Enhancing Aerial Pedestrian Detection via High-Resolution P2 Feature Integration in YOLOv12

Anonymous CVPR submission

Paper ID *****

Abstract

Detecting pedestrians in UAV (Unmanned Aerial Vehicle) imagery poses several challenges due to factors such as significant scale variation, low resolution, dense crowds, and cluttered backgrounds. In aerial views, pedestrians often occupy only a few pixels, making them difficult to detect using standard object detection architectures that rely on high-level feature maps. Most modern detectors begin predictions at a stride of 8 resolutions, which limits their ability to detect extremely small objects. In this study, we revisit the feature pyramid architecture of YOLOv12 and introduce a high-resolution P2 detection head to improve supervision at the early stages of the network. Our proposed modification extends the pyramid to a stride of 4 resolution and incorporates bidirectional feature refinement to maintain semantic consistency across different scales. This design remains lightweight and preserves practical inference speed while improving the representation of tiny objects. We evaluate our approach on two aerial pedestrian benchmarks: VisDrone and TinyPerson. The proposed model improves the mean Average Precision (mAP) at IoU 0.5 from 0.63 to 0.69, which represents a 9.5% relative gain on the VisDrone dataset. On the TinyPerson dataset, mAP@0.5 increases from 0.40 to 0.45, indicating a 12.5% relative gain. Additionally, there is a 25% relative increase in the tiny-scale AP50, rising from 0.24 to 0.30. The experimental results demonstrate consistent improvements in detection performance, particularly for small and tiny pedestrians, without significant computational overhead. Ablation studies further confirm that early-resolution detection is crucial in enhancing recall for small objects in UAV imagery. These findings indicate that revisiting the starting level of feature pyramids is a straightforward yet effective strategy for improving small-object detection in aerial scenarios.

1. Introduction

Unmanned aerial vehicles (UAVs) are increasingly utilised in various applications, including traffic monitoring, public safety, search and rescue, and crowd analysis [1, 13]. A critical aspect of these applications is the reliable detection of pedestrians from aerial imagery [16]. However, detecting pedestrians in UAV scenarios poses several challenges, including extreme scale variations, dense scenes, complex backgrounds, and the presence of large numbers of small objects occupying only a few pixels[uavchallenge]. Recent advancements in small object detection have focused on improving multi-scale feature fusion, enhancing attention mechanisms, and designing better localization losses. While these methods show improvements on general benchmarks, they often struggle with aerial pedestrian detection. Pedestrians captured from UAV platforms frequently appear at very small scales, often below the effective receptive field of standard detection heads. Several recent studies have specifically tackled pedestrian detection in UAV imagery by introducing feature enhancement modules, multi-scale attention mechanisms, or context-aware designs. These approaches have shown improved robustness against occlusion and complex backgrounds. However, most existing work employs conventional detection heads that start at P3 (1/8 resolution), thereby overlooking fine-grained spatial information essential for detecting tiny pedestrians. Lightweight and real-time UAV detectors aim to reduce computational costs through techniques such as backbone compression, simplified detection heads, or pruning strategies. Unfortunately, these reductions can lead to decreased performance when identifying extremely small objects. Transformer-based methods, including Swin transformer approaches for aerial detection [3], improve global context modeling through hierarchical attention. However, the use of transformers often increases computational overhead and does not specifically address the loss of early spatial resolution for detecting tiny pedestrians. To tackle these challenges, we propose a high-resolution P2 detection branch integrated into the YOLOv12 framework. By

extending the feature pyramid to include P2 (1/4 scale) and introducing bidirectional refinement, our proposed head preserves early spatial details while maintaining computational efficiency. We evaluate our design on two challenging aerial benchmarks: VisDrone [27] and TinyPerson [25]. Extensive experiments, including size-wise analysis, demonstrate consistent improvements in the detection of tiny and small object categories while maintaining competitive performance across various scales. Our contributions can be summarized as follows:

1. A P2-enhanced multi-scale detection head that extends YOLOv12 with high-resolution supervision for tiny pedestrian detection.
2. A structured bidirectional feature refinement mechanism for preserving high-resolution spatial cues.
3. Comprehensive evaluation on two challenging UAV datasets, including size-wise analysis.

2. Related Work

2.1. Small Object Detection in UAV Imagery

Recent years have seen growing interest in small-object detection in aerial imagery. Many studies focus on improving feature representation through techniques like edge enhancement and context modeling. Early research used cascade architectures with deformable convolutions to enhance the detection of small objects [26]. The work [20] highlights the importance of specialised datasets such as AI-TOD, which reveal the limitations of traditional detectors for tiny objects and suggest targeted learning strategies. For instance, CGD-YOLO [2] introduces global edge propagation and multi-dimensional attention mechanisms to improve the extraction of fine contours. EMSANet [8] further develops edge-enhanced multi-scale alignment to address semantic inconsistencies across different feature levels. Other approaches involve modifications to the backbone and feature pyramid designs. UAVDet [24] incorporates CNN-Mamba hybrid modules and redesigns the pyramid structure specifically for the perception of tiny objects. WCDB-YOLO [12] employs a dual-backbone strategy combined with wavelet-based context modeling and explicitly integrates P2 features into the detection head. Advancements such as High-Resolution Feature Pyramid Networks [4] improve multi-scale feature representation, while QueryDet [22] enhances efficiency by selectively applying high-resolution computation. Lightweight architectures and attention-based modules, like the NOVA [15] module in the RealDroneVision framework, further improve YOLO-based detection Collectively, these studies emphasize the importance of maintaining high-resolution spatial information when dealing with extremely small targets. However, accurately detecting extremely small pedestrians in aerial imagery remains challenging due to limited spatial information, underscoring the need for improved high-resolution detection heads for localising tiny objects.

2.2. UAV Pedestrian detection

Pedestrian detection from UAV platforms has been studied as a specialized sub-problem due to its critical importance for safety. Recent works have introduced techniques such as scale-sensitive feature enhancement [14], attention-guided fusion [9], and context-aware modules [16] specifically designed for aerial pedestrian detection. These methods aim to improve robustness in scenarios involving occlusion and dense crowds. Most UAV detectors focused on pedestrian detection employ a conventional P3-P5 detection head structure. However, this structure does not directly supervise the earliest high-resolution feature maps for detection, which limits the ability to accurately localize extremely small pedestrians.

Lightweight and real-time UAV detectors [7, 23] have been developed to reduce computational complexity while achieving high inference speeds. Unfortunately, they often struggle to maintain strong performance on tiny objects due to reduced feature capacity and limitations in high-resolution modeling. Transformer-based detectors [3] have shown effectiveness in capturing global relationships, but they also introduce significant computational overhead. Recent approaches address this issue by incorporating attention-based detection heads. For example, EHDC-YOLO [5] employs a Dynamic Head (DyHead) with multi-dimensional attention to improve adaptability across varying object scales. FDE-YOLO [10] integrates scale-aware, spatial-aware, and task-aware refinements in its detection head to enhance discrimination in dense scenes. CMA-Net [23] proposes collaborative multi-attention structures for efficient real-time UAV detection. Although attention mechanisms improve feature weighting, many methods still rely on the conventional P3–P5 detection hierarchy, which limits its sensitivity to ultra-small objects. Recent studies have focused on attention mechanisms and transformer-based approaches to improve pedestrian detection across scales. For example, CoSTAA [18] uses a convolutional Swin Transformer with channel and spatial attention to enhance feature interaction in complex environments. Similarly, MECSA [19] incorporates a multi-scale attention module for better feature representation. While these methods improve detection performance across a range of object sizes, they often struggle with extremely small pedestrians in aerial imagery because of insufficient high-resolution spatial features.

To better handle tiny objects, several recent works have introduced high-resolution branches. WCDB-YOLO [12] and DPN-YOLO [11] incorporate P2-level features to strengthen small-object localization. MSAD-YOLO [6] reconstructs a shallow detection head to preserve spatial details while maintaining a lightweight design. UAVDet

[24] replaces large-object detection heads with tiny-object-oriented heads to re-balance feature fusion. These approaches demonstrate that early-stage features contain essential spatial cues that should not be discarded through down-sampling.

Beyond architectural designs, several studies have modified bounding box regression losses to better adapt to small-scale targets. Inner-PIoUv2 [2], Wise-IoU variants [16], and composite IoU [21] losses have been proposed to adjust gradient contributions based on object scale and anchor quality. While these strategies improve localization robustness, they do not directly address the structural limitations of the feature pyramid.

All these works enhance feature aggregation and attention modeling, but they still rely on a feature pyramid that begins at a coarse resolution, limiting their ability to detect extremely small pedestrians. To address this gap, we have redesigned the detection head to include a stride-4 prediction branch tailored for small-object detection.

3. Proposed P2-Enhanced Multi-Scale Detection Head

To improve tiny pedestrian detection in aerial imagery, we extend the original YOLOv12n [17] detection head by introducing a high-resolution P2 branch. The proposed design enhances spatial sampling density while maintaining strong semantic representation through bidirectional feature fusion.

3.1. Backbone Feature Representation

Given an input image

$$I \in \mathbb{R}^{H\times W\times 3}, \tag{1}$$

the backbone extracts multi-scale feature maps:

$$\{F_2, F_3, F_4, F_5\}, \tag{2}$$

where

$$F_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_2}, \tag{3}$$

$$F_3 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_3}, \tag{4}$$

$$F_4 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_4}, \tag{5}$$

$$F_5 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_5}. \tag{6}$$

3.1.1. Top-Down Hierarchical Aggregation

To enhance semantic consistency across scales, a top-down fusion strategy is adopted.

First, the P5 feature is upsampled and fused with P4:

$$\tilde{F}_4 = \phi_4\left(\text{Concat}\left(\text{Up}(F_5), F_4\right)\right), \tag{7}$$

where $\text{Up}(\cdot)$ denotes $2\times$ nearest-neighbor upsampling, and $\phi_4(\cdot)$ represents the A2C2f refinement block.

Similarly, P4 is fused with P3:

$$\tilde{F}_3 = \phi_3\left(\text{Concat}\left(\text{Up}(\tilde{F}_4), F_3\right)\right). \tag{8}$$

3.1.2. High-Resolution P2 Construction

To better preserve fine-grained spatial details for tiny objects, we introduce a novel P2 branch:

$$\tilde{F}_2 = \phi_2\left(\text{Concat}\left(\text{Up}(\tilde{F}_3), F_2\right)\right). \tag{9}$$

This produces a high-resolution feature map:

$$\tilde{F}_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_2}, \tag{10}$$

which provides denser spatial sampling for small-object localization.

3.1.3. Bottom-Up Semantic Refinement

To maintain multi-scale semantic consistency, we introduce a bottom-up refinement path.

From P2 to P3:

$$\hat{F}_3 = \psi_3\left(\text{Concat}\left(\text{Down}(\tilde{F}_2), \tilde{F}_3\right)\right), \tag{11}$$

From P3 to P4:

$$\hat{F}_4 = \psi_4\left(\text{Concat}\left(\text{Down}(\hat{F}_3), \tilde{F}_4\right)\right), \tag{12}$$

From P4 to P5:

$$\hat{F}_5 = \psi_5\left(\text{Concat}\left(\text{Down}(\hat{F}_4), F_5\right)\right), \tag{13}$$

where $\text{Down}(\cdot)$ denotes stride-2 convolution and $\psi_i(\cdot)$ represents the feature refinement block.

3.1.4. Four-Scale Detection Strategy

The final detection is performed on four feature scales:

$$\mathcal{F}_{det} = \{\tilde{F}_2, \hat{F}_3, \hat{F}_4, \hat{F}_5\}. \tag{14}$$

For each scale $i \in \{2, 3, 4, 5\}$, the detection head predicts:

$$P_i = f_{det}(F_i), \tag{15}$$

where

$$P_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times(4+1+C)}. \tag{16}$$

Here, 4 corresponds to bounding box regression, 1 to objectness, and C to the number of object categories.

3.1.5. Spatial Sampling Density Analysis

Compared to conventional three-scale detection (P3–P5), the proposed P2 branch increases spatial sampling density by a factor of four relative to P3:

$$\frac{\text{Density}_{P2}}{\text{Density}_{P3}} = 4. \tag{17}$$

This denser sampling significantly improves the localization capability for extremely small pedestrian instances in UAV imagery. Figure 1 shows the architecture of P2 enhanced detection head. Figure 2 shows the architecture of the proposed YOLOv12 with P2-enhanced multi-scale head.

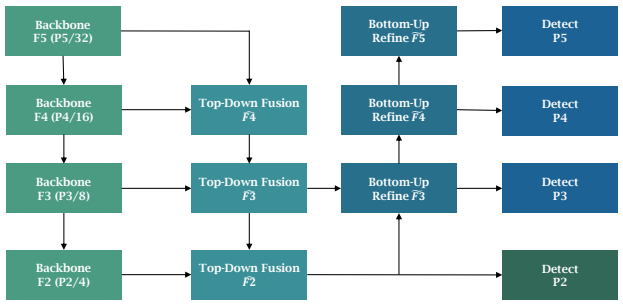

Figure 1. Architecture of the proposed P2-enhanced YOLOv12 detection head. The design extends conventional P3–P5 detection by introducing a high-resolution P2 branch and bidirectional feature refinement via FPN (top-down) and PAN (bottom-up) pathways, improving tiny-object localization in UAV imagery

4. Experimental Setup

4.1. Datasets

The proposed model was evaluated using the VisDrone 2019 [27] and TinyPerson [25] datasets. The VisDrone dataset comprises aerial images captured by UAV platforms under various real-world conditions, including varying altitudes, lighting conditions, occlusions, and background clutter. It contains 6,471 training images and 548 validation images, with a maximum image resolution of 2000 × 1500 pixels and 10 object categories. For the pedestrian-focused evaluation, only the " person" class is considered. This dataset presents significant challenges due to the presence of dense, small objects and scale variations. The TinyPerson dataset is specifically designed for detecting extremely small individuals in aerial and maritime scenes. It includes 736 training images, 796 validation images, and over 72,000 annotations. Most instances in this dataset are smaller than 32 × 32 pixels. To ensure a fair comparison, we merge the "sea_person" and "earth_person" categories into a single pedestrian class. This dataset highlights challenges posed by ultra-small objects, dense scenes, and complex backgrounds.

4.2. Implementation Details

The experiments were conducted using the YOLOv12n framework. We compared the baseline YOLOv12n model with the proposed version, which is equipped with the P2 enhanced detection head, on the VisDrone and TinyPerson datasets. The models were trained at a resolution of 1536 x 1536, with an additional ablation at a resolution of 1280. The training is conducted on NVIDIA A100 GPU with 40 GB memory. Table 1 shows the parameters used for training the models.

Table 1. Training configuration.

Parameter	Value	Augmentation
Optimizer	SGD	Mosaic = 0.7
Initial LR	0.003	Scale = 0.3
Momentum	0.937	MixUp = 0.0
Weight Decay	default	Copy-Paste = 0.1
Epochs	100	Close Mosaic = 10 epochs
Batch Size	8–12	
Warmup Epochs	5	
AMP	Enabled	
Cache	Disk	

4.3. Evaluation metrics

The metrics used to evaluate the model include precision, recall, average precision, and mean average precision, which are defined as follows:

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

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

$$AP = \int_0^1 P(R)dR \tag{20}$$

Where P: Precision, R: Recall, TP: True Positives, FP: False Positives, FN: False Negatives, AP: Average Precision, P(R): Precision as a function of recall, dR: differential change in recall.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{21}$$

Where mAP: mean Average Precision and n refers to the number of classes, while AP_k stands for the average precision for class k. To analyze performance on small-scale objects, we performed a size-based AP50 evaluation by categorizing ground-truth instances by bounding-box area (in pixels). Let A = w x h denote the area; the instances for TinyPerson datasets are grouped as follows:

$$S(A) = \begin{cases} \text{Tiny}, & A < 32^2 \\ \text{Small}, & 32^2 \le A < 64^2 \\ \text{Medium}, & A \ge 64^2 \end{cases}$$

For VisDrone datasets, the instances are categorized as:

$$S(A) = \begin{cases} \text{Small}, & h < 20 \\ \text{Medium}, & 20 \le h < 40 \\ \text{Large}, & h \ge 40 \end{cases}$$

where h denotes the height (in pixels) of the pedestrian bounding box.

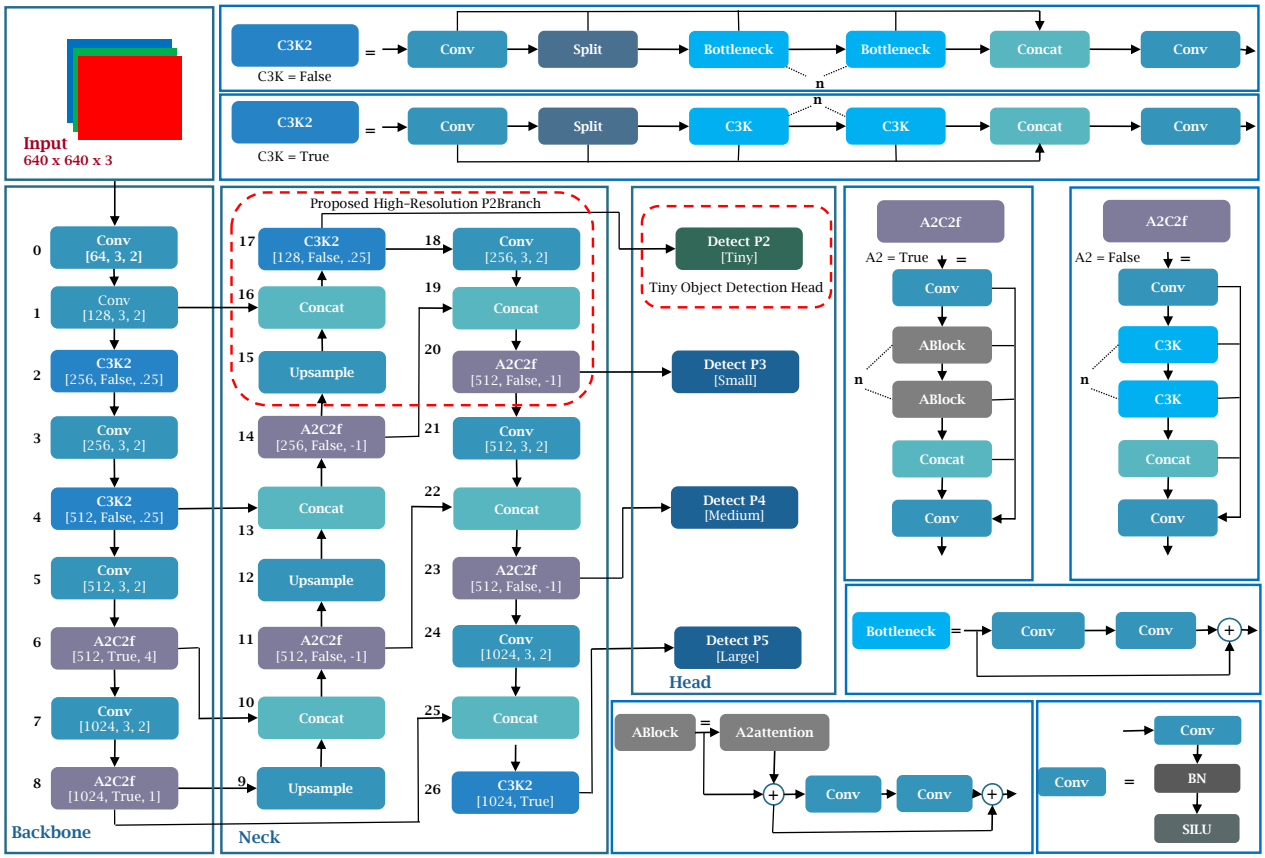

Figure 2. Overall architecture of the proposed YOLOv12-P2 detection framework for aerial pedestrian detection. The baseline YOLOv12n [17] backbone and neck (left) are extended by introducing a high-resolution stride-4 P2 branch (red dashed region) to enhance tiny-object representation. The proposed P2 branch is generated via top-down fusion from P3 features followed by C3k2 refinement, and is integrated into the bidirectional feature aggregation pathway. Final detection is performed at four scales (P2–P5), enabling improved localization of tiny and small pedestrians in UAV imagery.

5. Results

We assess the proposed P2-enhanced detection head using two challenging UAV benchmarks: VisDrone [27] and TinyPerson [25].

5.1. Results on VisDrone

Table 2 demonstrates that the P2-detection branch enhances overall detection performance on the VisDrone dataset. The size-wise analysis in Table 3 shows that the proposed model consistently improves performance across both small and medium scales. Although the performance for large scales remains competitive (0.80 vs. 0.86), this indicates that the high-resolution P2 detection branch significantly improves fine-grained localization without causing a major decline in performance for larger scales. These results confirm that strengthening high-resolution feature representation primarily benefits small-scale pedestrian detection in UAV imagery.

Table 2. Overall performance on VisDrone (Pedestrian class).

Model	Params (M)	GFLOPs	P	R	mAP@ 0.5	mAP@ 0.5:0.95
YOLOv12n	2.56	6.3	0.72	0.55	0.63	0.29
YOLOv12n + P2	2.86	10.3	0.76	0.62	0.69	0.33

Table 3. Size-wise AP50 performance on VisDrone (Pedestrian class).

Model	Small	Medium	Large
YOLOv12n	0.58	0.80	0.86
YOLOv12n + P2	0.63	0.84	0.80

5.2. Results on TinyPerson

Table 4 presents the performance results of the tiny person datasets, which include extremely small pedestrians

captured from aerial viewpoints. The proposed model improves the map@0.5 and recall values, indicating enhanced detection sensitivity. The size-wise evaluation in Table 5 shows that significant improvements occur at the tiny and small scales. These results confirm that the additional high-resolution P2 branch effectively enhances fine-grained feature representation, which is crucial for detecting tiny pedestrians in UAV imagery.

Table 4. Overall performance on TinyPerson.

Model	Params (M)	GFLOPs	P	R	mAP@ 0.5	mAP@ 0.5:0.95
YOLOv12n	2.56	6.3	0.54	0.42	0.40	0.14
YOLOv12n + P2	2.86	10.3	0.58	0.46	0.45	0.19

Table 5. Size-wise AP50 performance on TinyPerson.

Model	Tiny	Small	Medium
YOLOv12n	0.24	0.23	0.17
YOLOv12n + P2	0.30	0.28	0.12

6. Qualitative Results

To assess the performance of the P2-enhanced detector, we conducted a qualitative analysis on challenging scenes from the VisDrone and TinyPerson datasets. Figure 3 shows a qualitative comparison of dense urban scenes from the VisDrone dataset. The proposed P2-enhanced model effectively detects more small-scale pedestrians in crowded areas compared to the baseline model. In the last row, while some pedestrians remain undetected compared to the ground truth, the proposed model still outperforms the baseline in detection accuracy. Figure 4 presents qualitative results for TinyPerson, which captures extremely small pedestrians from aerial viewpoints. The proposed model outperforms the baseline, especially at detecting distant, tiny instances in large-scale beach scenes with high crowd density. In the final row, the proposed model misses fewer tiny pedestrians compared to the ground truth, consistently surpassing the baseline. These qualitative results demonstrate that integrating P2 enhances high-resolution feature representation for extremely small targets. Figure X shows qualitative comparisons on the TinyPerson and VisDrone datasets. In the highlighted areas, pedestrians appear very small, causing some to be missed by the baseline model. The proposed model, utilizing high-resolution features from the P2 detection head, detects these tiny pedestrians, resulting in denser and more consistent detection, even in distant aerial regions.

7. Ablation Study

To evaluate the contribution of each design component, we conducted ablation experiments using the VisDrone dataset. We assessed the impact of the P2 detection branch, its computational overhead, and the effect of input resolution on detection performance.

7.1. Impact of P2 Detection Branch

Table 6 illustrates the effect of utilizing the high-resolution P2 branch. The inclusion of P2 increases the overall mAP@0.5 from 0.63 to 0.69 (9.5% relative gain), showing consistent improvement across both small and medium scales. The enhancement from 0.58 to 0.63 (8.6% relative gain) for small pedestrians confirms that extending the feature pyramid to a stride of 4 enhances fine-grained spatial information. While large-scale performance remains competitive, there is a slight decrease (0.80 vs. 0.86), indicating that the architectural modification primarily improves sensitivity to small objects without significantly degrading detection of larger instances.

Table 6. Impact of the proposed P2 detection branch on VisDrone 2019 (AP50).

Model	mAP@0.5	Small	Medium	Large
YOLOv12n	0.63	0.58	0.80	0.86
YOLOv12n + P2 Head	0.69	0.63	0.84	0.80

7.2. Computational Overhead

The computational overhead of the P2 detection branch is analyzed in Table 7. The proposed model increases the number of parameters from 2.56 M to 2.86 M and the GFLOPs from 6.3 to 10.3, which reflects the inclusion of an additional high-resolution processing stage. Consequently, the inference latency has increased moderately from 4.6 ms to 5.7 ms per image. Despite this increase, the model remains lightweight and is suitable for deployment in UAV-based scenarios.

Table 7. Complexity Analysis on VisDrone 2019.

Model	Parms(M)	GFLOPs	Inference(ms)
YOLOv12n	2.56	6.3	4.6
YOLOv12n + P2	2.86	10.3	5.7

7.3. Resolution Sensitivity

The model's performance at different input resolutions is analyzed in Table 8. At a resolution of 1280, the proposed model outperforms the baseline, achieving an mAP@0.5 of 0.64 compared to the baseline's 0.61. Increasing the resolution to 1536 results in a 13.1% relative gain in mAP@0.5.

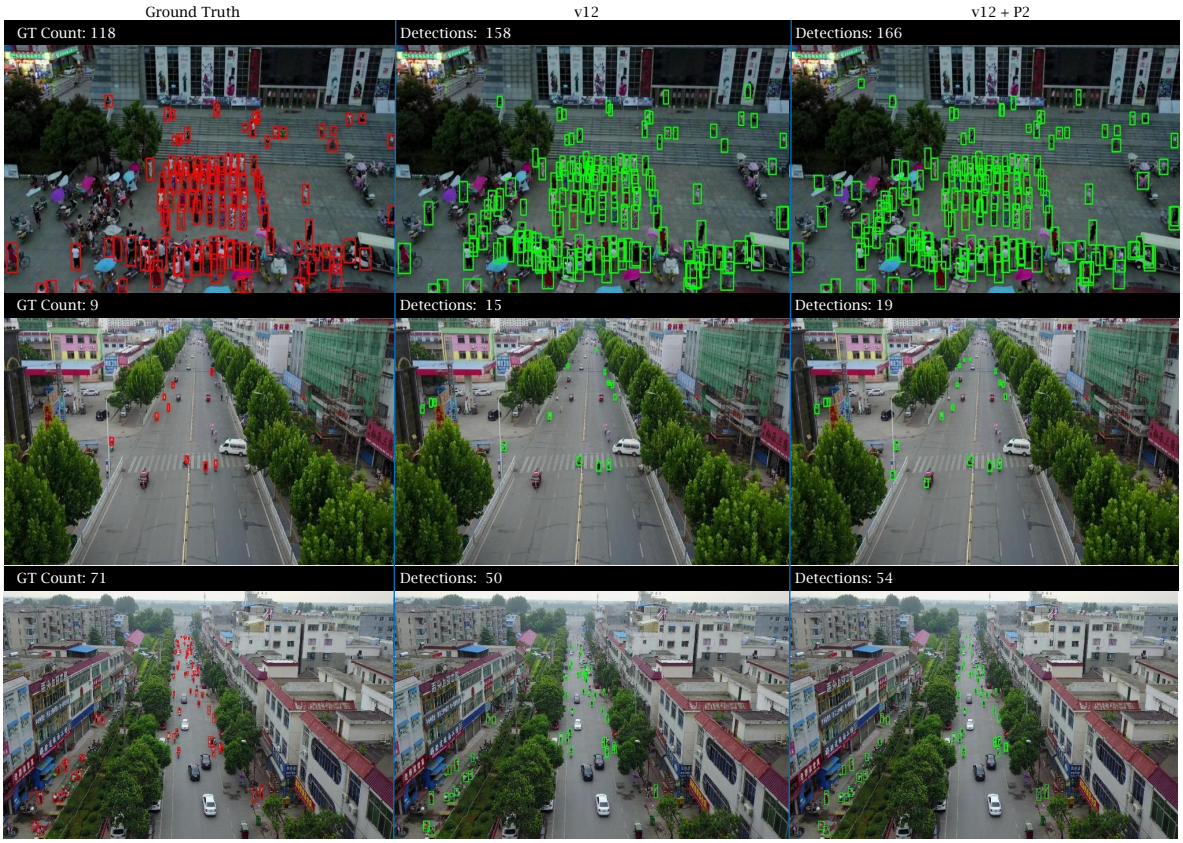

Figure 3. Qualitative results on the VisDrone dataset. From left to right: Ground Truth, YOLOv12n baseline, and YOLOv12n + P2. The proposed model improves detection density and small-object coverage in complex urban scenes. In the last row, the proposed model misses some pedestrians compared to the ground truth, but it detects more small pedestrians than the baseline, indicating improved sensitivity to small-scale pedestrians.

The results indicate that higher input resolutions provide richer spatial information to the P2 branch, which is particularly beneficial for detecting extremely small pedestrians. Overall, the findings demonstrate that the proposed design effectively utilizes both architectural refinement and resolution scaling to enhance small-object detection

Table 8. Resolution Sensitivity on VisDrone 2019.

Model	Resolution	P	R	mAP@ 0.5	mAP@ 0.5:0.95
YOLOv12n	1280	0.70	0.53	0.61	0.26
YOLOv12n	1536	0.72	0.55	0.63	0.29
YOLOv12n + P2	1280	0.74	0.58	0.64	0.30
YOLOv12n + P2	1536	0.76	0.62	0.69	0.33

8. Conclusion

In this study, we examined the limitations of conventional feature pyramids for pedestrian detection in UAV imagery, particularly under conditions involving extremely small objects. We found that standard YOLOv12 detection starts at P3 resolution, limiting its effectiveness at detecting ultra-small pedestrians common in aerial scenes. To overcome this limitation, we proposed a high-resolution, P2-enhanced detection head that extends the feature pyramid to a stride-4 resolution while maintaining computational efficiency. Our proposed detection head leverages top-down feature fusion and bottom-up feature refinement to maintain semantic consistency across scales. Experiments conducted on the VisDrone and TinyPerson datasets showed consistent improvements in average precision (AP) and recall for tiny objects, confirming the advantages of early-resolution supervision in aerial pedestrian detection. Importantly, this modification retains real-time performance with only a moderate increase in computational overhead. These results suggest that refining the architecture at early pyramid levels is an effective and practical strategy for enhancing small-object detection in UAV applications. Future work will focus on adaptive scale-aware feature routing and lightweight attention mechanisms to further improve detection robustness in dense, cluttered UAV scenarios.

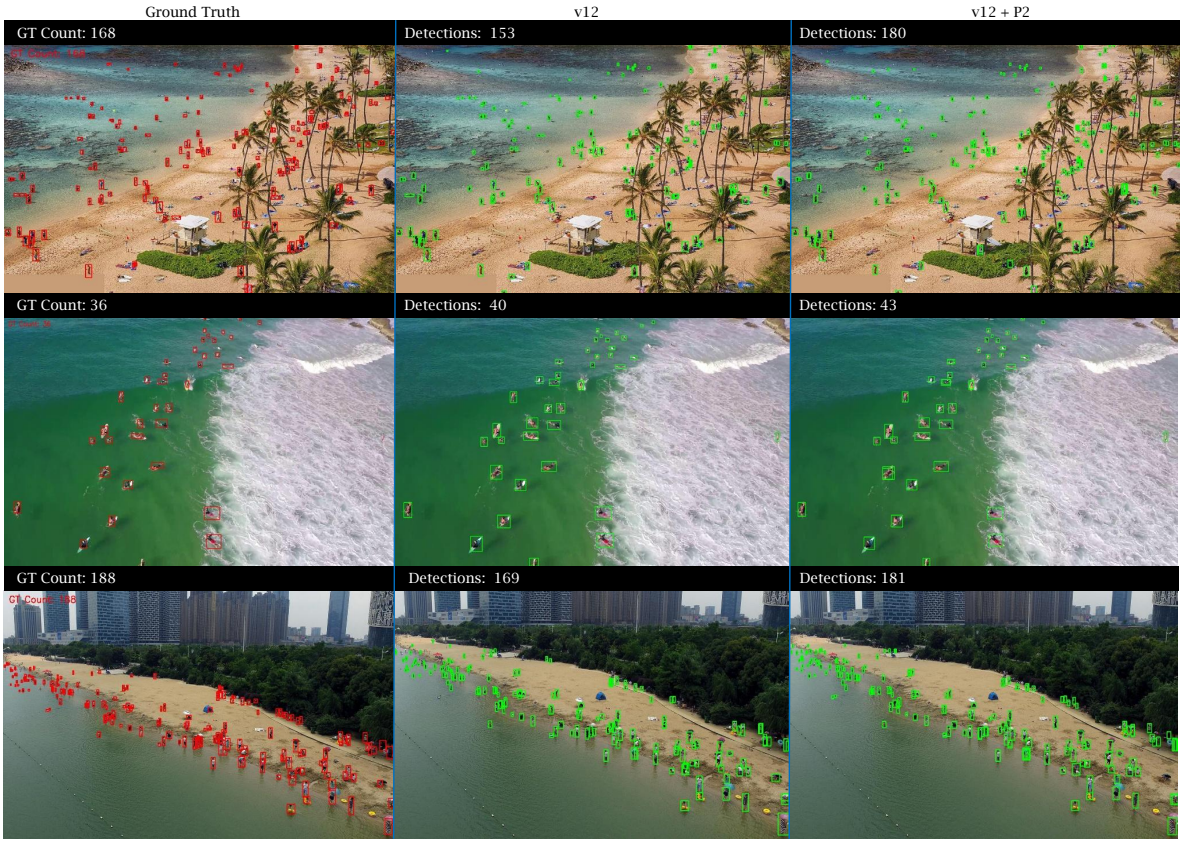

Figure 4. Qualitative results on the TinyPerson dataset. From left to right: Ground Truth, YOLOv12n baseline, and YOLOv12n + P2. The proposed model improves detection of extremely small pedestrians in dense aerial scenes. Some distant pedestrians are undetected in the last row compared to the ground truth. The proposed P2 enhanced model consistently detected more tiny pedestrians than the baseline, indicating a stronger high-resolution feature representation for extremely small pedestrians.

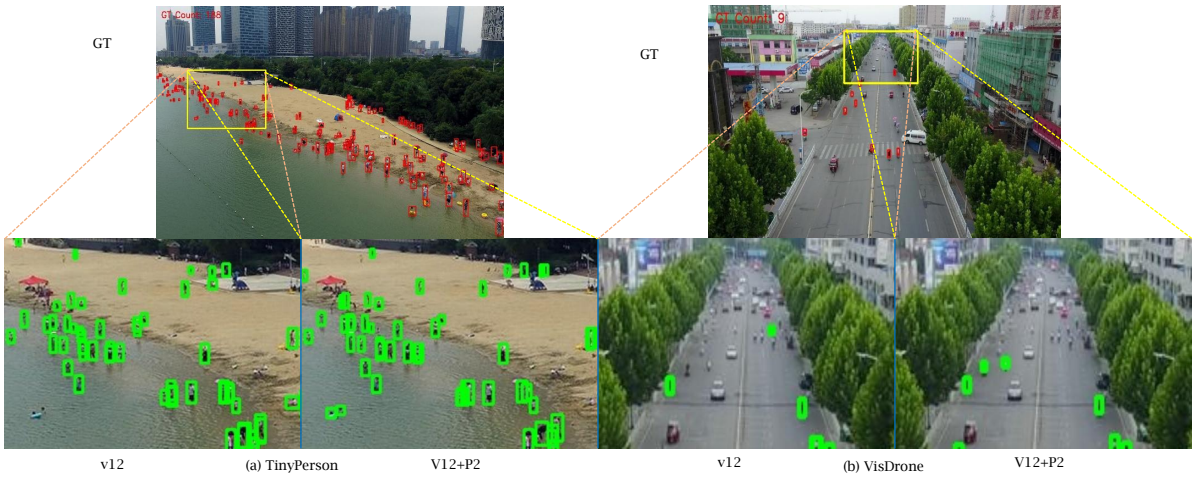

Figure 5. Qualitative comparison on TinyPerson and VisDrone datasets. The top row shows the ground-truth scene with the selected zoom region highlighted. The bottom row compares detection results between YOLOv12 and the proposed YOLOv12+P2 model. The proposed model significantly improves the detection of very small pedestrians, especially in distant aerial regions where baseline detectors often overlook such instances.

References

[1] Hongbo Bi, Rui Dai, Fengyang Han, and Cong Zhang. Drm-yolo: A yolov11-based structural optimization method for small object detection in uav aerial imagery. *Image and Vision Computing*, page 105894, 2025. 1

[2] Hui Chang, Yuru Long, Yiwen Guo, Yilin Li, Jinrui Wang, Kun Zhang, and Liangsong Huang. Enhancing small object detection in uav aerial imagery through integration of global edge information and multi-scale feature enhancement. *Measurement Science and Technology*, 37(6):066106, 2026. 2, 3

[3] Yajun Chang, Bingqian Suo, Tian Wang, Keping Wang, and Yi Yang. Entropy-enhanced swin transformer for small object detection in uav images. *Systems Science & Control Engineering*, 14(1):2600212, 2026. 1, 2

[4] Zhaodong Chen, Hongbing Ji, Yongquan Zhang, Zhigang Zhu, and Yifan Li. High-resolution feature pyramid network for small object detection on drone view. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):475–489, 2024. 2

[5] Zhiyong Deng, Yanchen Ye, and Jiangling Guo. Ehdc-yolo: Enhancing object detection for uav imagery via multi-scale edge and detail capture. *Computers, Materials, & Continua*, 86(1):1, 2026. 2

[6] Zhibin Fan, Liangyan Guo, and Linan Fan. Taming the tiny: A lightweight hierarchical attention network for small object detection in uav aerial images. *IEEE Access*, 2026. 2

[7] Jianglei Gong, Zhe Yuan, Wenxing Li, Weiwei Li, Yanjie Guo, and Baolong Guo. A lightweight upsampling and cross-modal feature fusion-based algorithm for small-object detection in uav imagery. *Electronics*, 15(2):298, 2026. 2

[8] Jingxiang Hu, Yongyi Chen, Dan Zhang, Zehui Mao, Zongda Wu, and Qinghua Ma. Emsanet: Edge-enhanced multi-scale aligned network for small object detection in uav imagery. *Neurocomputing*, page 132802, 2026. 2

[9] Hao Kong, Zhi Chen, Wenjing Yue, and Kang Ni. Improved yolov4 for pedestrian detection and counting in uav images. *Computational intelligence and neuroscience*, 2022 (1):6106853, 2022. 2

[10] Jialiang Li, Xu Guo, Xu Zhao, and Jie Jin. Fde-yolo: An improved algorithm for small target detection in uav images. *Mathematics*, 14(4):663, 2026. 2

[11] Zehua Li, Min Liu, Bohang Lv, Binrui Xu, Jincan Zhang, and Liwen Zhang. Dpn-yolo: an enhanced algorithm for small target detection in uav imagery. *The Journal of Supercomputing*, 82(2):56, 2026. 2

[12] Di Luan, Yuna Dong, Jian Zhou, Ang Li, Ling Xie, Hongying Liu, and Jun Zhu. Wcdb-yolo: Wavelet-enhanced contextual dual-backbone network for small object detection in uav aerial imagery. *Drones*, 10(3):155, 2026. 2

[13] Ghulam Mujtaba, Wenbiao Liu, Mohammed Alshehri, Yahya AlQahtani, Nouf Abdullah Almujally, and Hui Liu. Aerial images for intelligent vehicle detection and classification via yolov11 and deep learner. *Computers, Materials, & Continua*, 86(1):1, 2026. 1

[14] Zhenfeng Shao, Gui Cheng, Jiayi Ma, Zhongyuan Wang, Jiaming Wang, and Deren Li. Real-time and accurate uav pedestrian detection for social distancing monitoring in covid-19 pandemic. *IEEE transactions on multimedia*, 24: 2069–2083, 2021. 2

[15] Arun Kumar Sivapuram, Pranav R T Peddinti, Harish Puppala, Komuravelli Prashanth, Jaladi Sri Harsha, and Rama Krishna Sai Gorthi. Realdronevision: Dataset and architecture advancements for small-object drone detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6687–6695, 2026. 2

[16] Noor Ul Ain Tahir, Li Kuang, Melikamu Liyih Sinishaw, and Muhammad Asim. Pv3m-yolo: A triple attention-enhanced model for detecting pedestrians and vehicles in uav-enabled smart transport networks. *Journal of Visual Communication and Image Representation*, page 104701, 2026. 1, 2, 3

[17] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 3, 5

[18] Sukesh Babu V.S. and Rahul Raman. Costaa yolo: Convolutional swin transformer with attention and anchor box optimization on yolov7 for robust pedestrian detection. *Image and Vision Computing*, 168:105942, 2026. 2

[19] Sukesh Babu V.S. and Rahul Raman. Mecsa: a multi-scale enhanced channel and spatial attention module for robust pedestrian detection. *Pattern Analysis and Applications*, 29, 2026. 2

[20] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3791–3798, 2021. 2

[21] Zi Wang and Jun Yang. Re-yolo: a lightweight small object detection method for uav remote sensing imagery. *Annals of GIS*, pages 1–20, 2026. 3

[22] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13668–13677, 2022. 2

[23] Jianxiu Yang, Xiangmei Yue, and Liang Wu. A collaborative multi-attention network for real-time small object detection in uav imagery. *Scientific Reports*, 2026. 2

[24] Yiming Yang, Feng Guo, and Pei Niu. Uavdet: A cnn-mamba hybrid network for efficient small object detection in uav imagery. *Computer Vision and Image Understanding*, page 104637, 2026. 2, 3

[25] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020. 2, 4, 5

[26] Xindi Zhang, Ebroul Izquierdo, and Krishna Chandramouli. Dense and small object detection in uav vision based on cascade network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2

[27] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7380–7399, 2021. 2, 4, 5