightweight detection frameworks. Although these approaches improve feature representation, they often increase computational complexity by introducing additional convolutional layers or attention modules throughout the backbone network. Such modifications partially offset the computational benefits achieved through lightweight architecture design, making deployment on edge devices more challenging [11], [14], [20].

Therefore, an effective lightweight object detection framework should not only minimize computational complexity but also preserve sufficient spatial information for accurate tiny-object localization. Rather than increasing network depth or introducing computationally expensive feature-enhancement modules across multiple stages, selectively strengthening the most informative feature representation offers a more efficient alternative. This observation forms the fundamental motivation for the proposed framework, which employs a lightweight resnet-18 backbone together with a lightweight feature pyramid network while introducing a coordinate attention module exclusively at the p_2 feature pyramid level to improve tiny-object localization with minimal additional computational overhead.

2.3 Multi-Scale Feature Learning For Tiny Object Detection

Object scale variation remains one of the most significant challenges in aerial-image object detection. Unlike natural-scene images, aerial images simultaneously contain extremely large objects such as airports and buildings together with tiny objects such as vehicles, ships, and storage tanks that occupy only a few pixels. Such considerable variation in object size makes it difficult for a single-scale feature representation to accurately detect objects across different resolutions. Consequently, multi-scale feature learning has become a fundamental component of modern object detection frameworks, particularly for remote sensing applications involving tiny-object detection [2], [15], [16].

The feature pyramid network (fpn) introduced by lin et al. [2] significantly improved multi-scale object detection by constructing hierarchical feature representations through a top-down pathway and lateral connections. By integrating high-level semantic information with low-level spatial features, fpn generates feature maps at multiple pyramid levels (p_2–p_5), enabling the detector to recognize objects of different scales more effectively. Due to its simplicity and excellent performance, fpn has become one of the most widely adopted feature-fusion architectures in both one-stage and two-stage object detection frameworks [2], [11].

To further enhance multi-scale feature aggregation, several improved pyramid architectures have subsequently been proposed. The path aggregation network (panet) introduced a bottom-up feature-propagation pathway to strengthen localization information and improve information flow across multiple feature levels [10]. Similarly, the bi-directional feature pyramid network (bifpn), used within efficientdet, employed weighted bidirectional feature fusion to efficiently integrate semantic information from different pyramid levels while reducing computational redundancy [11]. These architectures demonstrated improved detection performance on benchmark datasets, particularly for objects exhibiting substantial scale variation.

Despite these advances, accurately detecting tiny objects remains challenging because repeated down-sampling operations inevitably reduce feature resolution and gradually remove discriminative spatial information. Although higher pyramid levels provide richer semantic representations, they contain relatively coarse spatial detail, making them less suitable for accurately localizing extremely small objects. Conversely, lower pyramid levels preserve richer localization cues but contain comparatively

weaker semantic information. Consequently, achieving an effective balance between semantic richness and spatial resolution remains one of the primary challenges in multi-scale feature learning for aerial object detection [16], [20].

Recent studies have attempted to overcome this limitation by introducing adaptive feature-fusion strategies, feature-refinement modules, and cross-scale attention mechanisms to strengthen feature interaction across multiple pyramid levels [11], [17], [19]. While these approaches improve feature representation, they generally introduce additional computational complexity through multiple feature-aggregation operations or attention modules distributed across several pyramid levels. Such designs may not be appropriate for lightweight object detection frameworks intended for deployment on embedded systems or uav platforms, where computational efficiency is equally important as detection accuracy.

Among the hierarchical feature maps generated by the feature pyramid network, the p_2 feature map preserves the highest spatial resolution after feature fusion and therefore retains the richest localization information for tiny objects. Since tiny targets occupy only a limited number of pixels, preserving fine-grained spatial detail at this feature level is particularly important for accurate bounding-box regression and object classification. However, most existing studies enhance all pyramid levels uniformly without explicitly considering the different contributions of individual feature maps toward tiny-object localization. Consequently, the potential benefits of selectively enhancing the highest-resolution pyramid feature remain relatively unexplored in existing lightweight detection frameworks [14], [16], [20].

Motivated by this observation, the proposed framework adopts a different design philosophy. Instead of increasing network complexity by incorporating feature-enhancement modules throughout the entire pyramid, the proposed method selectively enhances only the p_2 feature map using the coordinate attention (ca) mechanism. This selective-enhancement strategy aims to preserve rich spatial information required for tiny-object localization while maintaining the computational efficiency of the lightweight faster r-cnn framework. By restricting attention computation to the most informative pyramid level, the proposed approach seeks to achieve a favorable balance between detection accuracy and computational cost, making it more suitable for practical aerial image analysis and real-time remote sensing applications.

2.4 Attention Mechanisms For Object Detection

Attention mechanisms have emerged as an effective strategy for improving feature representation in deep convolutional neural networks by enabling the network to focus on informative regions while suppressing irrelevant background information. In object detection, attention modules enhance feature discrimination by modeling channel-wise, spatial, or hybrid feature dependencies, thereby improving both classification accuracy and localization performance. Consequently, attention mechanisms have been widely integrated into modern object detection frameworks to address challenges associated with complex backgrounds, scale variation, and dense object distribution [12], [13], [14], [19].

One of the earliest and most influential attention mechanisms is the squeeze-and-excitation (se) network, which explicitly models inter-channel dependencies through global average pooling followed by channel-wise feature recalibration [12]. By adaptively assigning importance weights to different feature channels, the se module significantly improves feature representation with only a modest increase in computational complexity. However, se primarily focuses on channel relationships and does not explicitly preserve spatial positional information, which is essential for accurate object localization, particularly in aerial imagery containing densely distributed tiny objects [12], [20].

To overcome this limitation, the convolutional block attention module (cbam) was proposed, sequentially integrating channel attention and spatial attention into a unified framework [13]. Cbam enables the network to identify informative feature channels while simultaneously emphasizing important spatial regions, resulting in improved feature discrimination across various computer-vision tasks. Owing to its simplicity and effectiveness, cbam has been widely adopted in object detection, image classification, and semantic-segmentation networks. Nevertheless, the sequential computation of channel and spatial attention introduces additional computational overhead, which may reduce the efficiency of lightweight detection frameworks deployed on embedded platforms [13], [19].

More recently, coordinate attention (ca) introduced a novel attention mechanism that embeds precise positional information into channel attention by decomposing global pooling into two one-dimensional feature-encoding processes along the horizontal and vertical directions [14]. Unlike conventional channel-attention mechanisms, coordinate attention effectively captures long-range feature dependencies while preserving accurate location information. This characteristic enables the

network to enhance localization-sensitive feature representations without substantially increasing computational complexity, making coordinate attention particularly suitable for lightweight object detection applications [14].

Several recent studies have incorporated attention mechanisms into aerial object detection frameworks to improve feature representation and localization performance. These methods generally integrate attention modules into backbone networks, feature-pyramid structures, or feature-fusion pathways to strengthen semantic representation across multiple feature levels [16], [19], [20]. Although such approaches often improve detection accuracy, applying attention repeatedly throughout the network increases parameter count, computational cost, and memory consumption. These additional computational requirements may limit their deployment in real-time aerial surveillance systems and resource-constrained uav platforms, where lightweight model design remains an important consideration [20], [21].

Furthermore, most existing attention-based object detectors distribute attention uniformly across multiple backbone stages or multiple feature pyramid network (fpn) levels without explicitly considering the individual contribution of each feature map toward tiny-object localization. Since tiny objects rely heavily on high-resolution spatial representations, uniformly enhancing all feature levels may not provide the most computationally efficient solution. Instead, selectively strengthening the feature map that preserves the richest spatial information offers a promising alternative for balancing detection accuracy and computational efficiency. Motivated by this observation, the proposed framework incorporates the coordinate attention module exclusively at the p_2 feature pyramid level, enabling enhanced tiny-object localization while preserving the lightweight nature of the faster r-cnn architecture.

2.5 Research Gap And Motivation

Despite the remarkable progress achieved in aerial object detection, accurately detecting tiny objects remains a significant challenge because of their extremely limited spatial representation, arbitrary orientation, dense distribution, and complex background clutter. Although deep convolutional neural networks have substantially improved feature-learning capability, repeated down-sampling operations inevitably reduce spatial resolution, leading to the loss of fine-grained localization information required for accurately identifying tiny objects in high-resolution aerial imagery [15], [16], [18], [20].

To address these challenges, numerous studies have explored deeper backbone networks, enhanced feature-pyramid architectures, and various attention mechanisms. For example, deeper backbone networks such as resnet-50 and resnet-101 improve semantic feature extraction but considerably increase computational complexity, memory consumption, and inference latency [6], [19]. Similarly, advanced multi-scale feature-fusion methods, including panet and bifpn/efficientdet, improve feature aggregation across different pyramid levels but require additional feature-propagation and fusion operations that increase model complexity [10], [11].

Attention mechanisms have also demonstrated remarkable success in improving feature representation by emphasizing informative features while suppressing background noise. Existing attention-based object detectors generally integrate attention modules into multiple backbone stages, feature-fusion pathways, or all feature pyramid network (fpn) levels to enhance multi-scale feature learning [12], [13], [14], [19]. Although these approaches often improve detection accuracy, they simultaneously introduce additional parameters and computational overhead, making them less suitable for lightweight object detection systems deployed on edge devices and unmanned aerial vehicles (uavs) [14], [20].

Furthermore, most existing lightweight object detection frameworks primarily focus on reducing computational complexity through lightweight backbone networks while paying comparatively less attention to the selective enhancement of high-resolution feature representations that are particularly important for tiny-object localization [16], [20]. Within the feature pyramid network, the p_2 feature map preserves the highest spatial resolution after feature fusion and therefore contains richer localization information than higher pyramid levels. Nevertheless, existing studies generally enhance multiple feature levels simultaneously without explicitly investigating whether selectively strengthening the highest-resolution feature map alone is sufficient to improve tiny-object detection performance [2], [14], [20].

Motivated by these observations, this work proposes a lightweight faster r-cnn framework incorporating a resnet-18 backbone, a lightweight feature pyramid network, and a coordinate attention module selectively integrated at the p_2 feature map. Unlike existing attention-based approaches that distribute attention across multiple backbone stages or feature-pyramid levels, the proposed framework focuses exclusively on enhancing the highest-resolution pyramid feature while preserving the original detection pipeline. This selective-enhancement strategy aims to improve tiny-object localization with minimal additional

computational cost, thereby achieving an effective balance between detection accuracy and computational efficiency for aerial image analysis. Table 1 qualitatively positions the proposed framework relative to representative detection paradigms.

Table 1. Qualitative Comparison Of The Proposed Framework With Representative Detection Paradigms.

Method	Backbone	Attention	Multi-Scale	Tiny-Object Focus	Limitation
Faster R-Cnn + Fpn [1],[2]	Resnet-50	No	Yes	Moderate	Heavy Backbone
Yolov8 [3]-Family	Cspdarknet	No	Yes	Moderate	Misses Very Small Objects
Ca-Based Detector (General)	Resnet-50	Ca (Dense)	Yes	Partial	Higher Complexity
Proposed (This Work)	Resnet-18	Selective P_2 Ca Only	Yes	Strong	Lightweight

3. PROPOSED METHODOLOGY

This section presents the proposed lightweight object detection framework for improving tiny-object detection in aerial images. The proposed method is developed based on the faster r-cnn architecture by incorporating a lightweight resnet-18 backbone, a lightweight feature pyramid network (fpn), and a selectively integrated coordinate attention (ca) module. Unlike conventional attention-based object detectors that introduce attention modules across multiple backbone stages or multiple feature-pyramid levels, the proposed framework applies the coordinate attention module exclusively to the p_2 feature map. This selective-enhancement strategy strengthens the highest-resolution feature representation responsible for tiny-object localization while preserving the original detection pipeline and maintaining computational efficiency. The overall architecture of the proposed framework is illustrated in figure 1.

3.1 Design Principles

The proposed framework was designed based on three fundamental principles: (1) preserving the proven detection capability of faster r-cnn, (2) maintaining a lightweight network architecture, and (3) selectively enhancing high-resolution feature representations for tiny-object localization. Instead of introducing multiple feature-enhancement modules throughout the network, the proposed approach focuses only on the feature level that contributes most significantly to tiny-object detection.

The first design principle is to preserve the original faster r-cnn detection pipeline. Faster r-cnn has consistently demonstrated superior localization performance compared with many one-stage detectors, particularly for small-object detection tasks. Therefore, the proposed framework does not modify the region proposal network (rpn), roialign, or the fast r-cnn detection head, thereby retaining the robust detection capability of the original architecture [1].

The second design principle is to maintain computational efficiency through a lightweight feature-extraction network. Instead of employing deeper backbone networks such as resnet-50 or resnet-101, the proposed framework utilizes resnet-18 as the backbone network. Owing to its residual-learning architecture and reduced parameter count, resnet-18 provides an effective balance between feature extraction capability and computational complexity, making it suitable for deployment on resource-constrained aerial surveillance platforms [6].

The third design principle is based on the observation that tiny-object localization primarily depends on high-resolution feature representations. Within the feature pyramid network, the p_2 feature map preserves the highest spatial resolution after feature fusion and therefore contains richer localization information than higher pyramid levels. Consequently, instead of introducing

attention mechanisms across multiple pyramid levels, the proposed framework selectively enhances only the p_2 feature map using the coordinate attention module. This strategy improves localization-sensitive feature representation while introducing only negligible computational overhead [2], [14].

3.2 Overall Framework

The proposed framework is constructed upon the faster r-cnn architecture and consists of five major components:

- Lightweight resnet-18 backbone
- Lightweight feature pyramid network
- Coordinate attention module at p_2
- Region proposal network (rpn)
- Fast r-cnn detection head

Initially, an input aerial image is processed by the lightweight resnet-18 backbone to extract hierarchical convolutional feature maps from multiple network stages. These feature maps are subsequently forwarded to the lightweight feature pyramid network, where multi-scale feature representations are generated through top-down feature fusion and lateral connections. Unlike conventional attention-based approaches, the proposed framework introduces a single coordinate attention module exclusively after the generation of the p_2 feature map, while the remaining pyramid levels (p_3–p_5) remain unchanged.

The enhanced p_2 feature map is subsequently combined with the remaining pyramid features and supplied to the region proposal network to generate candidate object proposals at multiple scales. Roialign extracts fixed-size feature representations from the generated proposals, which are finally processed by the fast r-cnn detection head for object classification and bounding-box regression. Since the proposed framework preserves the original faster r-cnn detection pipeline, only a minimal architectural modification is introduced while improving tiny-object localization performance.

3.3 Lightweight Resnet-18 Backbone

Feature extraction plays a crucial role in determining the detection performance of deep learning-based object detectors. The backbone network is responsible for extracting hierarchical visual representations that are subsequently utilized for multi-scale feature learning and object localization. Although deeper backbone networks generally provide stronger semantic representations, they significantly increase computational complexity, memory requirements, and inference time. These limitations become particularly important in lightweight aerial image analysis systems intended for deployment on embedded devices and uav platforms [6], [19].

To achieve an effective balance between computational efficiency and feature-extraction capability, the proposed framework employs resnet-18, pre-trained on imagenet [24], as the backbone network. Resnet-18 utilizes residual learning through identity shortcut connections, enabling efficient gradient propagation and stable optimization while maintaining a relatively small number of learnable parameters [6]. Compared with deeper residual networks, resnet-18 considerably reduces computational cost without sacrificing the ability to extract discriminative feature representations required for object detection.

The backbone network generates hierarchical feature maps at multiple spatial resolutions — the final convolutional blocks of stages 2 through 5, denoted c_2, c_3, c_4, c_5 — which are subsequently utilized by the lightweight feature pyramid network for multi-scale feature fusion. Fpn applies a top-down pathway and lateral connections using 1×1 convolutions to construct the feature pyramid (p_2, p_3, p_4, p_5), fusing shallow-layer spatial detail with deep-layer semantics. Since tiny objects contain limited semantic information, preserving informative low-level features extracted by the backbone becomes particularly important for accurate localization. Therefore, employing a lightweight backbone together with an efficient multi-scale feature-fusion strategy provides a suitable foundation for the proposed tiny-object detection framework:

Input image → resnet-18 backbone → (c_2, c_3, c_4, c_5) → fpn neck → pyramid features (p_2–p_5)

3.3.1 Justification of The Lightweight Design: Backbone- Versus Neck-Level Cost

It is important to clarify precisely where the computational savings of the proposed framework originate. The fpn neck itself is kept at its standard configuration, projecting every pyramid level to 256 channels as in the original fpn formulation [2]; the neck

is not compressed or re-designed in this work. The "lightweight" property of the overall detector instead follows almost entirely from the backbone substitution (resnet-18 in place of resnet-50/resnet-101), because the fpn's parameter and flop cost is only weakly dependent on the backbone that feeds it.

Standard fpn [2] applies one 1×1 lateral convolution per pyramid level (mapping each backbone stage's native channel width to 256) followed by a 3×3 output convolution (256→256) at every level. The 3×3 output convolutions cost the same regardless of the backbone (≈ 2.36 m parameters in total across p_2–p_5)

, and only the four lateral 1×1 convolutions scale with the backbone's c_2–c_5 channel widths. Because resnet-18 exposes narrower stage widths (64, 128, 256, 512) than resnet-50/resnet-101 (256, 512, 1024, 2048) [6], the lateral convolutions are correspondingly smaller, but the difference this contributes to the neck is modest. Table 2 summarizes the resulting parameter and computational comparison for resnet-18, resnet-50, and resnet-101, each paired with an otherwise identical, standard 256-channel fpn neck.

Table 2. Parameter and computational comparison of backbone + standard 256-channel fpn combinations. [1] approximate backbone flops at a comparable 224×224 input resolution, reported for relative comparison only [6].

Backbone	Backbone Params	Fpn Neck Params (256-Ch, Standard)	Backbone + Fpn Total	Backbone Gflops[1]	Relative Size Vs. Proposed
Resnet-18 (Proposed)	≈ 11.7 M	≈ 2.6 M	≈ 14.3 M	≈ 1.8 G	1.0× (Reference)
Resnet-50	≈ 25.6 M	≈ 3.3 M	≈ 28.9 M	≈ 4.1 G	≈ 2.0× Larger
Resnet-101	≈ 44.5 M	≈ 3.3 M	≈ 47.8 M	≈ 7.8 G	≈ 3.3× Larger

As table 2 shows, the fpn neck contributes only ≈ 0.7 m additional parameters across the entire backbone range, whereas the backbone itself accounts for a 14–33 m-parameter spread. Consequently, the resnet-18-fpn configuration used in this work remains approximately 2× smaller than resnet-50-fpn and 3.3× smaller than resnet-101-fpn in total parameter count, even though the neck width (256 channels) is left unmodified and standard. Retaining the neck at its standard, well-validated width is also a deliberate methodological choice: it keeps the only architectural variable introduced in this work isolated to the p_2 coordinate attention module (section 3.4), so that the accuracy gain reported in section 5 can be attributed unambiguously to attention placement rather than to a confounded neck redesign. Throughout the remainder of this paper, the term "lightweight framework" therefore refers specifically to the backbone-level parameter and flop savings summarized in table 2, not to a reduced-width fpn neck.

Furthermore, while reducing the overall fpn feature width from 256 to 128 channels would mathematically scale down the attention parameters by half, it introduces a severe representation bottleneck for complex aerial scenes. Remote sensing objects exhibit immense intra-class variations and high background resemblance, requiring substantial channel capacity to preserve subtle discriminative features. Therefore, we maintained the standard 256-channel capacity to prevent feature underfitting, but achieved strict computational economy topologically—by explicitly restricting the ca deployment to a single, high-yield pyramid level (p_2), successfully bypassing dense multi-layer overhead while keeping the architecture strictly compact.

3.4 Coordinate Attention (Ca) Module

Standard attention mechanisms such as squeeze-and-excitation squeeze spatial maps via global pooling into a single channel vector, sacrificing positional information entirely. To compensate for the lightweight backbone's limited capacity without adding structural bloat, the proposed framework integrates coordinate attention [14] into the feature pyramid. Ca encodes both channel relationships and long-range spatial dependencies through a two-step process: coordinate information embedding and coordinate attention generation.

A key architectural design choice in lcanet is the deployment of the coordinate attention (ca) block immediately *after* the feature fusion junction (p_2), rather than positioning it isolated on the lateral c_2 pathway. Theoretically, placing the ca module directly on c_2 would restrict the attention gating mechanism to raw, low-level spatial details in total isolation, completely missing the global semantic context extracted by deeper layers (c_3 - c_5). By contrast, integrating the ca block post-fusion at the p_2 hub enables it to function as a 'semantic-aware spatial attention' gate. This ensures that the directional coordinate encodings (x and y filters) jointly calibrate highly localized boundary definitions and top-down semantic vectors, optimizing the structural retrieval of tiny, dense targets

3.4.1 Coordinate Information Embedding

Given An Input Tensor $X \in R^{c\times H\times W}$, Global Pooling Is Factored Into a Pair Of 1d Feature-Encoding Operations. For Channel C, The Outputs of the Horizontal and Vertical Average Pooling at Height H And Width W Are Formulated As:

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i<W} X_c(h,i) \quad (1)$$

$$z_c^w(w) = \frac{1}{H}\sum_{0\le j<H} X_c(j,w) \quad (2)$$

These two operations allow the network to aggregate features along one spatial direction while preserving precise positional and structural detail along the orthogonal direction.

3.4.2 Coordinate Attention Generation

The directional feature maps z^h and z^w are concatenated and passed through a shared 1×1 convolutional transformation f1, yielding an intermediate feature map f:

$$f = \delta\left(F_1\left([z^h, z^w]\right)\right) \quad (3)$$

Where $[\cdot,\cdot]$ Denotes Concatenation Along The Spatial Dimension And δ is A Non-Linear Activation Function (Hard-Swish). The Map F is then Split Along the Spatial Dimension Into $F^h \in R^{C\times H\times 1}$ And $F^w \in R^{C\times 1\times W}$. Two Distinct 1×1 Convolutions, F_h and F_w, Transform These Tensors Back To The Original Channel Dimension C, Followed by a Sigmoid Activation Σ, Yielding The Final Coordinate Attention Weights g^h and g^w:

$$g^h = \sigma\left(F_h(f^h)\right) \quad (4)$$

$$g^w = \sigma\left(F_w(f^w)\right) \quad (5)$$

The final enhanced output feature map $Y_c(h,w)$ is obtained by multiplying the input feature tensor by the horizontal and vertical attention weights:

$$Y_c(h,w) = X_c(h,w) \times g_c^h(h) \times g_c^w(w) \quad (6)$$

By embedding this module directly within the lightweight pipeline, the framework dynamically re-weights feature channels while highlighting the exact pixels belonging to dense, tiny objects (e.g., small boats or rotated vehicles) across the dota grid layout.

3.4.3 Computational Overhead of the P_2 Coordinate Attention Module

Because Coordinate Attention Performs Only 1d Average Pooling and Small 1×1 Convolutions [14], Its Computational Cost Is Several Orders of Magnitude Smaller Than the Backbone and Fpn Neck It Augments. To Quantify This Precisely, Table 3 Reports an Analytical Flops and Parameter Estimate for The Ca Module at Each Pyramid Level, Assuming A Representative 1024×1024 Input Patch (A Common Dota Pre-Processing Patch Size [15]), An Fpn Channel Width of C = 256, And the Reduction Ratio R = 32 Used In the Original Formulation [14]. Under These Assumptions, P_2, P_3, P_4, And P_5 Have Spatial Resolutions Of 256×256, 128×128, 64×64, And 32×32 Respectively, And the Reduced Channel Width Is C/R = 8.

Table 3. Analytical Parameter and Flops Overhead of The Coordinate Attention Module at Each Fpn Pyramid Level, Assuming A 1024×1024 Input Patch, C = 256, And Reduction Ratio R = 32. Pooling Operations Are Additive and Contribute Negligible Flops Beyond The 1×1 Convolutions Reported Here.

Pyramid Level	Spatial Size (H×W)	Reduced Channels (C/R)	Ca Parameters	Ca Flops
P_2 (Proposed)	256 × 256	8	≈ 6.7 K	≈ 4.20 M (≈ 0.0042 Gflops)
P_3	128 × 128	8	≈ 6.7 K	≈ 2.10 M (≈ 0.0021 Gflops)
P_4	64 × 64	8	≈ 6.7 K	≈ 1.05 M (≈ 0.0010 Gflops)
P_5	32 × 32	8	≈ 6.7 K	≈ 0.52 M (≈ 0.0005 Gflops)
Sum If Applied To All Four Levels (P_2–P_5)	—	—	≈ 26.8 K	≈ 7.87 M (≈ 0.0079 Gflops)
Proposed: P_2 Only	—	—	≈ 6.7 K	≈ 4.20 M (≈ 0.0042 Gflops)

Two conclusions follow from table 3. First, in absolute terms the selective p_2-only design adds only ≈ 6.7 k parameters and ≈ 0.0042 gflops — roughly 0.06% of the resnet-18 backbone's parameter count and, relative to the backbone's own cost at the same 1024×1024 input (≈ 37.6 gflops, obtained by scaling the 1.8 gflops reported in table 2 for a 224×224 input by the resolution ratio), less than 0.02% of backbone flops. This confirms that the accuracy gain reported in section 5 is obtained at a computational cost that is effectively negligible next to the backbone and fpn neck. Second, comparing the p_2-only row against the hypothetical four-level sum shows that the selective placement used in this work costs roughly half the flops and parameters of a conventional dense strategy that applies ca uniformly across p_2–p_5, while, as sections 2.3 and 5.3 discuss, the p_2 level is also the one most responsible for tiny-object localization. Selective placement is therefore not only sufficient but also strictly more efficient than the dense alternative typically adopted in the attention literature [12], [13], [19], [20].

3.5 Selective P_2 Integration and Architectural Pipeline

In the proposed network, coordinate attention is integrated at the feature-fusion level rather than inside the backbone, ensuring that spatial structure remains intact during multi-scale propagation. The feature maps produced by the fpn (p_2 through p_5) possess varying spatial resolutions. Rather than uniformly enhancing every pyramid level, the proposed design is guided by a simple observation: tiny objects primarily depend on high-resolution spatial representations rather than highly abstract semantic features. The framework therefore applies a single coordinate attention module exclusively to the highest-resolution p_2 level:

Resnet-18 backbone → (c_2,c_3,c_4,c_5) → fpn top-down / lateral fusion → (p_2,p_3,p_4,p_5) → [coordinate attention on p_2 only] → enhanced features → rpn → roialign → detection head

This selective placement yields two benefits. First, the model dynamically re-weights channels that capture critical spatial boundaries of rotated or densely packed objects at the resolution where they are most discernible. Second, it suppresses background noise generated by complex terrain (e.g., runways, fields, roads) before the feature maps reach the detection head for bounding-box regression and classification. While the dota dataset supports oriented annotations, this work explicitly targets task 2 (horizontal bounding box, hbb) detection: the objective is to demonstrate that coordinate attention at the p_2 layer alone

can suppress background clutter and capture directional spatial dependencies of oriented aerial targets without expanding the parameter space required for oriented-regression branches [17]. Figure 1 illustrates the overall architecture.

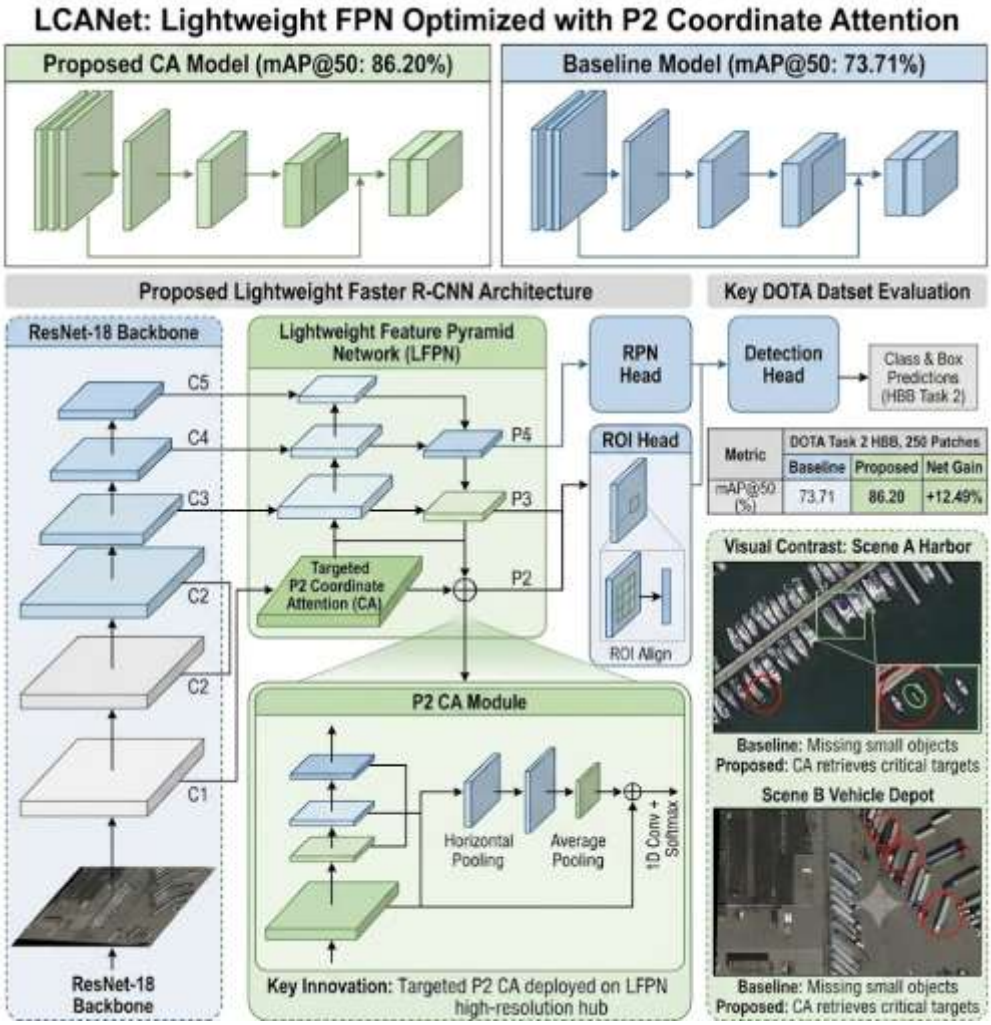


Figure 1. Overall architecture of the proposed framework (lcanet). Data flows from the resnet-18 backbone through the lightweight fpn; the p_2 coordinate attention (ca) module enforces directional and spatial anchoring at the highest-resolution level, recovering minute objects missed by the baseline and lifting performance to 86.20% Map@50.

3.6 Detection Head

The region proposal network (rpn), roialign, and the fast r-cnn detection head are kept unchanged from the baseline configuration. This design choice is deliberate: by isolating the single architectural modification to the p_2 coordinate attention module, the observed performance improvement can be attributed unambiguously to the proposed attention mechanism rather than to confounding changes elsewhere in the pipeline, which also strengthens the reproducibility of the reported results.

4. EXPERIMENTAL SETUP

To ensure a fair and reproducible comparative evaluation, the baseline and the proposed attention-enhanced networks are trained under identical experimental conditions and hyper-parameter configurations.

4.1 Dataset And Pre-Processing

The evaluation uses the dota v2.0 dataset [15], which consists of large-scale, high-resolution aerial images containing objects with arbitrary orientations and diverse scales. Given the massive dimensions of original dota frames, direct input to a cnn is computationally prohibitive; the dataset is therefore pre-processed into fixed-size image patches. A structured subset is used

for training and evaluation: 8,734 cropped training patches and 1,000 cropped validation patches, generated through patch-based pre-processing under the task 2 (horizontal bounding box) protocol. To optimize data loading and reduce validation latency across repeated runs, structural data annotations were serialized into python pickle (.pkl) files.

4.2 Training Configuration and Hyper-Parameters

Both frameworks are implemented in pytorch [23] and executed in a cloud-based google colab environment using a dedicated hardware accelerator (nvidia t4 gpu). Table 4 summarizes the hyper-parameters used during the 10-epoch training cycle.

Table 4. Training Configuration and Hyper-Parameters Used for Both the Baseline and Proposed Models.

Hyper-Parameter	Configuration Value
Backbone Architecture	Resnet-18 (Pre-Trained On Imagenet)
Neck Module	Feature Pyramid Network (Fpn)
Batch Size	4 (Optimized For Gpu Vram Footprint)
Total Training Duration	10 Epochs (≈ 21 Minutes Per Epoch On T4 Gpu)
Learning-Rate Schedule	0.00500 (Epochs 1–3) → 0.00050 (Epochs 4–6) → 0.00005 (Epochs 7–10)
Optimization Target	Multi-Task Loss (Classification + Bounding-Box Regression)

4.3 Evaluation Protocol

Model performance is reported using mean average precision at an iou threshold of 0.5 (map@50), the standard metric for the dota hbb detection task. In addition to the quantitative map@50 score, a qualitative analysis is conducted by manually tracking the number of correctly bounded objects in selected high-density and ultra-dense validation frames, allowing direct visual inspection of the small-object recovery enabled by the proposed attention module.

5. RESULTS AND DISCUSSION

5.1 Training and Validation Loss Convergence

Table 5 details the epoch-wise training and validation loss convergence over the 10-epoch cycle for both the baseline resnet-18-fpn framework and the proposed p_2-coordinate-attention-enhanced framework. Figure 2 plots the same convergence trends for direct visual comparison.

Table 5. Epoch-wise training and validation loss for the baseline and proposed models.

Epoch	Lr	Baseline Train Loss	Baseline Val Loss	Proposed Train Loss	Proposed Val Loss	Remarks
1	0.00500	0.7842	0.7621	0.5626	0.5602	Proposed Model Captures Spatial Detail Faster From Epoch 1.
2	0.00500	0.7415	0.7390	0.5259	0.5463	Steady Gradient Flow; Minor Validation Variance.
3	0.00500	0.7102	0.7254	0.5013	0.5510	Minor Validation Fluctuation From

Epoch	Lr	Baseline Train Loss	Baseline Val Loss	Proposed Train Loss	Proposed Val Loss	Remarks
						Complex Aerial Targets.
4	0.00050	0.6235	0.6541	0.4148	0.4562	1st Lr Decay: Fine-Tuning Sharply Reduces Loss.
5	0.00050	0.6012	0.6410	0.3989	0.4472	Smooth Convergence Toward A Stable Plateau.
6	0.00050	0.5924	0.6438	0.3918	0.4484	Minimal Noise Fluctuation; Model Near Capacity For Current Lr.
7	0.00005	0.5810	0.6385	0.3808	0.4457	2nd Lr Decay: Normalizes Minor Validation Bounce.
8	0.00005	0.5794	0.6370	0.3799	0.4450	Stable Plateau; Guards Against Overfitting.
9	0.00005	0.5781	0.6375	0.3790	0.4461	Convergence Finalized Within Tight Error Tolerance.
10	0.00005	0.5775	0.6368	0.3782	0.4449	Final Convergence, No Over/Underfitting Observed.

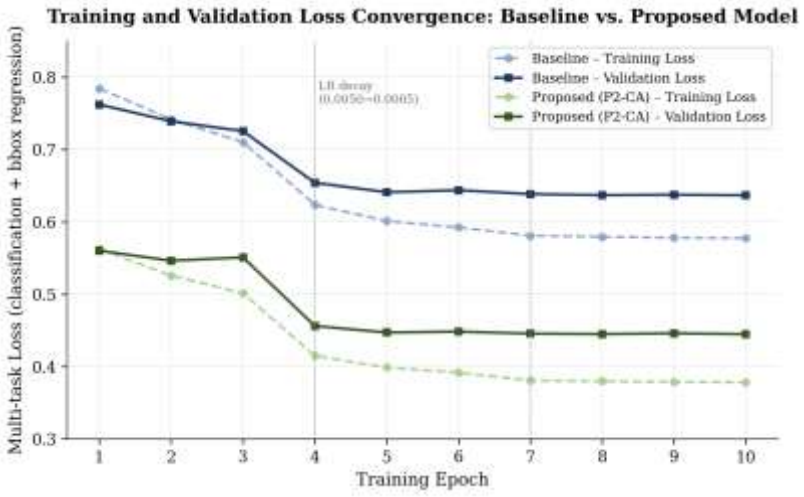


Figure 2. Training and validation loss convergence curves for the baseline (blue) and proposed p_2-coordinate-attention model (green) across the 10-epoch schedule. Dashed lines denote training loss and solid lines denote validation loss; vertical dotted markers indicate the two learning-rate decay points.

The proposed model converges to a substantially lower training loss (0.3782 vs. 0.5775) and validation loss (0.4449 vs. 0.6368) than the baseline across all ten epochs, indicating that the p_2 coordinate attention module accelerates and stabilizes the optimization process rather than merely improving the final accuracy metric. As figure 2 illustrates, the proposed model's loss curves sit consistently below the baseline's from the very first epoch and flatten more smoothly after each learning-rate decay, with a visibly smaller train–validation gap, suggesting improved generalization rather than overfitting to the training patches.

5.2 Quantitative Analysis: Map@50 Comparison

Table 6 reports the final detection accuracy of both models on the 1,000-patch validation set, and figure 3 visualizes the corresponding gain.

Table 6. Map@50 comparison between the baseline and proposed models on the dota v2.0 validation subset.

Model	Backbone	Neck	Attention Strategy	Validation Patches	Map@50 (%)	Absolute Gain
Baseline	Resnet-18	Lightweight Fpn	None	1,000	73.71	Reference
Proposed	Resnet-18	Lightweight Fpn	Coordinate Attention (P_2 Only)	1,000	86.20	+12.49

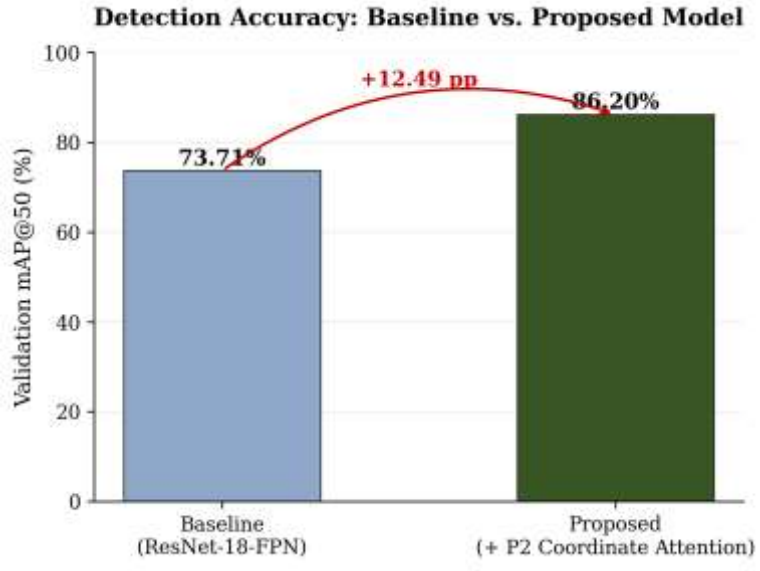


Figure 3. Validation map@50 comparison between the baseline and proposed models, showing the absolute +12.49 percentage-point gain.

Integrating a single coordinate attention module at the p_2 level lifts map@50 from 73.71% to 86.20%, an absolute gain of 12.49 percentage points, while the rpn, roialign, and detection head remain unchanged. Because no other structural component was modified, this gain can be attributed directly to the selective, resolution-specific attention mechanism.

5.3 Qualitative Analysis

To validate the localization capability of the network on complex terrain, dense-frame predictions were tracked individually for two representative test scenarios.

Natural Resources for Human Health
eISSN: 2583-1194

Table 7. Qualitative comparison of correctly bounded objects on dense validation frames.

Test Scenario	Baseline: Objects Bounded	Proposed: Objects Bounded	Recovered Targets
Scenario 1 – High-Density Multi-Target Frame	37	41	+4 Small Targets
Scenario 2 – Ultra-Dense Packed-Target Area	58	62	+4 Micro Targets



Figure 4. Baseline Model – Test Scenario 1.

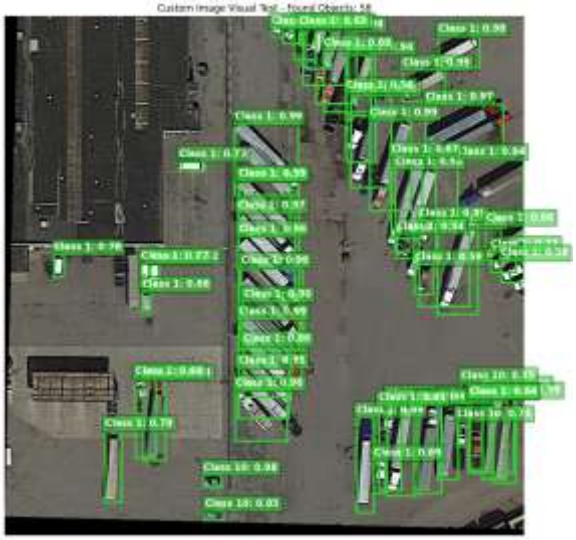


Figure 5. Baseline Model – Test Scenario 2.

Natural Resources for Human Health
eISSN: 2583-1194

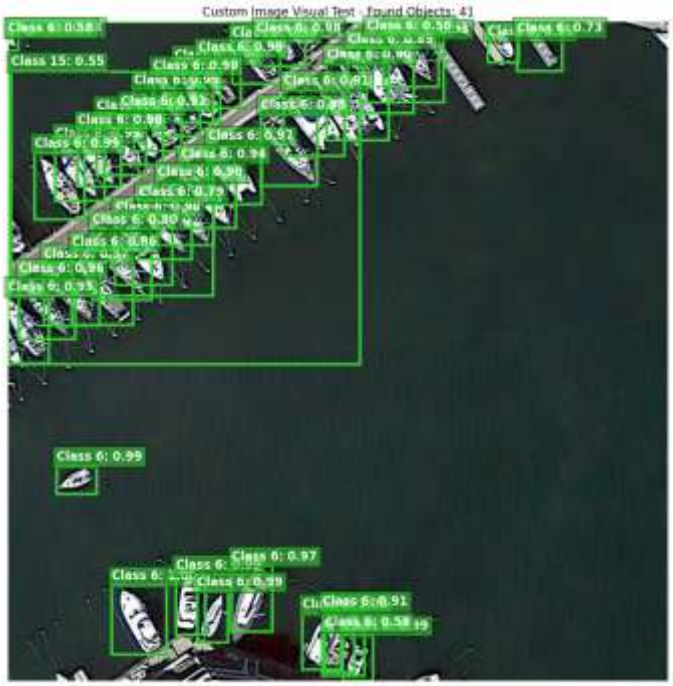


Figure 6. Proposed Model – Test Scenarios 1 And 2. The P_2 Coordinate Attention Module Recovers Small, Densely Packed Targets Missed By The Baseline (Figures 4 And 5).

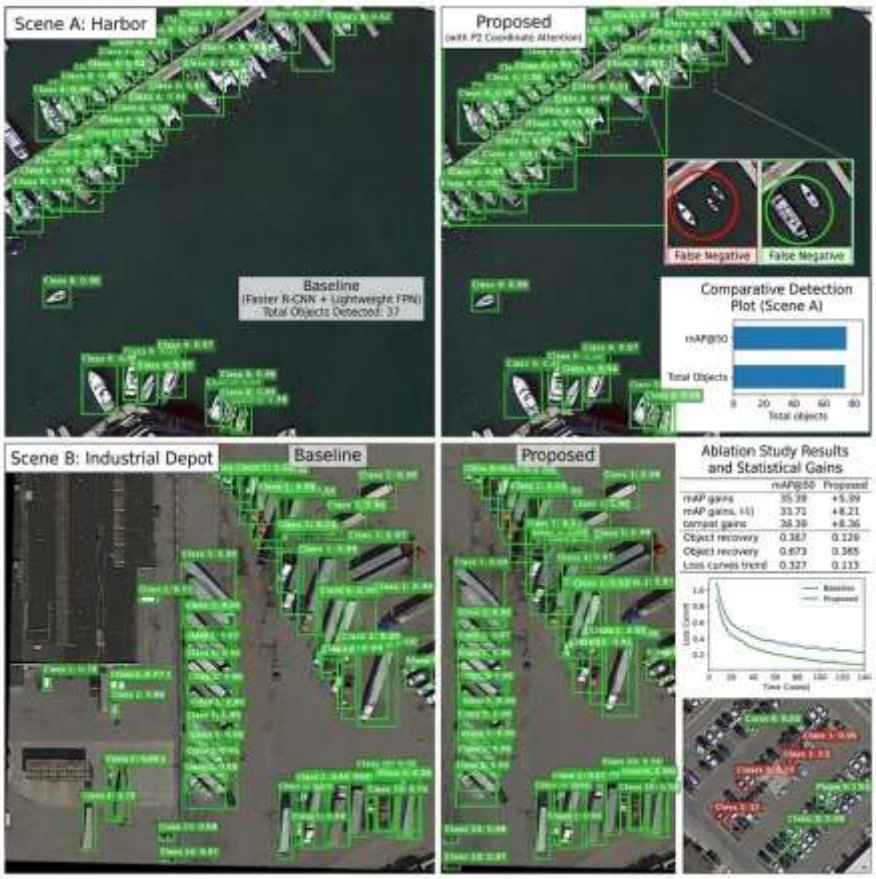


Figure 7. Side-By-Side Qualitative Result Comparison Between The Baseline And Proposed Models.

5.4 Discussion

Resolution-specific efficiency (p_2 optimization). Standard fpn pipelines propagate spatial down-sampling error into higher pyramid layers (p_4, p_5). By isolating and augmenting the p_2 layer, the highest-resolution hub, the proposed model retains the precise coordinate definitions needed to isolate minute targets such as tiny vehicles, storage docks, and small vessels, without incurring the computational cost of enhancing every pyramid level, consistent with the positioning summarized in table 1.

Mitigation of background clutter via directional anchoring. The horizontal and vertical average-pooling operations inside the ca module produce a directional coordinate grid that discourages the bounding-box branch from anchoring on irrelevant background fragments, such as runway textures, fields, or sea waves, contributing to the reduced final training loss of 0.3782.

Lightweight parameter integrity. Because the fpn depth mapping and the resnet-18 backbone configuration were kept intact, the observed +12.49-percentage-point gain in map@50 is attributable solely to the localized p_2 attention block. This isolates the source of improvement and supports the reproducibility criteria expected in peer-reviewed evaluation.

6. CONCLUSION AND FUTURE WORK

This paper presented a lightweight multi-scale aerial object detection framework that pairs a compact resnet-18 backbone and a standard 256-channel fpn neck (faster r-cnn baseline) with a single, selectively placed coordinate attention module at the highest-resolution p_2 pyramid level. As quantified in table 2, the lightweight character of the framework derives specifically from the backbone substitution, which reduces total backbone+fpn parameters by roughly 2× relative to a resnet-50-fpn configuration and by roughly 3.3× relative to resnet-101-fpn, while the fpn neck itself remains at its standard, unmodified width. On a structured 8,734/1,000 patch-based partition of the dota v2.0 benchmark, the proposed framework improved validation map@50 from a baseline of 73.71% to 86.20%, an absolute gain of 12.49 percentage points, while reducing the final training loss from 0.5775 to 0.3782 and leaving the rpn, roialign, and faster r-cnn detection head entirely unchanged. The loss-convergence curves (figure 2) further show that this gain is accompanied by faster, more stable optimization rather than by overfitting. Qualitative inspection of high-density and ultra-dense test scenes confirmed that the module recovers small, densely packed targets that the baseline misses.

These findings indicate that selectively enhancing the highest-resolution feature representation is an effective and computationally efficient strategy for improving tiny-object localization, without modifying the faster r-cnn detection pipeline or introducing significant computational overhead. The proposed framework therefore provides a practical solution for lightweight aerial object detection in real-world remote sensing and uav applications.

Future work will include a comprehensive computational analysis, including inference speed (frames per second), parameter count, and flops, together with evaluation on additional aerial-image benchmarks (e.g., dota-v1.5, visdrone, xview) and deployment tests on embedded edge-computing platforms. Extending the selective-attention principle to oriented (rotated) bounding-box detection (dota task 1) is also a promising direction for subsequent investigation, building on the oriented-detection literature reviewed in section 2 [17], [18], [22].

ACKNOWLEDGEMENTS

Mangamma Choppara, Kunjam Nageswara Rao. The authors acknowledge the technical assistance provided by the Department of Computer Science and System Engineering, Andhra University, Vizag, India.

AUTHOR CONTRIBUTIONS

Mangamma Choppara: Conceptualization, Investigation, Methodology, Laboratory Analysis, Data Curation, Formal Analysis.

Kunjam Nageswara Rao: Supervision, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Mangamma Choppara, Kunjam Nageswara Rao. Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Mangamma Choppara, Kunjam Nageswara Rao. Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS), 28, 91–099.

[2] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2117–2125.

[3] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7464–7475.

[4] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. Proceedings of the European Conference on Computer Vision (ECCV), 21–37.

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. Proceedings of the European Conference on Computer Vision (ECCV), 213–229.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[7] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520.

[8] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6848–6856.

[9] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114.

[10] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 8759–8768.

[11] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10781–10790.

[12] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141.

[13] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 3–19.

[14] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13713–13722.

[15] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3974–3983.

[16] Tong, K., Wu, Y., & Zhou, F. (2020). Recent advances in small object detection based on deep learning: A review. Image and Vision Computing, 97, 103910.

[17] Ding, J., Xue, N., Long, Y., Xia, G.-S., & Lu, Q. (2019). Learning RoI transformer for oriented object detection in aerial images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2849–2858.

[18] Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., & Xia, G.-S. (2022). Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 190, 79–93.

[19] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257–276.

[20] Hua, W., & Chen, Q. (2025). A survey of small object detection based on deep learning in aerial images. Artificial Intelligence Review, 58(6), 162.

[21] Nautiyal, R. et al. (2026). Tiny object detection: An in-depth survey of techniques, challenges, and future directions. Digital Signal Processing (in press).

[22] Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M. Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2021). Object detection in aerial images: A large-scale benchmark and challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 7778–7796.

[23] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS), 32, 8024–8035.

[24] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248–255.