remote sensing

MDPI

Article

SMA-YOLO: An Improved YOLOv8 Algorithm Based on Parameter-Free Attention Mechanism and Multi-Scale Feature Fusion for Small Object Detection in UAV Images

Shenming Qu, Chaoxu Dang, Wangyou Chen and Yanhong Liu *

School of Software, Henan University, Kaifeng 475004, China; qushenming@whu.edu.cn (S.Q.); dcx@henu.edu.cn (C.D.); chenwangyou@henu.edu.cn (W.C.)
* Correspondence: yanhongliu@vip.henu.edu.cn

Abstract

With special consideration for complex scenes and densely distributed small objects, this frequently leads to serious false and missed detections for unmanned aerial vehicle (UAV) images in small object detection scenarios. Consequently, we propose a UAV image small object detection algorithm, termed SMA-YOLO. Firstly, a parameter-free simple slicing convolution (SSC) module is integrated in the backbone network to slice the feature maps and enhance the features so as to effectively retain the features of small objects. Subsequently, to enhance the information exchange between upper and lower layers, we design a special multi-cross-scale feature pyramid network (M-FPN). The C2f-Hierarchical-Phantom Convolution (C2f-HPC) module in the network effectively reduces information loss by fine-grained multi-scale feature fusion. Ultimately, adaptive spatial feature fusion detection Head (ASFFDHead) introduces an additional P2 detection head to enhance the resolution of feature maps to better locate small objects. Moreover, the ASFF mechanism is employed to optimize the detection process by filtering out information conflicts during multi-scale feature fusion, thereby significantly optimizing small object detection capability. Using YOLOv8n as the baseline, SMA-YOLO is evaluated on the VisDrone2019 dataset, achieving a 7.4% improvement in mAP@0.5 and a 13.3% reduction in model parameters, and we also verified its generalization ability on VAUDT and RSOD datasets, which demonstrates the effectiveness of our approach.

Keywords: UAV images; feature fusion; small objects; YOLOv8; SMA-YOLO

check for updates

Academic Editors: Qiang Li, Jing Li, Dan Wang and Tania Stathaki

Received: 11 June 2025
Revised: 8 July 2025
Accepted: 10 July 2025
Published: 12 July 2025

Citation: Qu, S.; Dang, C.; Chen, W.; Liu, Y. SMA-YOLO: An Improved YOLOv8 Algorithm Based on Parameter-Free Attention Mechanism and Multi-Scale Feature Fusion for Small Object Detection in UAV Images. *Remote Sens.* **2025**, *17*, 2421. https://doi.org/10.3390/rs17142421

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, more and more fields, including agriculture [1], transportation [2], and industry [3], are now utilizing unmanned aerial vehicle (UAV) detection technology [4–6]. However, due to the high-altitude and oblique-view imaging characteristics of UAVs, ground objects often appear extremely small, making them difficult for detection algorithms to accurately identify. Furthermore, as illustrated in Figure 1, factors such as complex backgrounds, dense small objects, blurriness, and night situations further exacerbate the challenges of accurate small object detection.

Progress in deep learning has catalyzed significant innovations in object detection methodologies [7] across computer vision applications. Currently, there are two mainstream categories of detection methodologies: two-stage and single-stage approaches. Representative two-stage detection methodologies, including Faster RCNN [8], Cascade RCNN [9],

and Mask RCNN [10], are recognized for their high detection accuracy but often suffer from slower processing speeds. In contrast, SSD [11] pioneered single-stage detection, and then came methodologies like RetinaNet [12] and YOLO [13]. In addition to these two detection algorithms, there exists another category of vision transformers, including DETR [14] and Swin Transformer [15], which process images by dividing them into sequential patches and applying self-attention mechanisms for object detection. While vision transformers can simplify the overall architecture, their high computational demands make them unsuitable for deployment on resource-constrained UAV devices. Single-stage detectors offer a better trade-off between accuracy and efficiency, making them more suitable for integration with UAV platforms.



(**a**) (**b**) (**c**) (**d**)

Figure 1. (**a**) Complex backgrounds. (**b**) Dense small objects. (**c**) Blurriness. (**d**) Night situation. The Chinese characters visible do not affect the scientific understanding.

Although YOLOv8 [16] demonstrates strong performance in general object detection tasks, its capability to detect small objects remains limited, particularly in UAV images, where small objects are prevalent and frequently obscured by complex backgrounds. Therefore, enhancing YOLOv8's small object detection capability in UAV images presents dual significance: advancing computer vision theory while addressing critical needs in aerial surveillance applications. In this context, this paper proposes SMA-YOLO, a YOLOv8-based algorithm designed for small object detection in UAV images, evaluated on the VisDrone2019 [17] dataset. In this paper, the primary innovations are as follows:

1. We propose a parameter-free simple slicing convolution (SSC) to take the place of standard convolutions in the backbone network. By strategically partitioning feature maps and incorporating SimAM [18] attention, this module effectively preserves and enhances the discriminative features of small objects.
2. A multi-cross-scale feature pyramid network (M-FPN) is designed to optimize feature fusion in the neck network. Through its unique multi-level and cross-scale connections combined with the C2f-HPC module, our approach achieves fine-grained multi-

scale feature integration, significantly reducing information loss for small objects in complex scenarios.

3. We develop an adaptive spatial feature fusion detection head (ASFFDHead) featuring an additional P2 detection head for small objects specifically. By implementing the ASFF [19] mechanism to resolve feature conflicts during multi-scale fusion, the proposed structure substantially improves detection accuracy for small objects.

2. Materials

2.1. YOLOv8

As illustrated in Figure 2, the YOLOv8 architecture comprises three key components: backbone, neck, and head. The backbone network employs an improved C2f module to replace the conventional Cross Stage Partial (CSP) structure, where each residual bottleneck integrates dual convolutional layers for enhanced contextual information flow. Notably, the more efficient Spatial Pyramid Pooling-Fast (SPPF) variant is developed as an upgrade to the traditional SPP [20], achieving superior feature representation through sequential pooling operations. For bounding box regression, YOLOv8 incorporates Complete IoU (CIoU) [21] loss to better handle geometric relationships and aspect ratio variations. For feature fusion, YOLOv8 adopts a Path Aggregation Network (PANet) [22]-based neck structure that systematically combines hierarchical features through bidirectional cross-scale connections. Integrating both spatial details at low levels and semantic information at high levels across different scales, this design demonstrates effective feature fusion. The detection head further implements Distribution Focal Loss (DFL) [23], modeling bounding box coordinates as probability distributions to enhance localization precision for small objects. The detection head maintains the efficiency advantages of single-stage detectors while optimizing the anchor-based mechanism. By strategically balancing between structural simplicity and detection accuracy, YOLOv8 avoids the computational overhead typically associated with conventional anchor methods without compromising performance. This combination of CIoU and DFL addresses both large-scale geometric alignment and fine-grained localization accuracy.

2.2. UAV Images Small Object Detection

Small object detection [24,25] in UAV images has emerged as a critical research challenge in computer vision. Recent advances have primarily focused on architectural modifications and optimization strategies, each presenting unique advantages and limitations.

Zhang et al. [26] proposed a joint optimization framework incorporating model fusion, cascade networks, and deformable convolutions, which achieved state-of-the-art performance on VisDrone2019. However, the integration of multiple complex mechanisms significantly increased computational overhead, reducing practical deployment efficiency. Similarly, LAR-YOLOv8 [27] enhanced feature representation through dual-branch attention and vision transformer blocks, but its sophisticated architecture resulted in notable inference latency.

In the domain of detection head optimization, DarkNet-RI [28] introduced a classification-based localization approach with refined NMS [29] processing. While effective for isolated objects, its performance degraded severely in high-occlusion scenarios. Xu et al.'s DotD [30] metric addressed IoU sensitivity for small objects through normalized centroid distance, yet the additional geometric computations increased processing time.

Anchor-based improvements have shown particular promise, as demonstrated by Zhan et al. [31], who redesigned anchor sizes and adopted GIoU [32] in YOLOv5, achieving mAP improvement. Nevertheless, their fixed anchor strategy struggled with extreme scale

variations common in UAV imagery. TPH-YOLOv5 [33] overcame this limitation through transformer prediction heads but required more training data to achieve stable convergence.

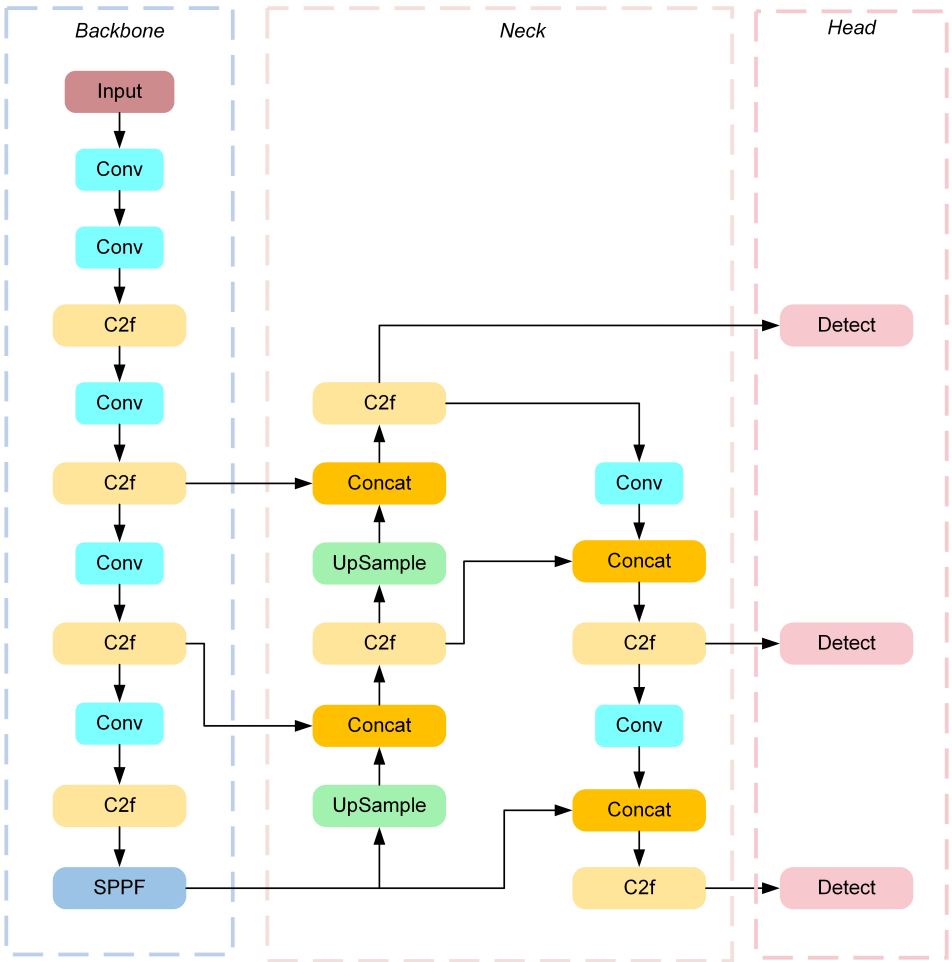


Figure 2. The structure of YOLOv8.

Shallow feature utilization remains a critical challenge. DS-YOLOv7 [34] employed SFN technology and LDSPP modules to enhance edge information, yet its multi-branch design increased parameters. IMCMD_YOLOv8_small [35] took an alternative approach by removing P5 layers and focusing on shallow feature fusion, but this came at the cost of completely losing large-object detection capability.

3. Methods

Small object detection in UAV images presents unique challenges due to the inherent limitations of limited pixel representation and complex environmental interference, which significantly degrade YOLOv8's feature representation capacity. To address the above issues, we propose SMA-YOLO, an enhanced architecture that systematically improves small object detection through three integrated innovations. As shown in Figure 3, our framework first introduces a Simple Slicing Convolution (SSC) module to replace standard backbone convolutions, leveraging feature map partitioning and SimAM attention mechanisms to amplify fine-grained details of small objects. The multi-cross-scale feature pyramid network (M-FPN) is then optimized with cross-scale fusion pathways and a novel C2f-HPC module, creating more discriminative feature representations that preserve critical small object information across different scales. Finally, the architecture incorporates an Adaptive Spatial Feature Fusion Detection Head (ASFFDHead) with a dedicated P2 detection branch, where the ASFF mechanism dynamically balances multi-scale feature contributions to

effectively filter conflicting information from multi-scale feature fusion. The comprehensive method maintains a certain processing speed while significantly enhancing its small object detection capability through coordinated improvements across all network components.

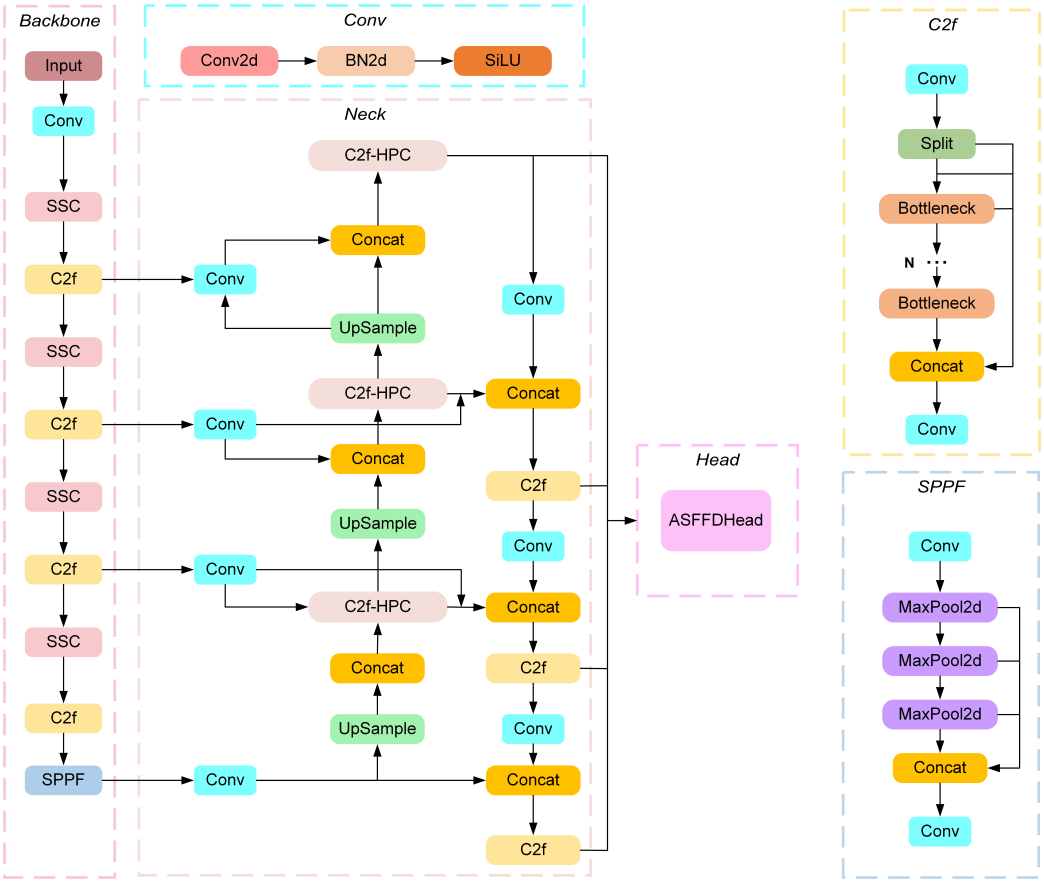


Figure 3. The structure of SMA-YOLO network.

3.1. Simple Slicing Convolution

The performance degradation of conventional CNNs in small object detection primarily stems from inadequate feature extraction architectures and progressive information loss during hierarchical processing. As illustrated in Figure 4, our Simple Slicing Convolution (SSC) module addresses these limitations through a multi-stage processing pipeline. The input feature map first undergoes symmetric partitioning into four sub-regions via strided convolution with batch normalization and SILU activation. Each partition then enters parallel enhancement blocks containing SimAM-augmented residual units and transition layers with feature division operations, where binary activations introduce controlled sparsity to preserve critical high-frequency details.

The three-dimensional attention mechanism depicted in Figure 5c fundamentally differs from conventional 1D channel-wise or 2D spatial attention approaches. By simultaneously modeling channel–height–width relationships through sequential feature generation, cross-dimensional fusion, and hierarchical expansion, SimAM generates more discriminative 3D weights that intrinsically amplify small object features. This process automatically assigns stronger weights to smaller objects whose features exhibit greater deviation from mean activation values, while larger objects with prominent textures receive proportionally less enhancement due to their inherent detectability.

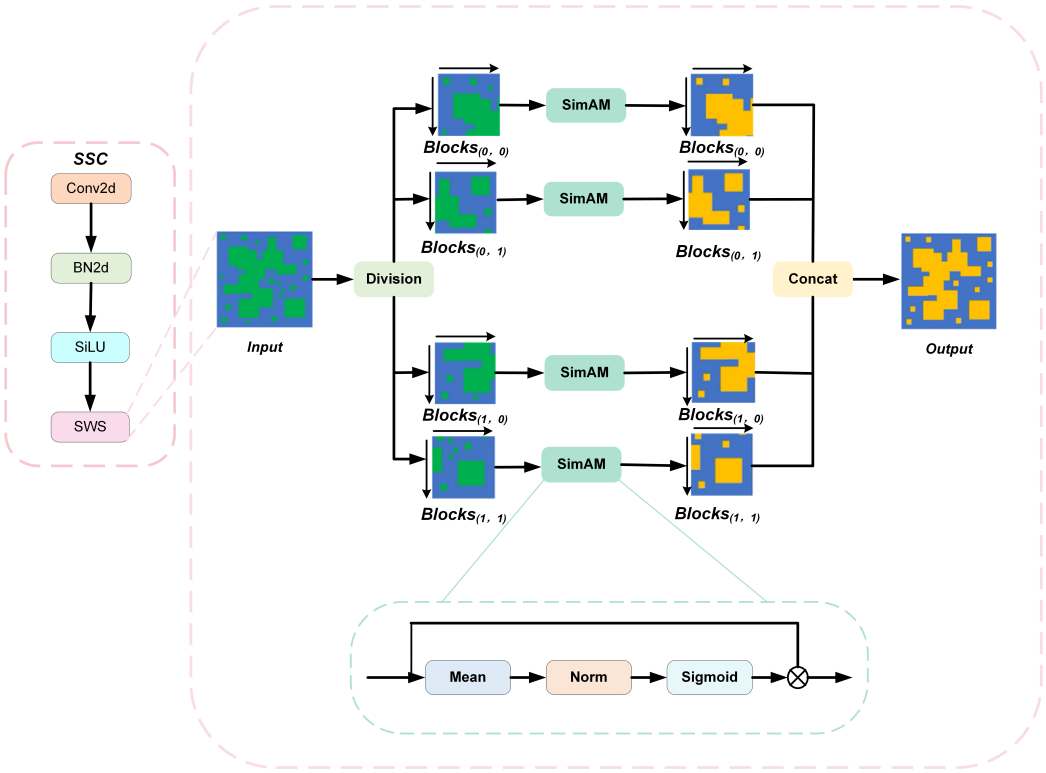


Figure 4. The structure of SSC and the internal principle of SWS module.

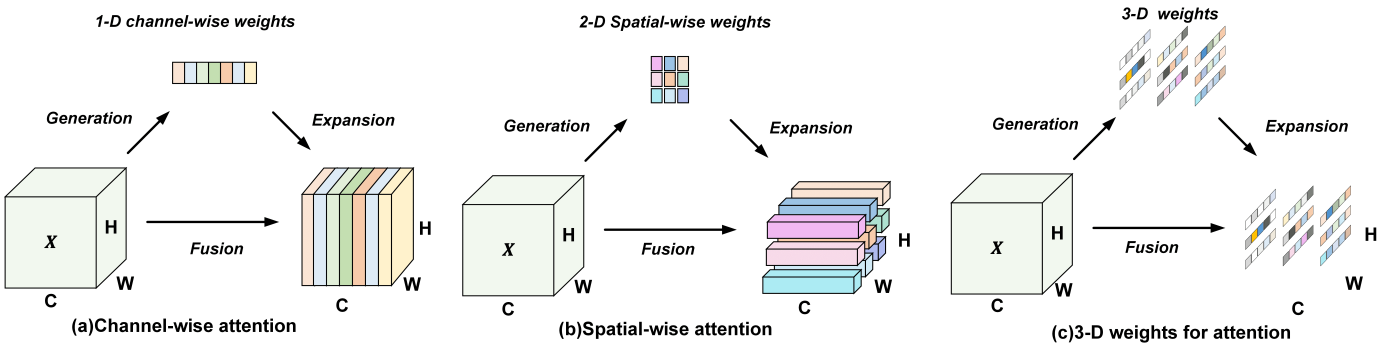


Figure 5. (**a**) Generating channel 1D weights. (**b**) Generating spatial two-dimensional weights. (**c**) Generating 3D weights. In each subfigure, X denotes the input feature map; C, H, and W indicate the number of channels, height, and width, respectively. Different colors represent distinct attention weights along different dimensions.

The SSC module offers the dual advantages of resolution preservation and adaptive feature enhancement without introducing any additional parameters. Although SimAM is referred to as a "parameter-free" attention mechanism, this term specifically refers to the absence of trainable parameters such as weights or biases during learning. Internally, SimAM does include lightweight operations, but these are fixed mathematical operations and do not involve any learnable parameters. The principle of the SimAM module comes from the spatial suppression phenomenon [36]. It posits that neurons exhibiting firing patterns distinct from their surroundings carry more critical information. SimAM defines an energy function e_t for each neuron t, measuring its linear separability from other neurons within the same channel. Crucially, SimAM derives a fast closed-form solution for the minimum energy e_t^*; the formula is as follows:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

Here, $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance computed over the entire channel of the input feature map, t is the activation value of the target neuron, and λ is a small, fixed hyperparameter (set to 10^{-4} in our experiments) for numerical stability. Notably, this solution relies solely on feature statistics and involves no learnable parameters. The importance weight for neuron t is given by $1/e_t^*$. Lower e_t^* indicates higher distinctiveness and importance. The final feature refinement is performed via element-wise multiplication with a sigmoid-gated version of the importance map; the formula is as follows:

$$\widetilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \tag{2}$$

where E aggregates all e_t^* values. This operation selectively enhances informative neurons while suppressing less relevant ones, mimicking the gain effect of attention in biological vision.

The slicing operation maintains fine-grained spatial information that would otherwise be lost in standard convolution, while the 3D attention mechanism dynamically recalibrates features across all dimensions. The parameter-free nature of SimAM is a key advantage. While it incurs computational cost for calculating channel statistics ($\hat{\mu}$, $\hat{\sigma}^2$) and applying Equations (1) and (2), it avoids the expensive parameterized operations. The enhanced features provide superior input for subsequent detection stages, particularly benefiting small objects in complex UAV scenarios where traditional methods often fail.

3.2. Multi-Cross-Scale Feature Pyramid Network

Effective multi-cross-scale feature fusion remains a fundamental challenge in UAV small object detection due to extreme scale variations and complex spatial distributions of objects. While YOLOv8's PANet architecture improves upon traditional FPN [37] with its bidirectional connections, several critical limitations hinder optimal performance. The current architecture employs simple replication of reverse fusion paths, resulting in redundant parameters that fail to establish meaningful cross-scale interactions. Moreover, its fixed fusion strategy cannot adapt to the differential importance of hierarchical features, while the bottleneck structures lack sufficient capacity for preserving fine-grained details—a particularly severe drawback for small objects where spatial information is easily lost during downsampling operations.

As illustrated in Figure 6, our proposed M-FPN addresses these limitations through three architectural innovations. First, we introduce adaptive 3 × 3 convolutions at each network level, dynamically adjusting receptive fields to capture scale-specific features. Second, the topology optimizes information flow through selective cross-layer connections that prioritize semantically complementary feature combinations rather than simple path duplication. Third, the integration of the C2f-HPC modules enables granular multi-scale fusion.

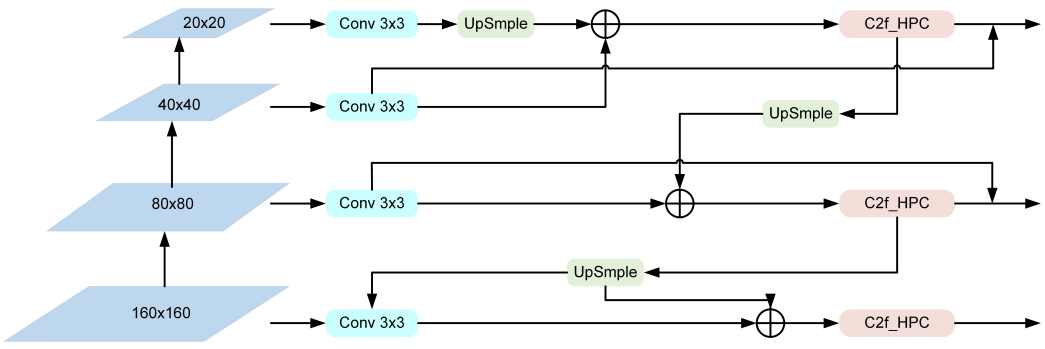


Figure 6. The structure of M-FPN. The ⊕ refers to Concat.

C2f-Hierarchical-Phantom Convolution

The C2f module in YOLOv8 serves as a key feature extraction component. However, its fixed receptive field struggles to capture the extreme scale variations of UAV objects, and the bottleneck structure inadequately preserves fine-grained spatial information, thus reducing the detection accuracy. As shown in Figure 7, our C2f-HPC module overcomes these constraints through a Hierarchical-Phantom Convolution (HPC) [38] architecture that enables progressive multi-scale feature learning.

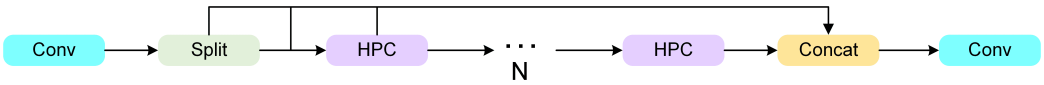


Figure 7. The construction of C2f-HPC.

As shown in Figure 8, the HPC module adopts a simple and efficient multi-scale feature extraction strategy that incorporates channel attention learning, significantly enhancing multi-scale representation capabilities at finer granularities. Concerning the input feature vector, the Split operator divides features into s homogeneous subsets F_i along the channel dimension, then each subset undergoes sequential 3 × 3 convolutions $\mathcal{T}_i(\cdot)$ in a residual hierarchy, where the i-th convolution operates on both the original subset and all previously refined features. This process is repeated, as formulated by the following equation:

$$\hat{F}_i = \begin{cases} \mathcal{T}_i(F_i), & i = 1 \\ \mathcal{T}_i(F_i \oplus \hat{F}_{i-1}), & 1 < i \leq s \end{cases} \tag{3}$$

where $\oplus$ represents the summation operation, and finally obtains the entire enhanced multi-scale feature map through Concat concatenation, as follows:

$$\hat{F} = \text{Concat}([\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_s]) \tag{4}$$

Each set of Conv operators in this module can extract feature information from a subset of feature vectors, and with each convolution operation, the output result will have a larger receptive field, which is subsequently fused with the input of the next layer, and after continuous iterative multi-scale fusion, due to the combinatorial explosion effect, the output receptive field of the HPC module contains rich fine-grained information, which can capture more small objects in the scene feature details.

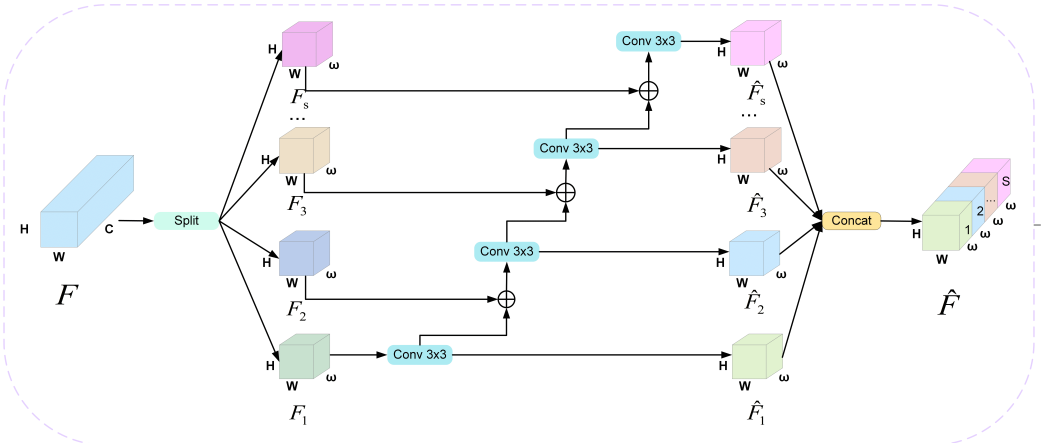


Figure 8. The construction of HPC. In the figure, H, W, and C represent the height, width, and number of channels of the input feature map, respectively. ω denotes the channel width after splitting, and s is the total number of splits. F and $\hat{F}$ denote the input and output features.

3.3. Adaptively Spatial Feature Fusion Detect Head

Single-stage detectors have shortcomings in handling scale variation due to the fundamental trade-off between resolution and semantic information across different feature levels. Shallow feature maps preserve fine spatial details essential for small object detection but lack semantic richness, while deeper features offer strong semantic representation at the cost of losing critical spatial information. This inconsistency leads to small object features being diluted or treated as background in deeper layers, conflicting information during multi-scale feature fusion, and suboptimal gradient propagation during training. As shown in Figure 9, we designed ASFFDHead to address these problems through an integrated architecture featuring a dedicated P2 detection head and an Adaptive Spatial Feature Fusion (ASFF) mechanism.

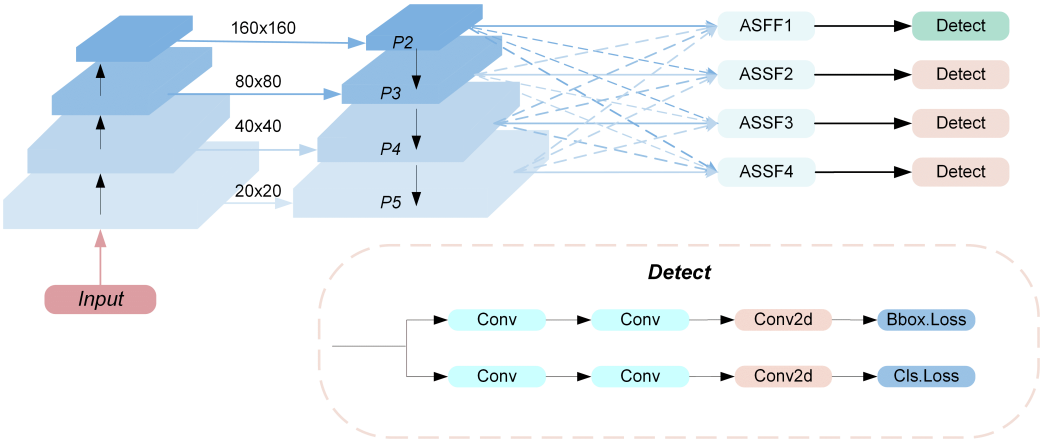


Figure 9. The structure of ASFFDHead.

The extended detection scale introduces a high-resolution P2 head specifically designed for small objects while maintaining the original P3-P5 heads for comprehensive scale coverage. Each detection head follows an identical structure, ensuring consistent processing across scales. The core challenge addressed by ASFF is the inconsistency or conflict that arises during multi-scale feature fusion in the detection head. Features from different pyramid levels (P2–P5) possess varying resolutions and semantic strengths. Shallow features (P2) excel in spatial detail but lack semantic richness, while deep features (P5) are semantically strong but spatially coarse. Therefore, when only naively fused, features from one level might contradict those from another level at the same spatial location—especially when that location corresponds to a small object primarily detectable at a specific scale. For example, a high-activation pixel in a shallow feature map (P2) indicating a small object might spatially coincide with a low-activation (background) pixel in a deeper feature map (P5) at the same location after resizing. This conflict confuses the detector during training and inference. The ASFF mechanism resolves this by adaptively learning spatial weight maps that dynamically filter out conflicting information and emphasize consistent, discriminative clues from each level during fusion. Crucially, these weights are learned directly from the data through standard backpropagation, allowing the network to discover the optimal fusion strategy for suppressing inconsistency.

As shown in Figure 10, we demonstrate the operating principle of the ASFF mechanism. Features from all source levels $n \in (P2, P3, P4, P5)$ are first resized (via interpolation for upsampling or strided convolution or pooling for downsampling) to match the spatial dimensions of the target level l. Simultaneously, 1×1 Covs are applied to ensure all resized features have the same number of channels. Subsequently, ASFF began adaptive fusion, and the overall fusion process formula is as follows:

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot x_{ij}^{n \to l} + \beta_{ij}^{l} \cdot x_{ij}^{n \to l} + \gamma_{ij}^{l} \cdot x_{ij}^{n \to l} + \delta_{ij}^{l} \cdot x_{ij}^{n \to l} \tag{5}$$

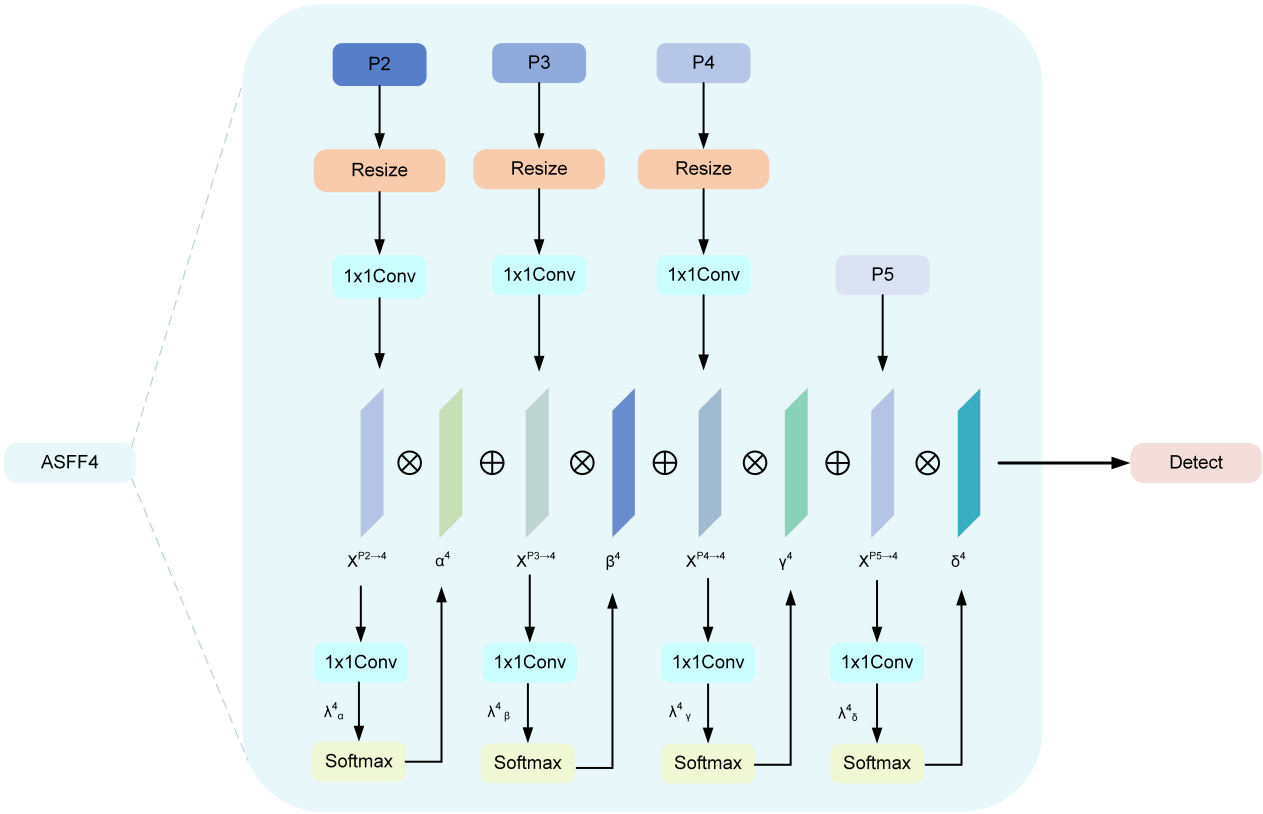


Figure 10. The details of ASFF4. The ⊗ represents dot product, and ⊕ represents fusion.

Here, $x_{ij}^{n \to l}$ denotes the feature vector adjusted to the size of the l-th layer on the j-th layer at the position (i,j). y_{ij}^{l} denotes the (i,j) vector of the inter-channel output feature map y^{l}. α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l}, and δ_{ij}^{l} are the spatial weights of the feature vectors learned by the network adaptively at the four different layers; notably, they are not predefined or heuristic, but rather learnable parameters derived from the feature maps themselves. We use the Softmax function to define them as follows:

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}} + e^{\lambda_{\beta_{ij}}^{l}} + e^{\lambda_{\gamma_{ij}}^{l}} + e^{\lambda_{\delta_{ij}}^{l}}} \tag{6}$$

Here, $\lambda_{\alpha_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$, $\lambda_{\gamma_{ij}}^{l}$, and $\lambda_{\delta ij}^{l}$ are the control parameters. For each source level, a dedicated 1×1 Conv layer is applied to its corresponding aligned feature map $x^{n \to l}$. This 1×1 Conv outputs a single-channel spatial map of control parameters ${\lambda_{ij}}^{l}$, used to fuse feature maps into each level. The purpose of softmax is to generate candidate weighting signals based on the ${\lambda_{ij}}^{l}$ the feature map $x^{n \to l}$ at each spatial location (i, j). The softmax operation ensures two key properties: The weights α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l}, and δ_{ij}^{l} are normalized between 0 and 1. Simultaneously, specify $\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} + \delta_{ij}^{l} = 1$. This normalization forces the network to compete for influence at each location. λ filters out conflicting information during the fusion process by controlling the weight parameters to approach 0 or 1.

By adaptively fusing features at all levels for each scale, irrelevant features are filtered out, while relevant features dominate by providing more distinctive clues. The modification results in considerably improved small object detection capabilities for the model.

4. Results

4.1. Experimental Basic Configuration

In Windows environment, we conducted all experiments. The YOLOv8 version is Ultralytics 8.1.20. In terms of hardware, the 13th Gen Intel(R) Core(TM) i9-13900K 3.00 GHz (Intel, CA, USA) was used for the CPU, the NVIDIA GeForce RTX 3090 (24 G) (Nvidia, CA, USA) was used for the GPU, and the software framework consists of Python3.8, torch1.10.0, and cuda11.6. During the training process, we started with pre-trained models and used a GPU for training. The input image size is 640 × 640. After multiple adjustments and attempts, we set epochs to 200 and patience to 150 to prevent resource waste. Subsequently, we set the batch size to 8, accelerating training time to the fullest extent; SGD as optimizer and control momentum to 0.937; and use mosaic for data enhancement. More experimental hyperparameters are shown in Table 1.

Table 1. Model hyperparameters.

Hyperparameters	Value
Image size	640 × 640
Epochs	200
Patience	150
Batch size	8
Optimizer	SGD
Momentum	0.937
Data enhancement	Mosaic
Workers	4
Learning rate	0.01
Weight decay	0.0005

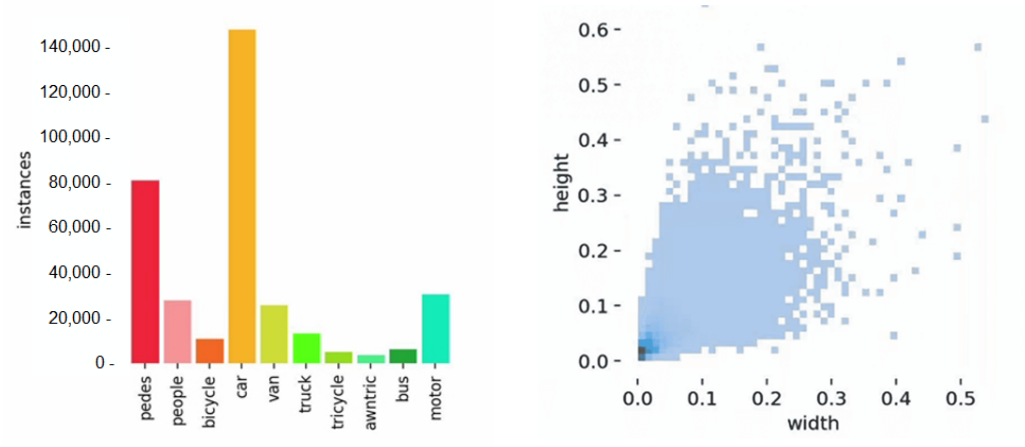
4.2. Dataset

We use the VisDrone2019 dataset tailored for UAV object detection, which is collected and produced by the AISKYEYE team in the Machine Learning and Data Mining Laboratory of Tianjin University with impartial authority. Unlike single-perspective collections, this dataset aggregates images from varied angles and mission contexts, including city streets, buildings, and plants, as shown in Figure 11, with more images for small objects. There are 10 classes in this dataset, as shown in Figure 12, 6471 images in the training set, 548 images in the validation set, and 1610 images in the test set, and the training set assigns class weights based on the number of corresponding labels. If we go to classify the object size according to the COCO [39] criterion, the proportion of small objects is more than 60% in the VisDrone2019 dataset, which is very suitable to be our benchmark dataset.



(**a**) (**b**)

Figure 11. (**a**) City street. (**b**) Building factory. The Chinese characters visible do not affect the scientific understanding.



(**a**) (**b**)

Figure 12. (**a**) Distribution of ten object categories. (**b**) Distribution of height and width of the objects.

4.3. Metrics

We evaluate the performance of our enhanced model using a set of metrics, including precision (P), recall (R), mAP@0.5:0.95, mAP@0.5, and an F_1 score, to facilitate a visual comparison of its effectiveness.

Precision (P) indicates the fraction of true positives among all predicted positives, whereas recall (R) quantifies the ability to find all relevant instances by computing the ratio of true positives to all actual objects. The F_1 score combines both as a harmonic mean to provide a balanced assessment. The mathematical formulas are as follows:

$$R = \frac{TP}{TP + FN} \tag{7}$$

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

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{9}$$

Here, True Positive (TP) indicates cases where the model correctly labels positive instances, False Negative (FN) refers to positive instances mistakenly classified as negative, and False Positive (FP) represents cases where the model erroneously predicts negative samples as positive.

The Mean Average Precision (mAP) provides a unified assessment of model performance by averaging the precision over all classes, offering a unified metric for model performance evaluation. The mathematical formulation is

$$mAP = \frac{1}{K} \sum_{i=1}^{K} AP_i \tag{10}$$

Here, K refers to the distinct classes, while Average Precision (AP) is employed to assess detection accuracy for each class individually, $AP_i = \int_0^1 P\left(R_i\right) dR$.

4.4. Comparison Experiments

The performance evaluation of SMA-YOLO involves three comparative experiments on the VisDrone2019 dataset. A direct comparison against the baseline YOLOv8n highlights the proposed model's performance gains. Further validation against other widely used YOLO variants confirms the effectiveness of the architectural improvements. Additional

benchmarks with recent YOLOv8n-based advanced models further establish the superiority of SMA-YOLO.

YOLOv8 comes in five different sizes. YOLOv8n, YOLOv8l, YOLOv8m, YOLOv8s, YOLOv8x. We balanced the accuracy and speed of model detection and combined it with our software and hardware environment, ultimately choosing YOLOv8n as our baseline model.

Table 2 shows the AP value of YOLOv8n and our model for each category, as well as the mAP@0.5 for all categories. As evidenced by the table data, our model's mAP@0.5 value increased by 7.4% for all categories, and the AP value of each category increased, with the three categories showing the greatest improvement being Pedes, Motor, and People, which are 12.7%, 10.0%, and 9.7%, respectively. These three categories all belong to the category of small objects, and the improvement in their detection accuracy confirms that SMA-YOLO's effectiveness in addressing small object detection challenges is specific to UAV-based images.

Table 2. Comparison results of SMA-YOLO with YOLOv8n. **(Bolding indicates the enhancement ratio)**.

Models	Pedes	People	Bicycle	Car	Van	Truck	Tricycle	Awntric	Bus	Motor	All
YOLOv8n	36.2	29.2	8.7	76.3	40.3	32.7	25.3	13.3	47.9	38.8	34.9
Ours	48.9	38.9	15.0	82.6	47.6	36.3	29.5	16.9	56.9	49.6	42.3
Increase(%)	**12.7**	**9.7**	**6.3**	**6.3**	**7.3**	**3.6**	**4.2**	**3.6**	**9.0**	**10.8**	**7.4**

Table 3 shows a comparative analysis between SMA-YOLO and other YOLO models. YOLOv5 and YOLOv7 underperform the baseline YOLOv8n on the VisDrone2019 dataset. YOLO-NAS adopts neural architecture search (NAS) to automatically optimize the backbone and head architecture, achieving a better balance between accuracy and efficiency for enhanced performance. Similarly, YOLOv10s and YOLOv11s achieve significant improvements over YOLOv8n through optimized architectures and advanced techniques. Subsequently, YOLOv12 builds upon YOLOv11 by introducing an attention-centric architecture with Area Attention and Residual Efficient Layer Aggregation Networks (R-ELAN), which reduces complexity and optimizes feature aggregation, achieving superior accuracy without sacrificing real-time speed. Notably, even YOLOv13s—the latest model in the YOLO series—proposes Hypergraph-based Adaptive Correlation Enhancement (HyperACE) for adaptive high-order correlation modeling via hypergraph computation, Full-Pipeline Aggregation-and-Distribution (FullPAD) for full-pipeline feature distribution, and depthwise separable convolutions for reducing computing overhead. Although it achieves a perfect 39.1% in mAP@0.5, considerably lower than SMA-YOLO's 42.3%, further validating its superior detection performance.

Table 3. Comparison results of SMA-YOLO with other YOLO models. **(Bolding indicates best results).**

Models	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (M)	GFLOPs
YOLOv5	30.6	15.6	7.2	17.1
YOLOv7 [40]	32.7	17.9	37.3	105.3
YOLOv8n	34.9	20.3	3.0	**8.2**
YOLO-NAS [41]	36.3	20.9	4.2	10.6
YOLOv10s [42]	37.8	22.4	7.2	20.9
YOLOv11s [43]	38.1	22.4	9.4	21.6
YOLOv12s [44]	38.8	23.0	9.3	21.4
YOLOv13s [45]	39.1	23.4	9.0	20.8
SMA-YOLO	**42.3**	**25.3**	**2.6**	20.9

A comparative analysis was conducted between SMA-YOLO and other advanced models to verify its superior effectiveness. As shown in Table 4. SSD demonstrates relatively poor detection performance among the compared methods. NanoDet is a lightweight,

fast, and easy-to-deploy detector based on an anchor-free design and an efficient backbone network. It exhibits extremely lightweight characteristics but therefore has low detection performance in complex scenarios and small objects. Faster RCNN, a representative two-stage detector, shows only marginal improvement over SSD while still underperforming significantly in comparison with the baseline YOLOv8n model. The performance benchmark, YOLOv8n, and all comparative models listed in the table are enhanced architectures specifically optimized for UAV-based small object detection. Notably, each modified model achieves varying degrees of performance enhancement over YOLOv8n. Among all compared algorithms, YOLO-MARS exhibits the best overall performance. However, SMA-YOLO surpasses YOLO-MARS by 1.4% in mAP@0.5 and 1.9% in mAP@0.5:0.95, while maintaining a parameter reduction of 0.3M. These results demonstrate that although SMA-YOLO did not achieve the best results on parameters and GFLOPs, the comprehensive evaluation shows that SMA-YOLO achieves excellent detection performance, particularly in small object detection scenarios for UAV images, confirming its effectiveness and advantages in this specialized domain.

Table 4. Comparison results of SMA-YOLO with other improved YOLOv8 models. **(Bolding indicates best results)**.

Models	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (M)	GFLOPs
SSD	25.3	14.6	58.0	99.2
NanoDet [46]	26.5	15.4	**1.8**	**1.5**
Faster RCNN	29.0	17.8	165.6	118.6
YOLOv8n	34.9	20.3	3.0	**8.2**
LGFF-YOLO [47]	38.3	22.8	4.2	12.4
RFAG-YOLO [48]	38.9	23.1	5.9	15.7
DASSF [49]	39.6	23.5	8.5	23.9
YOLO-GE [50]	40.7	23.7	3.5	15.9
SSE-YOLO [51]	40.8	23.6	3.6	10.9
YOLO-MARS [52]	40.9	23.4	2.9	13.7
SMA-YOLO	**42.3**	**25.3**	2.6	20.9

4.5. Ablation Experiments

Two sets of ablation experiments are carried out to examine the effect of each individual improvement on model performance. The first phase evaluates the isolated impact of the SSC module through ablation studies, while the second phase systematically examines the collective contribution of all proposed enhancements. The test comparisons are performed on the VisDrone2019 dataset using YOLOv8n as the benchmark model in the same experimental setting.

For the enhancement of feature maps in the SSC module, we use SimAM, which computes 3D attentional weights to enhance features, and to verify its superiority, we conduct experiments under the same settings and conditions using the SE [53] attentional mechanism, which focuses on computing one-dimensional weights, and CBAM [54], which computes one- and two-dimensional weights.

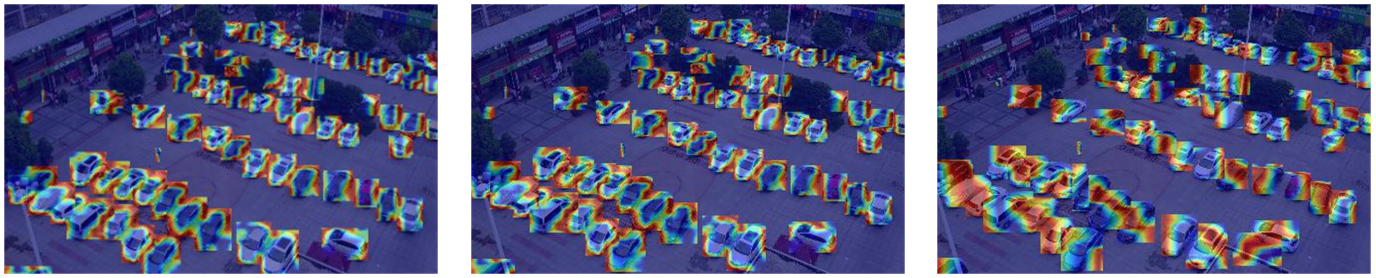
The ablation results are shown in Table 5; several key observations can be made regarding the effectiveness of various attention mechanisms. SimAM incurs a minimal increase in GFLOPs, primarily due to the calculation of channel statistics (mean $\hat{\mu}$ and variance $\hat{\sigma}^2$) and element-wise operations defined by its closed-form solution (Equation (1)). In contrast, SE and CBAM exhibit substantially higher computational costs, stemming from their parameterized layers (FC layers in SE, FC and Conv layers in CBAM). Most importantly, the parameter-free nature enables more efficient feature propagation during inference, and the lower computational complexity of SimAM preserves inference speed slightly higher than the baseline. SE and CBAM, however, suffer significant FPS degradation due to their

added parameters and more complex operations. Although the mAP@0.5 gain of SimAM (+1.4%) is only marginally higher than SE (+0.2%) and CBAM (+1.0%), this improvement is achieved with negligible parameter growth and minimal impact on inference latency. This exceptional efficiency makes SimAM particularly suitable for resource-constrained UAV platforms where real-time performance is paramount. The slight variance in mAP gains could be attributed to SimAM's 3D attention mechanism being inherently better at preserving fine-grained details critical for small objects.

Table 5. The results of introducing different attention mechanisms into SSC. **(Bolding indicates the best performance)**.

Models	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	Learnable Parameters	GFLOPs	FPS
YOLOv8n	45.8	34.3	34.9	3.0	–	**8.2**	189
+SE	45.9	34.6	35.1	4.1	2FC	10.7	172
+CBAM	46.6	35.2	35.9	4.5	FC + Conv	12.9	164
+SimAM	**46.9**	**35.6**	**36.3**	**3.0**	**0**	8.3	**197**

We employ the Grad-CAM technique to generate heatmaps, facilitating a visual comparison of the detection outcomes across three different attention mechanisms under identical scenes. As illustrated in Figure 13, a representative scenario containing numerous vehicles and substantial background interference is selected. The red regions in the activation maps denote areas of high response. Overall, the detection performance of SimAM exhibits clear superiority over the other two mechanisms. While the integration of the SE module enables the model to localize targets, the high-activation regions are predominantly concentrated along the object boundaries. When introducing CBAM, a broader distribution of red regions is observed; however, these activations still largely remain along the peripheries of the targets. Additionally, both SE and CBAM show weak activation responses for vehicles affected by background clutter. In contrast, the introduction of SimAM results in a marked increase in the number and intensity of activation regions, which are primarily focused on the central areas of the objects, indicating more accurate recognition. Furthermore, SimAM more effectively emphasizes vehicles partially occluded by trees or situated at considerable distances from the camera. These results demonstrate the superior capability of SimAM in enhancing object detection performance compared with SE and CBAM.



(**a**) (**b**) (**c**)

Figure 13. Heatmap results of introducing different attention in SSC. (**a**) +SE. (**b**) +CBAM. (**c**) +SimAM.

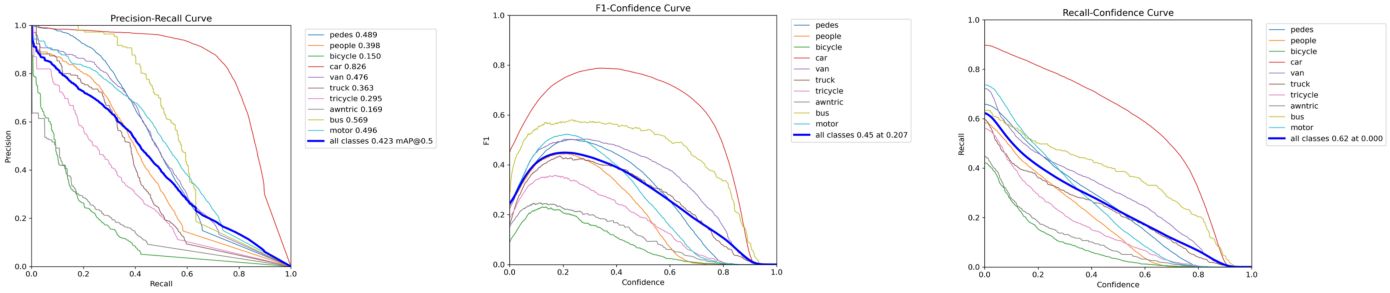
The tabulated results in Table 6 confirm significant improvements across all metrics of detection accuracy when contrasting the fully optimized model with the original YOLOv8n. Specifically, 7.2% in precision (P), 6.6% in recall (R), and 7.4% in mAP@0.5, as well as achieving a 0.4 M reduction in parameters. The detailed analysis of the architecture modification shows that the model detection accuracy is greatly improved after the introduction of

M-FPN and ASFFDHead. M-FPN replaces the original PANet-style feature fusion with a more streamlined and efficient structure, which focuses on scale-aligned fusion while avoiding redundant convolutions, thus saving parameters without compromising feature quality. Subsequently, ASFFDHead's attention-based adaptive spatial fusion mechanism helps the model filter conflict information while emphasizing discrimination information, leading to more accurate prediction. But at the same time, its multi-branch structure, in which each input feature from different scales is processed by an independent convolution layer, also introduces more parameters. Notably, the increase in GFLOPs mainly stems from the M-FPN and ASFFDHead components. M-FPN introduces additional multi-scale feature fusion paths, which significantly increase computations on feature maps. Meanwhile, ASFFDHead replaces the YOLOv8 detection head with a more complex fusion-based structure, introducing more layers and inter-feature computations. These changes, while improving small object perception, inevitably lead to higher GFLOPs and reduced FPS. The introduction of SSC slightly improves the decline of FPS. Nevertheless, FPS is still above 100, and while maintaining acceptable real-time performance, we have achieved significant improvements in detection accuracy.

To further validate these findings, as illustrated in Figure 14, we conducted a comprehensive performance analysis through Precision–Recall (P–R) curve, F1–Confidence (F1–C) curve, and Recall–Confidence (R–C) curve. These results collectively confirm that our proposed model architecture is both robust and effective.

Table 6. Ablation study of key components in the improved model. ASFFDH means ASFFDHead. **(Bolding indicates the best performance)**.

Models	SSC	M-FPN	ASFFDH	P (%)	R (%)	mAP@0.5 (%)	F1 (%)	Params (M)	GFLOPs	FPS
YOLOv8n				45.8	34.3	34.9	38	3.0	**8.2**	189
1	✓			46.9	35.6	36.3	38	3.0	8.3	**197**
2		✓		48.2	36.6	38.0	41	2.2	14.9	150
3			✓	48.8	36.9	38.3	41	4.3	17.6	137
4	✓		✓	50.4	39.2	41.5	42	4.3	17.6	142
5	✓	✓		50.1	38.9	41.1	42	**2.0**	14.9	156
6		✓	✓	51.7	40.1	42.1	44	2.6	20.9	102
SMA-YOLO	✓	✓	✓	**53.0**	**40.9**	**42.3**	**45**	2.6	20.9	106



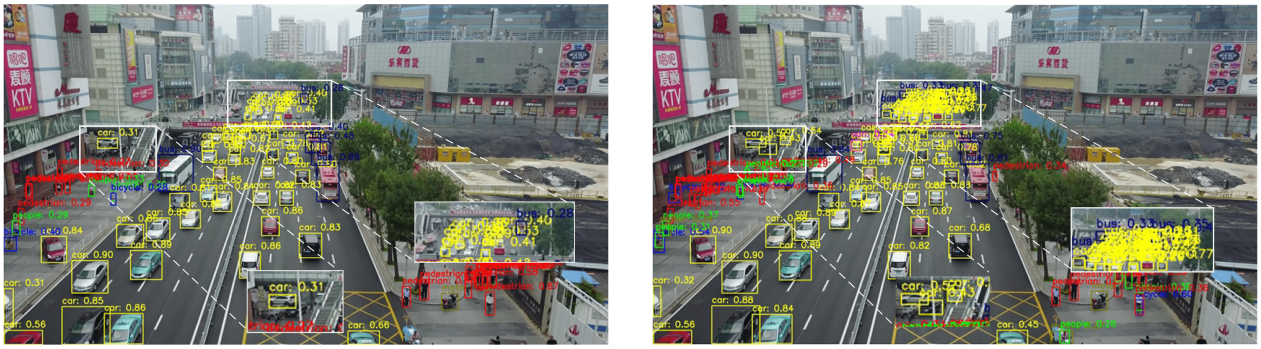
(**a**) (**b**) (**c**)

Figure 14. (**a**) Precision–Recall curve. (**b**) F1–Confidence curve. (**c**) Precision–Confidence curve.

4.6. Visualization

We present intuitive visual comparisons between SMA-YOLO and baseline YOLOv8n to demonstrate its enhanced effectiveness for UAV small object detection, visualizing its detection results with those of the baseline model YOLOv8n. Two representative images from the VisDrone2019 dataset are selected, each featuring distinct scenes, camera perspectives, and lighting conditions. Regions with notable differences in detection results are

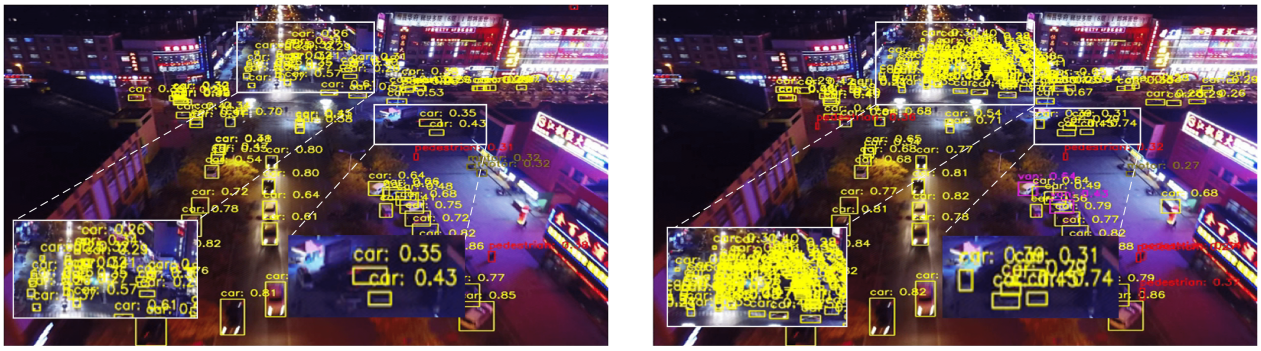
highlighted using white rectangular boxes and enlarged for clearer observation. The visualizations include object class names and confidence scores, with different colors indicating different categories. Figures 15 and 16 show visualization results under two markedly different conditions.



(**a**) (**b**)

Figure 15. An aerial image of urban streets acquired at an oblique angle during the daytime. (**a**) YOLOv8n. (**b**) SMA-YOLO.

Figure 15 depicts a UAV image captured during the daytime at an oblique angle over an urban street with a complex background. Both models detect most of the objects. However, YOLOv8n exhibits clear limitations in identifying small and distant objects. At least four regions in the image reveal substantial discrepancies, of which two representative areas are emphasized with white rectangles. In the upper section, where vehicles are densely packed, SMA-YOLO accurately detects nearly all of them, including a bus partially occluded by other objects. Additionally, cars in the upper-left corner, partially hidden by streetlights and stairs, are still correctly identified by SMA-YOLO. These results directly show the detection results of SMA-YOLO for occluded or small, dense objects, highlighting its superior detection sensitivity and accuracy.



(**a**) (**b**)

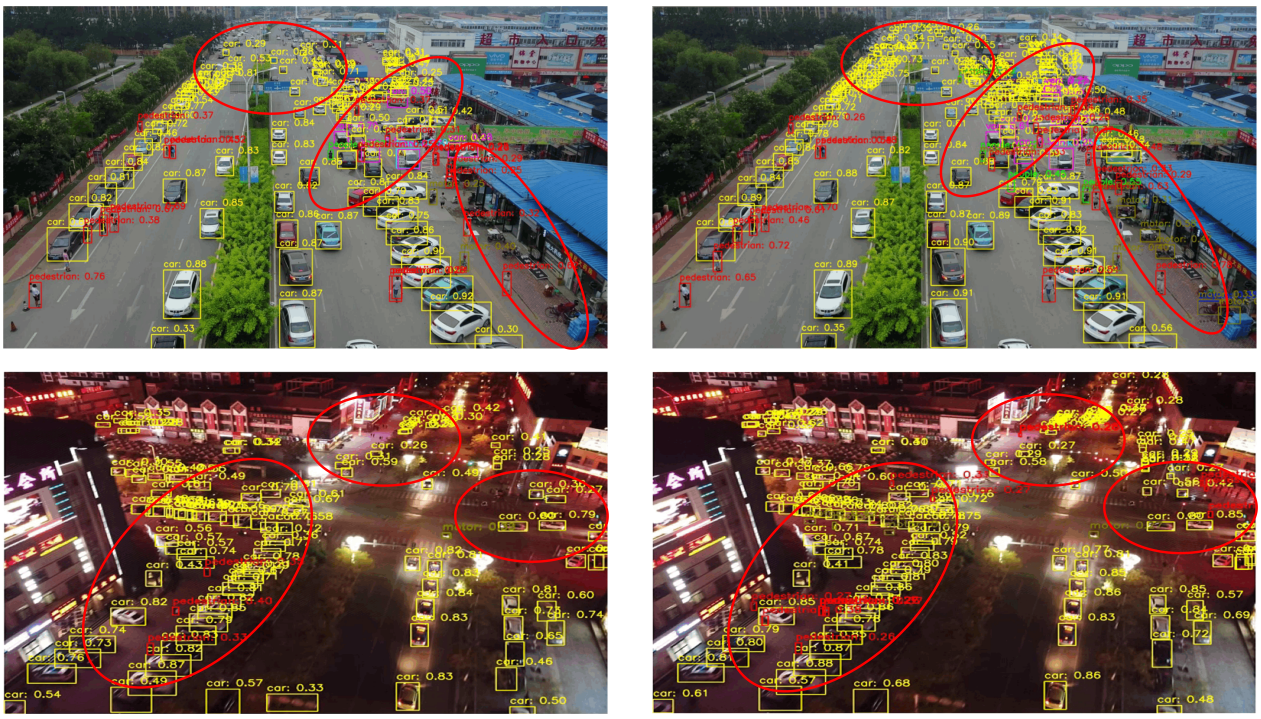
Figure 16. An aerial image of city intersection with poor and blurry lighting during the night. (**a**) YOLOv8n. (**b**) SMA-YOLO.

In contrast, Figure 16 shows a nighttime UAV image with poor visibility, motion blur, and minimal lighting conditions that pose a significant challenge for object detection. While both models perform similarly in the lower part of the image, the upper region reveals stark differences. Although several discrepancies are evident, we focus on two representative areas. In the central upper portion, a group of cars with headlights is located at a distance. Despite adverse factors such as distance, glare, and blur, SMA-YOLO successfully detects most vehicles, whereas YOLOv8n fails to identify the majority. In the top-right area, which lacks illumination, YOLOv8n detects only two vehicles, while SMA-YOLO correctly

identifies all. These results demonstrate the robustness of SMA-YOLO under challenging conditions such as low light, poor visibility, and image degradation.

To further validate the effectiveness of SMA-YOLO, an additional image similar to each of the two scenes is selected for supplementary comparison, as shown in Figure 17. In these cases, we focus on overall detection outcomes rather than analyzing specific regions. Areas with significant performance gaps are marked using red ellipses to highlight the differences.

We present the Precision–Confidence (P–C) curves for both YOLOv8 and our proposed SMA-YOLO in Figure 18. As shown, SMA-YOLO exhibits superior precision across a wide range of confidence thresholds. The blue curve, representing the average across all classes, is consistently higher and smoother in SMA-YOLO, indicating more reliable confidence calibration and fewer false positives. Moreover, SMA-YOLO achieves 100% precision at a slightly higher confidence threshold, implying a better capability to distinguish correct detections at high-confidence levels. Several small-object categories, such as pedes, people, and bicycles, also show significantly improved performance. In contrast, YOLOv8 exhibits unstable behavior in categories like awntri and tricyle, where the precision drastically drops at higher confidence levels. These results confirm that SMA-YOLO not only improves detection accuracy but also provides more trustworthy confidence estimates. However, the marginal improvement in the extreme confidence range suggests limited gains in suppressing false positives at very high thresholds. This implies that while SMA-YOLO calibrates confidence better than YOLOv8 overall, its capability to handle extremely ambiguous cases or rare object types still leaves room for further refinement.

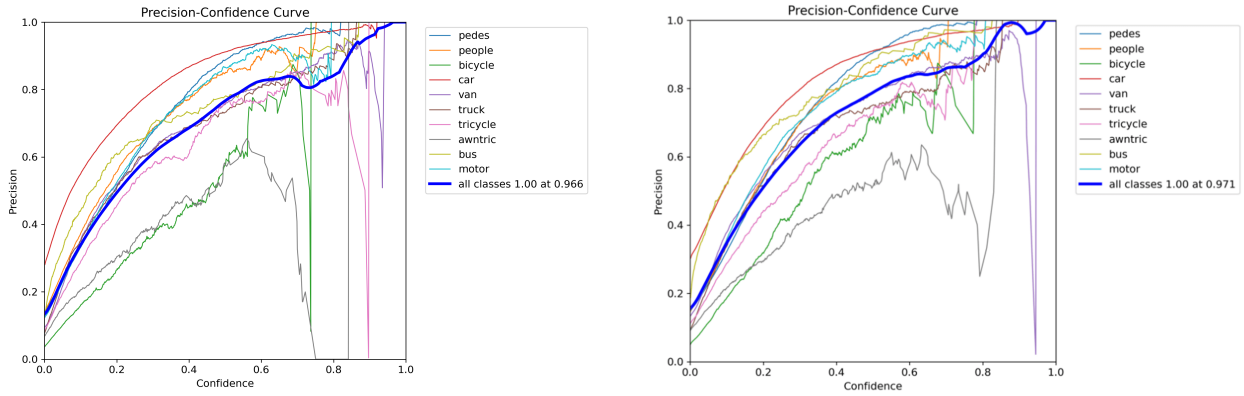


(**a**) (**b**)

Figure 17. (**a**) YOLOv8n. (**b**) SMA- YOLO.

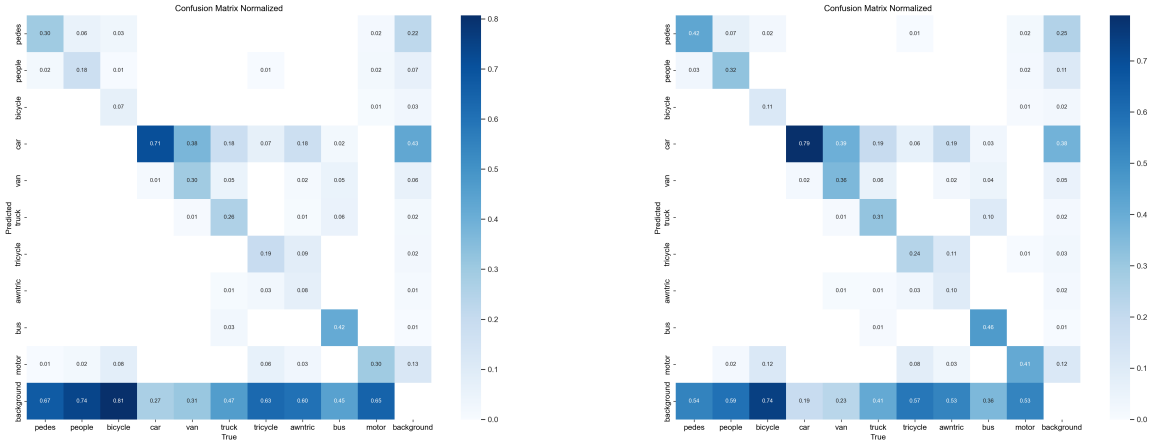
To offer a more detailed insight into the performance advantage of SMA-YOLO compared with the baseline YOLOv8, we conducted a comparative analysis using the confusion matrices of both models. As illustrated in Figure 19, an inspection of the diagonals reveals that YOLOv8 struggles to accurately recognize most categories, particularly those other than car and bus. In contrast, SMA-YOLO demonstrates noticeable improvements

across all categories, with the most significant gains observed in pedes, people, and motor, each achieving an increase of over 10%. These three classes predominantly consist of small objects, highlighting SMA-YOLO's enhanced capability in small object detection. Moreover, the background category has a substantial influence on the detection performance due to the dataset's diverse scenes and high background complexity. In many cases, target objects are easily misclassified as irrelevant background. However, SMA-YOLO significantly reduces the false positive rate associated with the background class, indicating its robustness in detecting small objects even under cluttered and complex visual environments.



(**a**) (**b**)

Figure 18. (**a**) Precision–Confidence curves of YOLOv8n. (**b**) Precision–Confidence curves of SMA-YOLO.



(**a**) (**b**)

Figure 19. (**a**) Confusion matrix of YOLOv8n. (**b**) Confusion matrix of SMA-YOLO.

4.7. Generalization Experiments

To demonstrate the generalization capability of our model, in addition to the VisDrone2019 dataset, we conducted comparative experiments on two remote sensing image datasets: UAVDT [55] and RSOD. The UAVDT dataset is a large-scale benchmark designed for object detection and tracking tasks in aerial videos captured by UAVs. It contains over 80,000 frames with annotations for vehicles (car, truck, bus) in challenging conditions such as high altitude, dynamic backgrounds, weather variations, and small object sizes. Moreover, the RSOD dataset is a remote sensing image benchmark focused on high-resolution aerial imagery. It includes four object categories—aircraft, oil tank, playground, and overpass—captured from satellite or aerial sensors, commonly used for evaluating detection performance on small and densely distributed objects.

The experimental settings, including training parameters and environments, were kept consistent with those used on VisDrone2019. We compared our proposed SMA-YOLO with NanoDet, Faster R-CNN, and various YOLO variants. As shown in Table 7, SMA-YOLO outperforms all models in both mAP@50 and mAP@50:95. Specifically, it surpasses YOLOv8n by 4.3% on UAVDT and 3.2% on RSOD in terms of mAP@50. Although our model does not achieve the lowest values in terms of parameters and GFLOPs, mainly due to comparison with extremely lightweight models such as NanoDet. Nevertheless, the slight increase in computational cost leads to a notable improvement in detection accuracy, which demonstrates the effectiveness of our design.

Furthermore, we conducted a qualitative comparison between the baseline and our improved model on the UAVDT and RSOD datasets to intuitively highlight the advantages of SMA-YOLO. As shown in Figures 20 and 21, the left and right respectively present the visualization results of YOLOv8 and SMA-YOLO on the UAVDT and RSOD datasets. In Figure 20, YOLOv8 incorrectly classifies a roadside solar panel as a car, misidentifies a car at the top of an intersection as a truck, and misses a truck on the right side. In contrast, SMA-YOLO accurately detects all of these instances. Moreover, as illustrated in Figure 21, the ground truth number of oil tank objects is 12, yet YOLOv8 falsely detects a building in the upper-right corner as an oil tank, whereas SMA-YOLO produces no false positives.

Table 7. The results of experiment on UAUDT and RSOD datasets. **(Bolding indicates the best performance)**.

Datasets	Models	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (M)	GFLOPs
UAVDT	SSD	23.9	15.5	58.0	99.2
	NanoDet	25.2	18.7	**1.8**	**1.5**
	Faster RCNN	27.5	21.3	165.6	118.6
	YOLOv5	28.3	21.5	7.2	17.1
	YOLOv7	29.8	21.9	37.3	105.3
	YOLOv8n	32.3	22.7	3.0	8.2
	YOLOv10s	33.4	23.1	7.2	20.9
	YOLOv13s	34.8	23.5	9.0	20.8
	SMA-YOLO	**36.6**	**24.1**	2.6	20.9
RSOD	SSD	91.5	68.4	58.0	99.2
	NanoDet	92.4	68.7	**1.8**	**1.5**
	Faster RCNN	93.7	69.5	165.6	118.6
	YOLOv5	94.8	69.2	7.2	17.1
	YOLOv7	93.9	68.9	37.3	105.3
	YOLOv8n	94.6	69.1	3.0	8.2
	YOLOv10s	94.7	69.2	7.2	20.9
	YOLOv13s	95.1	69.6	9.0	20.8
	SMA-YOLO	**97.8**	**70.2**	2.6	20.9

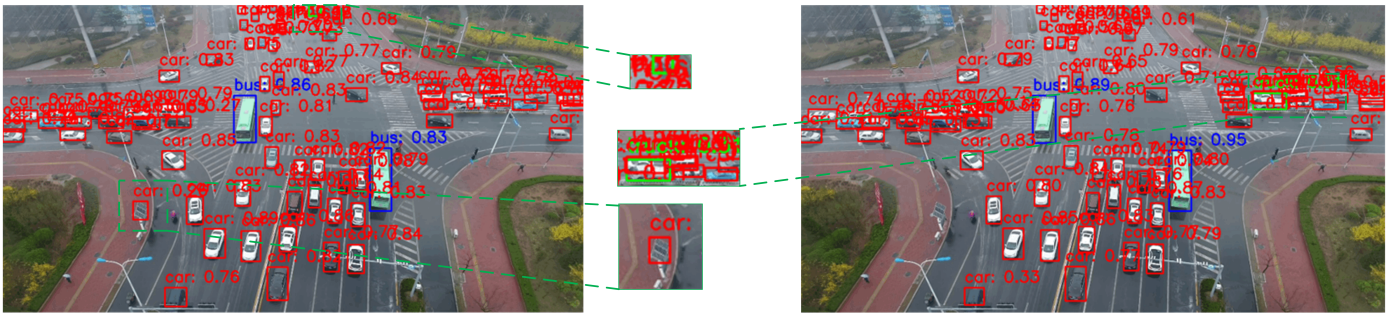


Figure 20. Visual comparison of YOLOv8 and SMA-YOLO on UAVDT dataset. (**left**) YOLOv8n, (**right**) SMA-YOLO.

The above experimental results and visual comparisons on the two datasets demonstrate that SMA-YOLO exhibits strong robustness and generalization capability. It effectively reduces both false positives and missed detections and achieves superior detection

performance across heterogeneous remote sensing conditions instead of being confined to a single dataset.

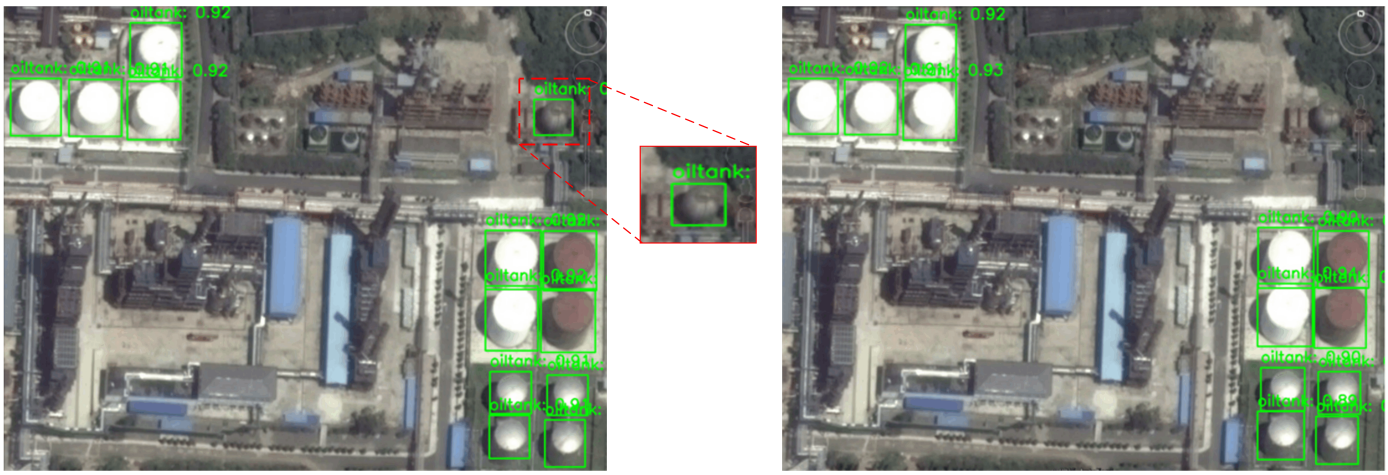


Figure 21. Visual comparison of YOLOv8 and SMA-YOLO on RSOD dataset. (**left**) YOLOv8n, (**right**) SMA-YOLO.

5. Dicussion

The proposed SMA-YOLO framework achieves notable improvements in small object detection within UAV imagery by systematically overcoming three critical limitations of existing methods. Specifically, the parameter-free SSC module preserves the discriminative features of small objects by leveraging spatial slicing and SimAM-guided 3D attention recalibration. As evidenced by the quantitative results in Table 5, the SSC module surpasses conventional parameterized attention mechanisms such as SE and CBAM, delivering a 1.4% improvement in mAP@0.5 without introducing additional learnable parameters while preserving real-time inference speed. Furthermore, the visualizations in Figure 13 highlight the module's capability to concentrate activation responses on object centers rather than on peripheral boundaries, thereby enhancing detection performance in cluttered and complex aerial scenarios.

The M-FPN architecture enhances cross-scale feature fusion by introducing multi-level connectivity alongside the novel C2f-HPC module. As demonstrated in Table 6, this design yields a 3.1% increase in mAP@0.5 compared with the baseline while simultaneously reducing the number of parameters by 26.7%. The hierarchical receptive field expansion mechanism facilitates combinatorial feature integration across scales, effectively preserving fine-grained details that are critical for detecting densely distributed small objects. This advancement effectively addresses a fundamental limitation of conventional FPN structures, which often suffer from significant information degradation during successive downsampling operations.

The ASFFDHead addresses feature conflicts commonly encountered in multi-scale fusion by employing adaptive spatial weighting. Specifically, it incorporates a dedicated high-resolution P2 detection head and introduces ASFF fusion strategies, which jointly contribute to a substantial reduction in missed detections under challenging conditions such as nighttime scenes and motion blur, as illustrated in Figure 1. This design is particularly effective for small object detection, where inconsistencies between shallow and deep feature representations often hinder accurate localization. Furthermore, the model's generalization ability is validated through cross-dataset evaluations presented in Table 7, where consistent performance gains—4.3% mAP@0.5 on UAVDT and 3.2% on RSOD—affirm the robustness and adaptability of the proposed design.

Although SMA-YOLO significantly enhances small object detection performance in UAV and remote sensing imagery, it introduces negative effects, such as increased computational complexity (GFLOPs) and a reduction in inference speed (FPS), both of which are critical considerations for deployment on resource-constrained aerial platforms. Future research will focus on reducing computational overhead and mitigating FPS degradation without compromising detection accuracy. In particular, model compression techniques and dynamic inference mechanisms are promising directions that may further optimize the efficiency of the proposed framework.

6. Conclusions

In this study, in order to deal with the vital issues of missed and false detection in UAV scenarios for small objects, particularly with low resolution and complex backgrounds. We propose SMA-YOLO, an enhanced YOLOv8-based detector incorporating several architectural innovations to improve small object detection performance. The proposed parameter-free SSC module enhances feature representation through strategic spatial slicing and SimAM-based feature enhancement, effectively preserving fine details of small objects while robustly suppressing irrelevant background features. To optimize multi-level and cross-scale feature fusion, we introduce an M-FPN with cross-layer connections for efficient information flow, complemented by the C2f-HPC module for hierarchical multi-scale feature extraction. Further improvements are achieved through our novel ASFFDHead design, which adds a high-resolution P2 detection head and introduces an ASFF mechanism to dynamically resolve scale conflicts during feature aggregation. Comprehensive evaluations on the VisDrone2019 dataset demonstrate superior performance of SMA-YOLO, achieving a 7.4% improvement in mAP@0.5 over YOLOv8n while reducing model parameters by 0.4 M. Meanwhile, we also demonstrated its generalization on the UAVDT and RSOD datasets. Future research directions include exploring model compression techniques for real-time edge deployment and investigating dynamic inference mechanisms, potentially further advancing UAV-based detection capabilities with a lighter and faster model.

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

Funding: This research was partially supported by the National Natural Science Foundation of China (Grant No. 12201185).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors thank all the editors and anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

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

References

1. Ahirwar, S.; Swarnkar, R.; Bhukya, S.; Namwade, G. Application of drone in agriculture. *Int. J. Curr. Microbiol. Appl. Sci.* **2019**, *8*, 2500–2505. [CrossRef]
2. Cvitanić, D. Drone applications in transportation. In Proceedings of the 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 23–26 September 2020; pp. 1–4.
3. Shahmoradi, J.; Talebi, E.; Roghanchi, P.; Hassanalian, M. A comprehensive review of applications of drone technology in the mining industry. *Drones* **2020**, *4*, 34. [CrossRef]

4. Fan, B.; Li, Y.; Zhang, R.; Fu, Q. Review on the technological development and application of UAV systems. *Chin. J. Electron.* **2020**, *29*, 199–207. [CrossRef]
5. Yao, H.; Qin, R.; Chen, X. Unmanned aerial vehicle for remote sensing applications—A review. *Remote Sens.* **2019**, *11*, 1443. [CrossRef]
6. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Vehicle detection from UAV imagery with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6047–6067. [CrossRef]
7. Al-lQubaydhi, N.; Alenezi, A.; Alanazi, T.; Senyor, A.; Alanezi, N.; Alotaibi, B.; Alotaibi, M.; Razaque, A.; Hariri, S. Deep learning for unmanned aerial vehicles detection: A review. *Comput. Sci. Rev.* **2024**, *51*, 100614. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
9. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
12. Ross, T.Y.; Dollár, G. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
15. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
16. Varghese, R.; Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6.
17. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
18. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
19. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
21. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
22. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
23. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
24. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
25. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A survey of object detection for UAVs based on deep learning. *Remote Sens.* **2023**, *16*, 149. [CrossRef]
26. Zhang, X.; Izquierdo, E.; Chandramouli, K. Dense and small object detection in uav vision based on cascade network. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
27. Yi, H.; Liu, B.; Zhao, B.; Liu, E. Small object detection algorithm based on improved YOLOv8 for remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1734–1747. [CrossRef]
28. Zand, M.; Etemad, A.; Greenspan, M. Oriented bounding boxes for small and freely rotated objects. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]

29. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference On Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
30. Xu, C.; Wang, J.; Yang, W.; Yu, L. Dot distance for tiny object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1192–1201.
31. Zhan, W.; Sun, C.; Wang, M.; She, J.; Zhang, Y.; Zhang, Z.; Sun, Y. An improved Yolov5 real-time detection method for small objects captured by UAV. *Soft Comput.* **2022**, *26*, 361–373. [CrossRef]
32. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
33. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
34. Sun, T.; Chen, H.; Liu, H.; Deng, L.; Liu, L.; Li, S. DS-YOLOv7: Dense Small Object Detection Algorithm for UAV. *IEEE Access* **2024**, *12*, 75865–75872. [CrossRef]
35. Feng, F.; Hu, Y.; Li, W.; Yang, F. Improved YOLOv8 algorithms for small object detection in aerial imagery. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102113. [CrossRef]
36. Webb, B.S.; Dhruv, N.T.; Solomon, S.G.; Tailby, C.; Lennie, P. Early and late mechanisms of surround suppression in striate cortex of macaque. *J. Neurosci.* **2005**, *25*, 11666–11675. [CrossRef]
37. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
38. Yu, Y.; Zhang, Y.; Cheng, Z.; Song, Z.; Tang, C. Multi-scale spatial pyramid attention mechanism for image recognition: An effective approach. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108261. [CrossRef]
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
40. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
41. Aharon, S.; Dupont, L.; Masad, O.; Yurkova, K.; Fridman, F.; Lkdci; Khvedchenya, E.; Rubin, R.; Bagrov, N.; Tymchenko, B.; et al. Super-Gradients. 2021. Available online: https://zenodo.org/records/7789328 (accessed on 1 July 2025).
42. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
43. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
44. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
45. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733. [CrossRef]
46. Lyu, R. Nanodet-Plus: Super Fast and High Accuracy Lightweight Anchor-Free Object Detection Model. 2021. Available online: https://github.com/RangiLyu/nanodet (accessed on 1 July 2025).
47. Peng, H.; Xie, H.; Liu, H.; Guan, X. LGFF-YOLO: Small object detection method of UAV images based on efficient local–global feature fusion. *J. Real-Time Image Process.* **2024**, *21*, 167. [CrossRef]
48. Wei, C.; Wang, W. RFAG-YOLO: A Receptive Field Attention-Guided YOLO Network for Small-Object Detection in UAV Images. *Sensors* **2025**, *25*, 2193. [CrossRef] [PubMed]
49. Li, H.; Qu, H. Dassf: Dynamic-attention scale-sequence fusion for aerial object detection. In Proceedings of the International Conference on Computational Visual Media, Hong Kong, China, 19–21 April 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 212–227.
50. Yue, M.; Zhang, L.; Zhang, Y.; Zhang, H. An Improved YOLOv8 Detector for Multi-scale Target Detection in Remote Sensing Images. *IEEE Access* **2024**, *12*, 114123–114136. [CrossRef]
51. Lu, Y.; Sun, M. SSE-YOLO: Efficient UAV Target Detection With Less Parameters and High Accuracy. *Preprints* **2024**, 2024011108.. [CrossRef]
52. Zhang, G.; Peng, Y.; Li, J. YOLO-MARS: An Enhanced YOLOv8n for Small Object Detection in UAV Aerial Imagery. *Sensors* **2025**, *25*, 2534. [CrossRef]
53. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

54. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
55. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.