

PRNet: Original Information Is All You Have

PeiHuang Zheng, Yunlong Zhao, Zheng Cui, Yang Li*

Nanjing University of Aeronautics and Astronautics
{bluce.8185,zhaoyunlong,cuizheng,liyangnuaa}@nuaa.edu.cn

Abstract

Small object detection in aerial images suffers from severe information degradation during feature extraction due to limited pixel representations, where shallow spatial details fail to align effectively with semantic information, leading to frequent misses and false positives. Existing FPN-based methods attempt to mitigate these losses through post-processing enhancements, but the reconstructed details often deviate from the original image information, impeding their fusion with semantic content. To address this limitation, we propose PRNet, a real-time detection framework that prioritizes the preservation and efficient utilization of primitive shallow spatial features to enhance small object representations. PRNet achieves this via two modules: the Progressive Refinement Neck (PRN) for spatial-semantic alignment through backbone reuse and iterative refinement, and the Enhanced SliceSamp (ESSamp) for preserving shallow information during downsampling via optimized rearrangement and convolution. Extensive experiments on the VisDrone, AI-TOD, and UAVDT datasets demonstrate that PRNet outperforms state-of-the-art methods under comparable computational constraints, achieving superior accuracy-efficiency trade-offs.

Code — <https://github.com/hhao659/PRNet>

Introduction

Small object detection in aerial imagery has become increasingly important in remote sensing and computer vision, enabling critical applications such as traffic monitoring (2018), rescue operations (2022), and precision agriculture (2022). These applications often require real-time inference on resource-limited edge devices while maintaining high accuracy for objects that occupy very few pixels and are challenging to discern.

Detecting small objects in aerial images is fundamentally difficult due to their extremely limited pixel representation and complex, cluttered backgrounds. Unlike natural scene detection, where objects typically occupy substantial portions of the image, aerial objects are exceptionally small—often under 32×32 pixels and occupying merely 0.1% to 1% of the total image area (2014). As demonstrated in Figure 1, when image resolution decreases (from

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

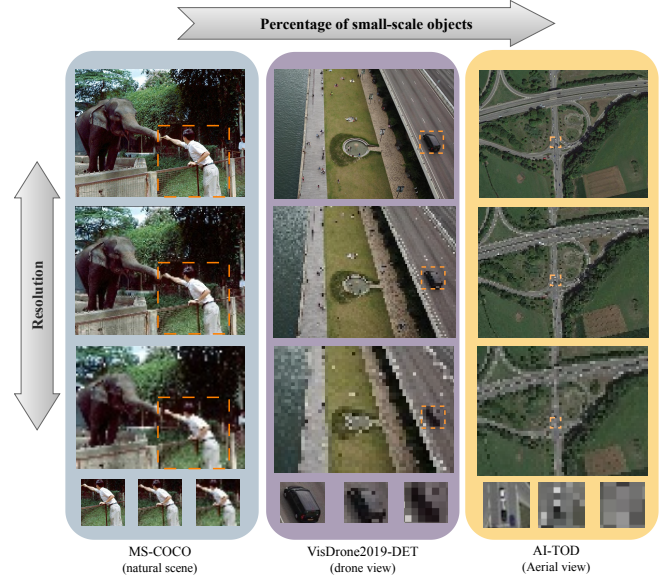


Figure 1: Comparative Analysis of Resolution Degradation on Object Visibility Across Datasets. Comparison of object visibility degradation across MS-COCO, VisDrone, and AI-TOD at original, 160×160 , and 80×80 resolutions (top to bottom). Small objects exhibit greater impact from losses in edges, textures, and shapes during degradation.

the original resolution to 160×160 and 80×80), small objects suffer catastrophic information loss compared to larger objects. This phenomenon mirrors the information degradation that occurs during the model’s forward propagation (2021; 2025), where the loss of shallow spatial details leads to semantic mismatches and, consequently, increased rates of false positives and missed detections (2025; 2025).

Contemporary object detection models comprise a backbone, neck, and head, with the neck—commonly known as the Feature Pyramid Network (FPN, 2017)—serving as the primary architecture for multi-scale feature aggregation (2021). FPN have become a foundational framework for multi-scale object detection due to their ability to aggregate features across different resolutions. Nevertheless, traditional FPNs are suboptimal for small object detection in

aerial imagery. This limitation stems from their reliance on fusing features that have already undergone multiple convolutional and sampling operations, resulting in substantial loss of high-resolution spatial details critical for identifying small objects.

Recent FPN variants (2025; 2025; 2025) primarily focus on enhanced feature fusion or additional refinement modules. However, these approaches do not adequately mitigate the progressive information degradation that accumulates during feature extraction, particularly in its early stages, leading to an irreversible loss of fine-grained details that cannot be fully restored through upsampling or conventional feature fusion. Specifically, FPN-based methods face two critical challenges: (1) Underutilization of shallow features: High-resolution shallow features (e.g., at the P_2 level) are typically used only once, leading to a permanent loss of spatial information crucial for small object discrimination. (2) Feature misalignment: Difficulty in fully integrating shallow spatial features with deeper semantic representations for small targets, thereby resulting in feature mismatches that reduce detection effectiveness.

To address these challenges, preserving and effectively utilizing high-resolution information from initial processing is critical. We propose PRNet, a novel framework tailored for aerial small object detection. First, to maximize the use of preserved shallow features, PRNet introduces the Progressive Refinement Neck (PRN), which iteratively refines high-resolution features through multi-stage backbone reuse, ensuring robust small object representation while alleviating feature misalignment. PRN is flexible and can be integrated into various detection frameworks. Second, to mitigate detail loss during downsampling, PRNet employs the Enhanced SliceSamp (ESSamp) module, which optimizes spatial rearrangement and enhances depthwise convolution for superior feature preservation.

Our contributions can be summarized as follows:

- We reveal key limitations of existing FPN-like methods, specifically information degradation and feature misalignment, which make them unsuitable for aerial image datasets.
- We design a novel neck architecture, PRN, which achieves efficient high-resolution detail retention through multi-stage feature reuse and progressive fusion. In addition, we develop an enhanced downsampling module, ESSamp, to improve the preservation of shallow spatial information.
- Experimental results show that our proposed PRNet significantly surpasses state-of-the-art methods, achieving superior detection accuracy while maintaining competitive efficiency.

Related Works

Small Object Detection

Object detection in aerial images is a representative small object detection task and has consistently posed a challenge. FFCA-YOLO (2024a) proposes a context-aware detection framework for remote sensing images, enhancing the

model’s ability to perceive semantic context. SFFEF-YOLO (2025) introduces a fine-grained feature extraction module to replace standard convolutions, aiming to reduce information loss during the sampling process. FBRT-YOLO (2025) incorporates a Feature Complementary Apping Module and a Multi-Kernel Perception Unit to improve semantic alignment and multi-scale object perception, achieving a better trade-off between detection accuracy and efficiency. Nevertheless, high-precision real-time detection of small objects remains a challenging task.

Feature Pyramid Networks

Feature Pyramid Networks (FPNs) are the dominant architecture for multi-scale detection. The original FPN design integrates deep semantic features with shallow spatial features via top-down pathways and lateral connections. Subsequent improvements, such as PANet (2018), which adds bottom-up pathways, and BiFPN (2020), which employs weighted bidirectional fusion, enhance integration efficiency. For small objects, DSP-YOLO (2024b) introduces a lightweight, detail-sensitive DsPAN, while E-FPN (2025) enhances semantic and fine-grained details bidirectionally. However, the issue of detail loss cannot be alleviated. Unlike these methods, which refine fusion post-extraction, our PRN iteratively reuses backbone features and applies progressive fusion to preserve high-resolution details, minimizing information loss for small object detection.

Feature-Preserving Downsampling

The ability of downsampling methods to retain critical information plays a vital role in overall model performance, especially for tiny objects. Content-Adaptive Downsampling (2023) attempts to preserve key regions during subsampling, but relies heavily on accurate importance masks, which are difficult to generate for small objects in complex aerial scenes. SliceSamp (2023) leverages spatial slicing and depthwise separable convolutions to improve computational efficiency while better preserving feature information. Diff-Stride (2023) introduces learnable stride parameters to adaptively control resolution loss, but its increased model complexity limits deployment in resource-constrained environments. While these methods have made notable progress, effectively preserving detailed features during downsampling remains a significant challenge.

Methodology

In this section, we present a detection framework built upon YOLO11 (2024), named PRNet. PRNet’s design follows a cohesive pipeline: first, ESSamp optimizes downsampling in the backbone to preserve shallow spatial details; second, PRN iteratively refines these features in the neck through backbone reuse and progressive fusion. Figure 2 illustrates the overall architecture where PRN replaces traditional PAN-FPN and ESSamp replaces the first two stride convolutions in the backbone. ESSamp complements PRN by ensuring high-quality shallow inputs for reuse.

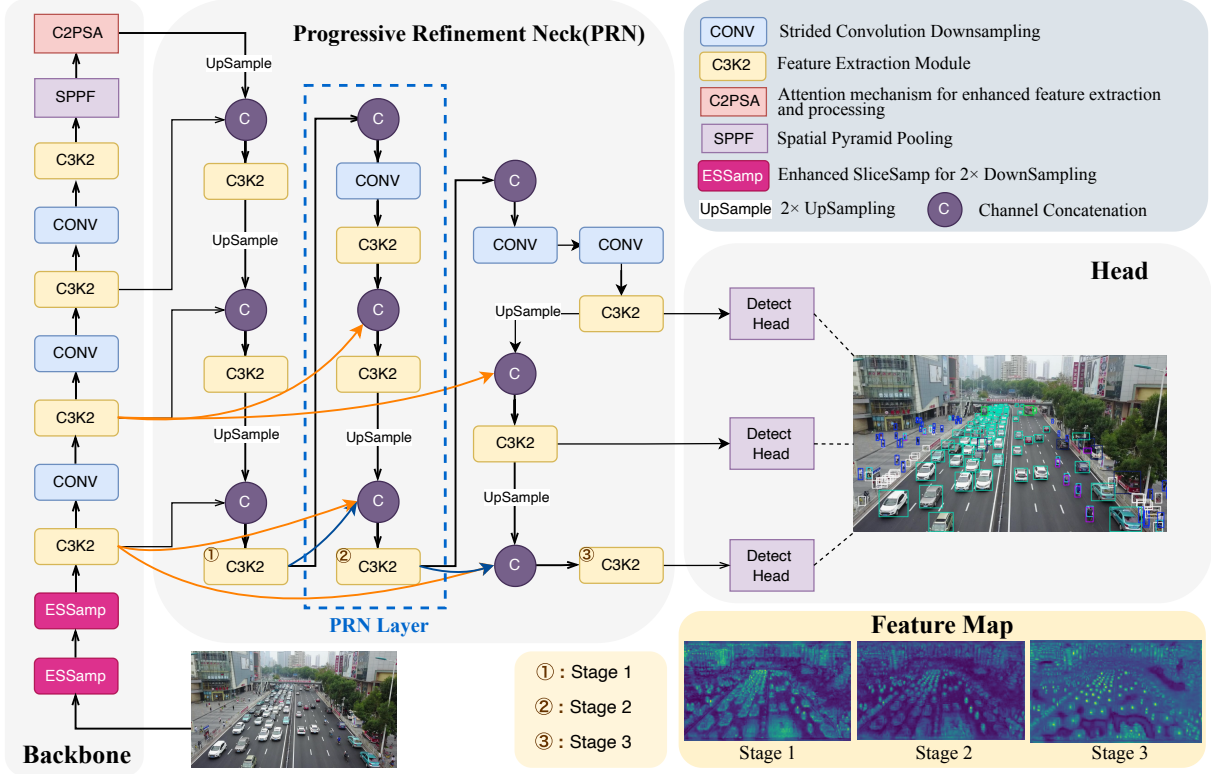


Figure 2: **Architecture of Progressive Refinement Network.** Using YOLO11 as the baseline model, we replace PAN-FPN with our proposed PRN and replace traditional stride convolution downsampling with the proposed ESSamp in the first two layers of the network. The bottom left shows comparisons of feature APs at different stages, demonstrating that the feature quality improves as the number of stages increases.

Progressive Refinement Neck

Traditional FPN-based methods suffer from insufficient utilization of high-resolution backbone features. As illustrated in Figure 3, shallow features containing critical spatial details are typically used only once during fusion. This single-use pattern limits the exploitation of preserved high-resolution information, potentially leading to suboptimal feature representations for small object detection. To address this limitation, we propose the Progressive Refinement Neck (PRN). This module maximizes the retention of original high-resolution details through multi-stage backbone feature reuse, enabling iterative refinement to fully exploit detailed information for enhanced small object representation. The detailed structure of PRN is shown in Figure 2, which utilizes a backbone feature reuse mechanism and progressive fusion strategy. The implementation of this module is detailed below.

Initial Feature Fusion. PRN begins with standard top-down fusion as in PAN-FPN to establish initial spatial-semantic integration:

$$P_i^{td} = \text{Conv}\{\text{Concat}(\text{Resize}(P_{i+1}), P_i^{in})\}, \quad i \in \{2, 3, 4\} \quad (1)$$

where Resize denotes upsampling or downsampling operations for resolution matching, and Conv represents convo-

lution operations for feature processing (using 3×3 kernels). This initial fusion maintains compatibility with the baseline YOLO11 while establishing a foundation for subsequent refinement stages.

Backbone Feature Reuse Mechanism. We noticed the information value of backbone features: The shallow features P_2^{in} and P_3^{in} from the backbone network contain relatively unprocessed original detail information, which is discarded after single use in traditional FPN, causing information waste. To this end, PRN compensates for information dilution in traditional fusion through multi-stage backbone feature reuse. Specifically, PRN downsamples the top-down fused P_2^{td} , then concatenates it with the unused backbone feature P_3^{in} , reintroducing mid-level details; subsequently, the result is upsampled and concatenated with the similarly pristine backbone feature P_2^{in} , maximizing the utilization of high-resolution original details. As shown in Equations below:

$$P_3^{td1} = \text{Conv}\{\text{Concat}(\text{Resize}(P_2^{td}), P_3^{in})\} \quad (2)$$

$$P_2^{refine1} = \text{Conv}\{\text{Concat}(\text{Resize}(P_3^{td1}), P_2^{td}, P_2^{in})\} \quad (3)$$

After initial feature fusion and subsequent processing, the fused features contain sufficient semantic information. Excessive up-and-down sampling operations would dilute the

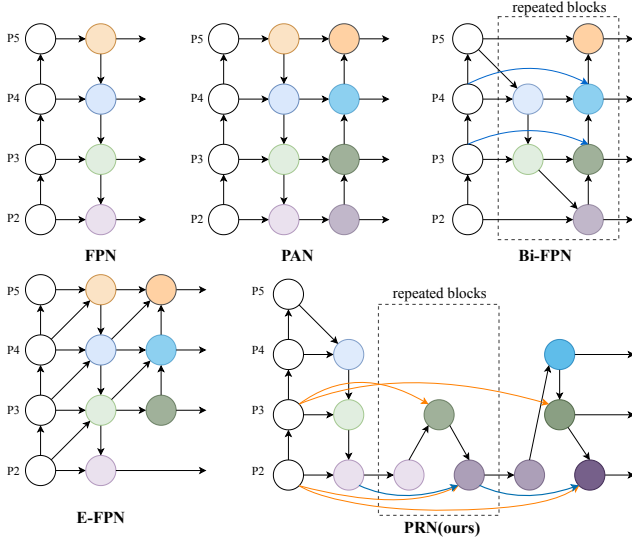


Figure 3: **Comparison of PRN and Traditional FPN Architectures.** PRN enables backbone feature reuse (orange lines) and progressive fusion (blue lines) for iterative high-resolution feature refinement.

scarce spatial information critical for small object detection. Therefore, we employ a single downsample-upsampling strategy to maximize spatial detail preservation while maintaining computational efficiency.

Progressive Fusion Strategy. PRN integrates refined features from earlier stages (e.g., P_2^{td}) into subsequent computations, thereby guiding the refinement process and preventing indiscriminate feature fusion by introducing contextual constraints. This progressive design enables high-resolution features to be iteratively enhanced across multiple stages while remaining consistent with deeper semantic representations. As illustrated by the blue connections in Figure 3, these progressive links form a gradually optimizing closed loop. To ensure efficient refinement, the progressive fusion process is structured into repeated blocks, where each block reuses backbone features and performs a downsampling-upsampling cycle. This design guarantees that each refinement stage receives original feature inputs from the backbone, mitigating recursive information decay and allowing semantic representations to be progressively enriched while high-resolution details are preserved.

Output Generation. To generate output features suitable for three detection scales, PRN performs structured processing on the final refined features, as shown in Equations (4)-(6):

$$P_4^{out} = Conv\{Resize(P_2^{refine^i})\} \quad (4)$$

$$P_3^{out} = Conv\{Concat(Resize(P_4^{out}), P_3^{in})\} \quad (5)$$

$$P_2^{out} = Conv\{Concat(Resize(P_3^{out}), P_2^{refine^i}, P_2^{in})\} \quad (6)$$

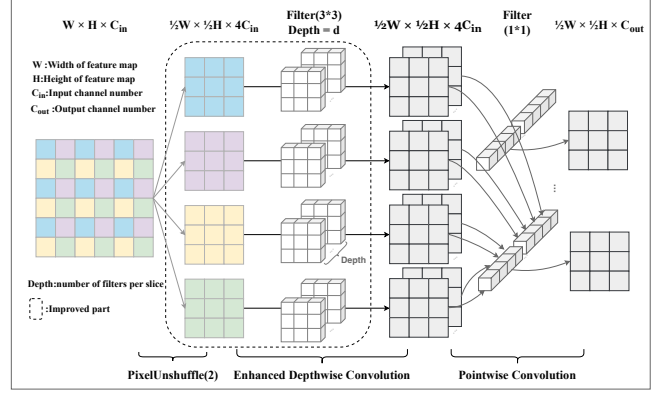


Figure 4: **ESSamp Module Structure.** Utilizes PixelUnShuffle for efficient spatial rearrangement and augmented depthwise convolution (depth multiplier $d=2$) to enhance feature expressiveness, preserving fine-grained details for small object detection.

Enhanced SliceSamp

While PRN maximizes the utilization of preserved features, effective downsampling is essential to ensure high-quality inputs from the backbone; to this end, we introduce ESSamp. Information loss during downsampling operations substantially impacts the quality of shallow backbone features, which are critical for effective feature reuse in small object detection. Conventional downsampling methods, such as strided convolution, often result in severe loss of fine-grained details. Meanwhile, existing detail-preserving approaches, such as SliceSamp, suffer from computational inefficiency and limited feature expressiveness. To overcome these limitations, we introduce the Enhanced SliceSamp (ESSamp) module. ESSamp enhances feature representation through improved depthwise convolution and further optimizes the spatial rearrangement process to increase computational efficiency. The overall structure of ESSamp is illustrated in Figure 4, and its design is described in detail below.

Enhanced Feature Expression. The primary limitation of existing detail-preserving downsampling methods lies in their insufficient feature modeling capability. Standard depthwise convolution in SliceSamp uses only a single kernel per input channel, severely limiting the ability to capture complex local patterns essential for small object discrimination. To overcome this bottleneck, we introduce enhanced depthwise convolution with depth multiplier d , which assigns multiple kernels to each input channel to enrich local feature representation.

This design is supported by empirical analysis and receptive field considerations. By introducing the depth multiplier d , the feature expressiveness is enhanced, with its impact validated through ablation studies (e.g., Table 7). For instance, when $d = 2$, the capacity for local structure modeling is effectively improved. Such an enhancement is particularly crucial for small objects, where discriminative information is extremely limited and subtle local patterns are

Model	Size	AP ₅₀	AP	Params	FLOPs
YOLOv8-s (2024)	640	39.6	23.6	11.2 M	28.6 G
YOLO11-s (2024)	640	40.4	24.2	9.4 M	21.3 G
FBRT-YOLOv8-s (2025)	640	41.7	25.5	2.9 M	23.1 G
PRNet-N(Ours)	640	43.4	26.7	2.2 M	17.8 G
YOLOv8-m (2024)	640	44.0	26.9	25.8 M	78.4 G
YOLO11-m (2024)	640	46.1	28.6	20.1 M	68.0 G
FBRT-YOLOv8-x (2025)	640	47.3	29.6	23.2M	187.1 G
PRNet-YOLOv8-s(Ours)	640	50.4	31.3	8.2 M	55.5 G
PRNet(Ours)	640	49.9	31.1	7.77 M	44.9 G
EMA attention [†] (2023)	640	49.7	30.4	91.18 M	315 G
yolov9c [†] (2024)	640	47.6	29.3	25.3 M	239.9 G
HV-SwinViT [†] (2025)	640	43.63	26.3	64 M	523 G
PRNet-L(Ours)	640	54.1	34.4	24.6 M	196.8 G
Larger Input Size					
DQ-DETR [†] (2024)	800×1333	60.9	37.0	58 M	904 G
HV-SwinViT [†] (2025)	1280	52.5	35.6	91.8 M	-
PRNet-L(Ours)	1024	61.0	38.3	24.6 M	505 G

Table 1: Comparison with state-of-the-art models under different resource constraints on VisDrone-Validation dataset. ”-” indicates that no data were available for this item. **Bold** indicates the best results. Results marked with [†] are reported from the original papers, while the others are reproduced by us under the same experimental settings.

essential for reliable detection.

Improved Spatial Rearrangement. In addition to the core feature enhancement, we also improve the spatial rearrangement process in SliceSamp to boost computational efficiency. SliceSamp’s explicit indexing operations(e.g., $X = \text{Concat}(X_{in}[:, :, i :: 2, j :: 2]), i, j \in \{0, 1\})$ incur high memory access overhead and cannot fully utilize GPU parallel computing capabilities. We replace these explicit operations with PixelUnShuffle, which improves memory coalescing and provides a constant-factor reduction in runtime while maintaining the detail-preserving property, as demonstrated in efficient sub-pixel convolutional networks (Shi et al. 2016). as shown in Equations below:

$$X = \text{PixelUnShuffle}(2, X_{in}) \quad (7)$$

$$Y = \text{GELU}(\text{BN}_2(W_2^{PW} * \text{GELU}(\text{BN}_1(W_1^{EDW} \odot X)))) \quad (8)$$

where $W_1^{EDW} \in \mathbb{R}^{4dC \times 4C \times 3 \times 3}$ is the expanded depthwise convolution kernel with groups=4C and output channels of 4dC, $d = 2$; $W_2^{PW} \in \mathbb{R}^{C_{out} \times 4dC \times 1 \times 1}$ is the pointwise convolution kernel; d is the depth multiplier; $*$ denotes standard convolution, $\odot$ denotes Depthwise Convolution.

Compared to traditional SliceSamp, ESSamp maintains the advantage of high-fidelity downsampling while significantly improving computational efficiency and feature expression capability, providing higher-quality feature input for PRN and achieving an optimized balance between detail preservation and computational efficiency. Channel expansion from C to $4dC$ significantly enhances feature expression but also increases the computational burden of depthwise convolution. The subsequent pointwise convolution needs to compress $4dC$ back to the object channel number with a compression ratio of 4d:1, posing a risk of information bottleneck. In subsequent experiments, we conduct

Model	AP ₅₀	AP	Params	FLOPs
YOLOX-M [†] (2021)	30.5	17.8	25.3 M	73.75 G
YOLOv7 [†] (2023)	39.2	21.3	36.5 M	103.3 G
YOLOv8-M (2024)	35.6	21.2	25.6 M	78.7 G
DCYOLO-M [†] (2025)	38	22.3	33 M	117 G
PRNet(Ours)	40.3	24.2	7.77 M	44.9 G

Table 2: Comparison of PRNet with latest methods on VisDrone test-dev dataset.

ablation studies on the hyperparameter d to achieve an optimal balance between feature expression capability and computational overhead. This integration of ESSamp with PRN forms a comprehensive pathway for preserving and utilizing original information, enhancing overall detection performance in aerial imagery.

Experiments

Implementation Details

We conduct comprehensive experiments on three aerial image object detection benchmarks: VisDrone (2019), AI-TOD (2022), and UAVDT (2018). Experiments are conducted on RTX 3090 GPU. Our network is trained for 400 epochs using the stochastic gradient descent (SGD) optimizer with a momentum of 0.937, a weight decay of 0.0005, a batch size of 8, a patience of 50, and an initial learning rate of 0.01. All other configurations remain the same as in YOLO11.

Results on Visdrone Dataset

Comparison with State-of-the-art Methods. As shown in Table 1, we conduct comprehensive comparisons be-

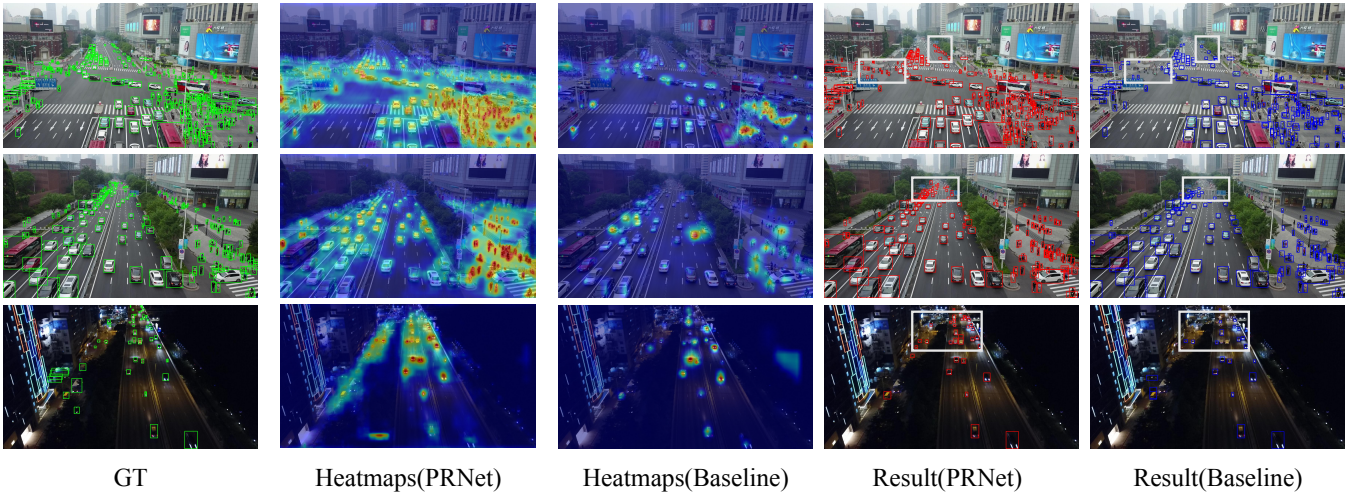


Figure 5: **Visualization of the detection results and heatmaps on VisDrone.** The highlighted areas represent the regions that the network is focusing on.

tween PRNet and existing state-of-the-art detection methods on the VisDrone validation dataset. The experimental results demonstrate that PRNet exhibits superior performance advantages across different resource constraints. For lightweight models, PRNet-N achieves 43.4% AP_{50} and 26.7% AP with only 2.2M parameters and 17.8G FLOPs, compared to YOLO11-s, it improves detection accuracy by 3.0% AP_{50} and 2.5% AP respectively while reducing parameters by 76.6%. Compared to FBRT-YOLO-S with similar parameter count, AP_{50} and AP are improved by 1.0% and 0.8% respectively. For medium-scale models, we present two variants: PRNet (based on YOLO11) achieves 49.9% AP_{50} and 31.1% AP with 7.77M parameters and 44.9G FLOPs, while PRNet-YOLOv8-s (based on YOLOv8-s backbone) achieves even better performance of 50.4% AP_{50} and 31.3% AP with 8.2M parameters and 55.5G FLOPs, demonstrating that even with the same YOLOv8 backbone as FBRT-YOLOv8-X, our framework achieves superior performance (50.4% vs. 47.3% AP_{50}); furthermore, as shown in Table 8, PRN can be effectively integrated into FBRT-YOLO, further improving its detection performance. Compared to YOLO11-m, it improves detection accuracy by 3.8% AP_{50} and 2.5% AP respectively while reducing parameters by 61.3% and computational cost by 34.0%. For large-scale models, PRNet-L achieves the best accuracy of 54.1% AP_{50} and 34.4% AP, comprehensively outperforming all comparison methods, validating the detection advantages of our approach in complex aerial scenarios.

Generalization Validation. To further validate the generalization capability of PRNet, we conducted comparisons with the latest methods on the VisDrone test-dev dataset, with results shown in Table 2. PRNet achieves 40.3% AP_{50} and 24.2% AP, outperforming YOLOv8-M (2024) by 4.7% AP_{50} and 3.0% AP, and DCYOLO-M (2025) by 2.3% AP_{50} and 1.9% AP. These results validate the robust stability and superiority of PRNet in aerial image detection.

Qualitative Results. To further illustrate the superior per-

Method	Size	AP	AP_{50}	AP_{75}
FFCA-YOLO [†] (2024a)	640	27.7	61.7	22.3
DQ-DETR [†] (2024)	800×1333	30.2	68.6	22.3
HS-FPN [†] (2025)	800×800	25.1	55.7	22.3
HV-SwinVit [†] (2025)	1280	<u>32.1</u>	62.3	<u>29.4</u>
PRNet(Ours)	640	30.3	61.4	28.1
PRNet-L(Ours)	640	35.6	<u>67.8</u>	33.1

Table 3: Comparison of PRNet with advanced methods on AI-TOD test dataset. **Bold** indicates the best-performing method, and underline indicates the second-best method.

Method	Size	AP	AP_{50}	AP_{75}
YOLO11-S (2024)	640	19.1	31.4	20.9
YOLC [†] (2024)	1024×640	19.3	30.9	20.1
AD-Det [†] (2025)	1024×540	<u>20.1</u>	34.2	<u>21.9</u>
PRNet(Ours)	640	20.8	<u>32.3</u>	23.8

Table 4: Comparison of PRNet with advanced methods on UAVDT dataset.

formance of PRNet in detecting small objects in aerial images, we present visualizations of heatmaps and detection results in Figure 5. The heatmaps show PRNet’s enhanced focus on small and densely packed objects compared to the baseline, while the detection results demonstrate more precise localization aligned with ground truth.

Results on AI-TOD Dataset

The AI-TOD dataset contains a large number of extremely small objects and high-density scenes, which impose higher demands on the fine-grained feature extraction capabilities of detection algorithms. To further validate the superiority of our method in remote sensing small object detec-

PRN	ESSamp	AP ₅₀	AP	Params	FLOPs
-	-	39.0	23.3	9.4 M	21.3 G
✓	-	49.3	30.4	7.71 M	41.1 G
-	✓	40.1	24.1	9.49 M	24.8 G
✓	✓	49.8	31.1	7.77 M	44.9 G

Table 5: Ablation study of the proposed method on VisDrone.

PRN layer	AP ₅₀	AP	Params	FLOPs
-	39.0	23.3	9.4 M	21.3 G
0	45.0	27.6	7.05 M	28.7 G
1	49.3	30.4	7.71 M	41.1 G
2	51.0	31.8	8.4 M	54.3 G
3	51.4	32.2	9.1 M	67.5 G

Table 6: Ablation study on progressive refinement iterations.

tion, we evaluated PRNet and PRNet-L on the AI-TOD test set. As shown in Table 3, at a compact 640×640 resolution, PRNet achieves 30.3% AP, 61.4% AP₅₀, and 28.1% AP₇₅, outperforming most benchmarks. PRNet-L sets new records with 35.6% AP and 33.1% AP₇₅, surpassing DQ-DETR (2024) and HV-SwinVit (2025) despite their larger input sizes. These results underscore the superior accuracy and efficiency of our framework in remote sensing scenarios.

Results on UAVDT Dataset

Table 4 presents the comparison results on the UAVDT dataset. Our proposed method surpasses existing methods, such as YOLC(2024) and AD-Det(2025). Utilizing a smaller input size of 640, PRNet achieves 20.8% AP, 32.3% AP₅₀, and 23.8% AP₇₅, outperforming other state-of-the-art methods in AP and AP₇₅ despite AD-Det’s larger resolution of 1024×540. This demonstrates the effectiveness of our detection framework.

Ablation experiments

We conduct ablation experiments on the VisDrone dataset using YOLO11-S as the baseline to validate PRNet’s core components.

Effect of Key Components. As shown in Table 5, the baseline YOLO11s achieves 39.0% AP₅₀ and 23.3% AP. Adding PRN alone improves AP₅₀ by 10.3% to 49.3% and AP by 7.1% to 30.4%, while reducing parameters by 18% (9.4M to 7.71M). Using ESSamp alone yields modest gains (40.1% AP₅₀, 24.1% AP). Combining PRN and ESSamp achieves the best performance (49.8% AP₅₀, 31.1% AP), with 7.77M parameters and 44.9G FLOPs, demonstrating their synergistic effect. Although PRNet increases FLOPs by 110.7% (21.3G to 44.9G) compared to the baseline, this computational overhead is strategically justified: PRN’s iterative refinement operates primarily on high-resolution features where small objects reside, directly translating increased computation into substantial accuracy gains (10.8% AP₅₀,

Baseline	Depth	AP ₅₀	AP	Params	FLOPs
PRNet	-	49.3	30.4	7.71 M	41.1 G
	1	48.6	30.2	7.75 M	42.9 G
	2	49.8	31.1	7.77 M	44.9 G
	3	49.4	30.6	7.82 M	46.9 G

Table 7: Ablation study on different depths of ESSamp on VisDrone.

Method	PRN	AP ₅₀	AP	Params	FLOPs
YOLOv5s	-	40.3	23.9	9.1 M	23.8 G
	✓	47.4	29.2	6.7 M	42.7 G
YOLOv5m	-	44.4	27.1	25.9 M	78.9 G
YOLOv8s	-	40.5	24.4	11.1 M	28.7 G
	✓	48.3	30.2	8.3 M	54.1 G
YOLOv8m	-	44.0	26.9	25.8 M	78.4 G
YOLO11s	-	39.0	23.3	9.4 M	21.3 G
	✓	49.3	30.4	7.71 M	41.1 G
YOLO11m	-	46.1	28.6	20.1 M	68.0 G
FBRT-YOLOv8-s	-	41.7	25.5	2.9 M	23.1 G
	✓	47.7	29.2	2.2 M	40.4 G
FBRT-YOLOv8-m	-	45.9	28.4	7.2 M	58.7 G
RT-DETR-R50	-	28.9	16.1	42 M	136 G
	✓	32.1	18.3	39.9 M	176.4 G
RT-DETR-R101	-	31.5	17.8	60.9 M	186.3 G

Table 8: Ablation study of PRN on different YOLO detectors on VisDrone. “-” indicates PRN method is not used.

7.8% AP), while our overall framework maintains superior efficiency compared to state-of-the-art methods—for instance, achieving comparable accuracy to YOLO11-m (46.1% AP₅₀) while using 34.0% fewer FLOPs (44.9G vs. 68.0G), demonstrating an advantageous accuracy-efficiency trade-off for aerial small object detection.

Effect of Progressive Refinement Stages. Table 6 shows that increasing PRN iterations from 0 to 3 improves AP₅₀ from 45.0% to 51.4% and AP from 27.6% to 32.2%. Although further increasing the repetition count can continue to improve accuracy, the computational overhead also grows significantly. Therefore, considering real-time performance, we select 1 repetition as the optimal configuration.

Effect of Depth Multiplier in ESSamp. Table 7 evaluates ESSamp’s depth multiplier. Depth=2 achieves the best performance (49.8% AP₅₀, 31.1% AP), improving over depth=1 (equivalent to SliceSamp) by 1.2% AP₅₀ and 0.9% AP. Further increasing depth to 3, although theoretically enhancing feature expression further, causes the dramatic increase in channel numbers to require larger compression ratios in subsequent pointwise convolutions, leading to key information loss and reduced detection accuracy. Therefore, we select depth=2 as the optimal configuration for ESSamp.

Effect of PRN Generalizability. Table 8 validates PRN’s versatility across YOLOv5s, YOLOv8s, YOLO11s, FBRT-YOLOv8-s and RT-DETR-R50. After introducing PRN to all tested detectors, detection accuracy is significantly im-

proved. YOLO11s+PRN achieves the highest gains, improving AP₅₀ by 10.3% and AP by 7.1% while reducing parameters by 18%. Notably, when applying PRN to the state-of-the-art FBRT-YOLOv8-s baseline (41.7% AP₅₀), our method achieves 47.7% AP₅₀ and 29.2% AP—a substantial improvement of 6.0% AP₅₀ and 3.7% AP—while reducing parameters from 2.9M to 2.2M, demonstrating that PRN can further enhance already optimized architectures. Furthermore, even compared to larger m-series models, lightweight models with PRN can achieve superior detection performance under fewer resource constraints, validating the universality and effectiveness of PRN across different architectures.

Due to page constraints, additional experimental results and visualizations are provided in the **Appendix**.

Conclusion

In this paper, we address the challenge of information loss in small object detection for aerial imagery by proposing PRNet, a novel real-time detection framework that prioritizes the preservation and efficient utilization of original shallow spatial features. The framework introduces two key innovations: the Progressive Refinement Neck (PRN), which enables multi-stage backbone feature reuse and iterative refinement for enhanced spatial-semantic alignment, and the Enhanced SliceSamp (ESSamp), which optimizes downsampling through improved spatial rearrangement and augmented depthwise convolution to minimize detail degradation. Extensive experiments on the VisDrone, AI-TOD, and UAVDT datasets demonstrate that PRNet achieves superior detection accuracy while maintaining competitive computational efficiency, outperforming state-of-the-art methods across various resource constraints.

References

- Bai, C.; Zhang, K.; Jin, H.; Qian, P.; Zhai, R.; and Lu, K. 2025. SFDEF-YOLO: Small object detection network based on fine-grained feature extraction and fusion for unmanned aerial images. *Image and Vision Computing*, 156: 105469.
- Bian, D.; Tang, M.; Ling, M.; Xu, H.; Lv, S.; Tang, Q.; and Hu, J. 2025. A refined methodology for small object detection: Multi-scale feature extraction and cross-stage feature fusion network. *Digital Signal Processing*, 105297.
- Bouraya, S.; and Belangour, A. 2021. Deep learning based neck models for object detection: A review and a benchmarking study. *International Journal of Advanced Computer Science and Applications*, 12(11).
- Deng, C.; Wang, M.; Liu, L.; Liu, Y.; and Jiang, Y. 2021. Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, 24: 1968–1979.
- Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 370–386.
- Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.
- Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.
- Gu, Q.; Han, Z.; Kong, S.; Huang, H.; Li, Y.; Fan, Q.; and Wu, R. 2025. DCYOLO: Dual negative weighting label assignment and cross-layer decouple head for YOLO in remote sensing images. *Expert Systems with Applications*, 281: 127595.
- He, L.; and Wang, M. 2023. SliceSamp: A promising downsampling alternative for retaining information in a neural network. *Applied Sciences*, 13(21): 11657.
- Hesse, R.; Schaub-Meyer, S.; and Roth, S. 2023. Content-adaptive downsampling in convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4544–4553.
- Hua, W.; and Chen, Q. 2025. A survey of small object detection based on deep learning in aerial images. *Artificial Intelligence Review*, 58(6): 1–67.
- Huang, Y.-X.; Liu, H.-I.; Shuai, H.-H.; and Cheng, W.-H. 2024. Dq-detr: Detr with dynamic query for tiny object detection. In *European Conference on Computer Vision*, 290–305. Springer.
- Kaleem, Z.; and Rehmani, M. H. 2018. Amateur drone monitoring: State-of-the-art architectures, key enabling technologies, and future research directions. *IEEE Wireless Communications*, 25(2): 150–159.
- Khanam, R.; and Hussain, M. 2024. YOLOv11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.
- Li, Z.; He, Q.; and Yang, W. 2025. E-FPN: an enhanced feature pyramid network for UAV scenarios detection. *The Visual Computer*, 41(1): 675–693.
- Li, Z.; Lian, S.; Pan, D.; Wang, Y.; and Liu, W. 2025. AD-Det: Boosting Object Detection in UAV Images with Focused Small Objects and Balanced Tail Classes. *Remote Sensing*, 17(9): 1556.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, C.; Gao, G.; Huang, Z.; Hu, Z.; Liu, Q.; and Wang, Y. 2024. Yolc: You only look clusters for tiny object detection in aerial images. *IEEE transactions on intelligent transportation systems*, 25(10): 13863–13875.
- Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.
- Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; and Huang, Z. 2023. Efficient multi-scale attention module

- with cross-spatial learning. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 1–5. IEEE.
- Rafif, S.; Azhar, M. F.; Wahyu Pratama, M. A. R.; Ibad, A. M.; Yudistira, N.; and Muflikhah, L. 2023. Hybrid of DiffStride and Spectral Pooling in Convolutional Neural Networks. In *Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology*, 210–216.
- Ren, X.; Sun, M.; Zhang, X.; Liu, L.; Zhou, H.; and Ren, X. 2022. An improved mask-RCNN algorithm for UAV TIR video stream target detection. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102660.
- Roy, A. M.; and Bhaduri, J. 2022. Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4. *Computers and Electronics in Agriculture*, 193: 106694.
- Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.
- Shi, Z.; Hu, J.; Ren, J.; Ye, H.; Yuan, X.; Ouyang, Y.; He, J.; Ji, B.; and Guo, J. 2025. HS-FPN: High frequency and spatial perception FPN for tiny object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6896–6904.
- Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790.
- Varghese, R.; and Sambath, M. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, 1–6. IEEE.
- Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475.
- Wang, C.-Y.; Yeh, I.-H.; and Mark Liao, H.-Y. 2024. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, 1–21. Springer.
- Wang, Y.; Li, Z.; Zhu, S.; and Wei, X. 2025. EFCNet for small object detection in remote sensing images. *Scientific Reports*, 15(1): 20393.
- Xiao, Y.; Xu, T.; Xin, Y.; and Li, J. 2025. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8673–8681.
- Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; and 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.
- Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; and Yan, J. 2024a. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.
- Zhang, Y.; Zhang, H.; Huang, Q.; Han, Y.; and Zhao, M. 2024b. DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. *Expert Systems with Applications*, 241: 122669.
- Zhao, X.; Wang, J.; Li, L.; Shao, X.; and Zhang, K. 2025. A unified solution for replacing position embedding in Vision Transformer for object detection. *Engineering Applications of Artificial Intelligence*, 152: 110679.