AIMS Mathematics

AIMS Mathematics, 11(2): 5172–5191.
DOI: 10.3934/math.2026211
Received: 02 November 2025
Revised: 06 February 2026
Accepted: 12 February 2026
Published: 28 February 2026

https://www.aimspress.com/journal/Math

Research article

FCA-YOLO: A small object detection method based on feature attention fusion for UAV remote sensing images

Qiming Li[1], **Yonghui Yan**[1] **and Shaohui Lan**[2,*]

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 20136, China
[2] China Unicom Data Intelligence Co, Ltd, Beijing 100800, China

* **Correspondence:** Email: mm2013_shmtu@163.com; Tel: +8602138282823.

Abstract: Small object detection remains a significant challenge in computer vision, especially in scenarios such as remote sensing and unmanned aerial vehicles (UAVs) application, where there is considerable room for improvement in detection accuracy. The difficulty primarily arises from factors such as low resolution of the images, complex backgrounds in the images, and insufficient feature representation. To address these challenges, we propose FCA-YOLO, a novel detection framework built upon YOLOv11, specifically optimized for small object detection from UAV perspectives. First, a down-sampling structure was designed to better preserve small object features, which integrated a redundant compression feature transformation strategy with an inverted bottleneck residual block to enhance feature flow and representation capacity. Second, we propose a cross-scale feature fusion module that integrates spatial and channel attention mechanisms to effectively align and optimize multi-scale features, thereby enhancing the model's focus on small objects. Finally, a specialized detection structure was designed to enhance sensitivity to small targets, combining a dedicated detection head with skip connections that fused deep semantic features and shallow details, thereby improving the model's ability to capture fine-grained small information. Experimental results on the VisDrone2019 dataset demonstrated that FCA-YOLO outperforms the baseline model, achieving improvements of 3.1% in precision, 4.7% in recall, and 5% in mAP@0.5, while reducing the number of parameters by 30%. Compared with other YOLO variants and state-of-the-art algorithms, the proposed method achieves superior performance in terms of detection accuracy. Further evaluations on the DOTAv1.0 and VEDAI datasets validated the robustness and consistent detection performance of the proposed model across aerial imaging scenarios.

Keywords: attention mechanism; remote sensing images; small object detection; unmanned aerial

vehicles (UAVs); YOLO
Mathematics Subject Classification: 68T07, 68U10

1. Introduction

As one of the core tasks in computer vision, object detection has demonstrated significant research value and technical challenges in applications based on unmanned aerial vehicle (UAV) platforms. The high mobility and wide-area coverage capabilities of UAVs offer dynamic perspectives and complex scene data, driving the development of UAV-based object detection in a range of real-world scenarios, such as public safety and emergency response [1], agricultural monitoring [2], power line inspection [3], and urban management [4]. However, UAV images are often affected by factors such as flight altitude and varying viewpoints, resulting in targets that are typically small, irregular shaped, and frequently occluded. These characteristics increase the difficulty of feature extraction and raise the false detection rate, particularly in dense scenes and long-range small object detection tasks.

Although researchers have sought to alleviate these issues through multi-scale feature fusion [5], attention mechanisms [6], and optimized upsampling strategies [7], critical challenges remain. These include spatial details loss due to downsampling, diminished shallow feature representation during feature fusion, and limited sensitivity of detection heads to small object details. To address these challenges, we propose an enhanced method aimed at improving fine-grained feature representation and suppressing background interference, thereby boosting small object detection performance. The major contributions of this work are as follows:

- A lightweight inverted bottleneck residual module is designed for small-scale feature preservation. It adopts an expand-and-squeeze structure with residual connections to efficiently retain crucial details of small objects.
- A novel three-branch feature attention fusion module is proposed, which aligns and integrates features from multiple layers to enhance the model’s focus on target regions.
- A fine-grained high-resolution structure is designed and incorporated into the network, which enhances the feature representation of small targets and controls model complexity by integrating a high-resolution detection head, backbone compression, and cross layer feature fusion.

2. Related work

Early object detection methods were primarily based on handcrafted features and traditional machine learning classifiers. A representative method is the Histogram of Oriented Gradients [8] descriptor combined with a Support Vector Machines [9] classifier, which achieved early success in pedestrian detection. Another influential method is the Deformable Part-based Model [10], which represents objects as a collection of parts with geometric constraints, providing improved robustness to deformation and partial occlusion. These traditional methods typically followed a multi-stage pipeline involving feature extraction, classification, and post-processing. However, their reliance on manually designed features limited their ability to cope with large variations in object appearance, cluttered backgrounds, and complex scenes.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), brought a transformative shift to the field of object detection. CNNs enabled end-to-end learning of feature representations directly from raw image data, eliminating the need for manual feature design and substantially improving detection accuracy and generalization. Current object detection methods are broadly categorized into two-stage and one-stage approaches. Two-stage detectors, such as Faster R-CNN [11], first generate region proposals, followed by feature extraction, classification, and bounding box regression. Although these methods typically yield high accuracy, their complex architecture and slower inference speed hinder their applicability in scenarios that require high response speed. In contrast, one-stage detectors bypass region proposals by directly predicting object classes and bounding boxes from the input image, thereby achieving faster inference suitable for real-time applications. Representative one-stage detectors include SSD [12], RetinaNet [13], CenterNet [14], EfficientDet [15], and the YOLO [16] series. The YOLO series has evolved to its latest version, YOLOv11 [17], which achieves an effective balance between detection accuracy and inference speed.

Many enhancements have been proposed based on the YOLO framework to improve the performance of small object detection. To better preserve the fine-grained features of small objects, YOLO-Extract [18] enhances the shallow feature extraction capability of YOLOv5 by incorporating residual structures and dilated convolutions. Similarly, Zhang et al. [19] proposed a multi-branch dilated convolution module with optimized channel reweighting and feature fusion strategies, effectively integrating local and global contextual information while suppressing background noise. To improve feature focus and separability, attention mechanisms and structured fusion methods have been explored. Li et al. [20] introduced YOLOSR-IST, which integrates coordinate attention with Swin Transformer blocks to strengthen focus on target regions in infrared remote sensing images. Wang et al. [21] applied deformable convolutions in feature fusion to enable sparse sampling and adaptive attention to key spatial regions, alleviating semantic bias and background interference. In addition, Kang et al. [22] proposed the Type-1 Fuzzy Attention module, which employs fuzzy entropy theory to suppress noise in feature maps and enhance detection robustness under uncertainty. To mitigate semantic compression and the loss of fine details caused by repeated downsampling, DSAA-YOLO [23] combines multi-scale downsampling with sub-pixel upsampling, preserving high-frequency details without significantly increasing computational cost. TPH-YOLOv5 [24] enhances small, dense object detection in UAV image by integrating Transformer-based prediction heads, Convolutional Block Attention Modules, and extra detection heads, demonstrating strong robustness in scenarios involving scale variation and motion blur. Lightweight design has also received attention. Gong et al. [25] proposed a simplified variant of YOLOv3-tiny tailored for thermal imaging, which significantly reduces parameter size and computational overhead while maintaining competitive performance. Shen et al. [26] developed CA-YOLO, which incorporates a cascaded pooling structure to reduce redundant computation via information reuse, thereby improving multi-scale feature representation.

In summary, methods have improved YOLO-based small object detection from various perspectives. However, challenges remain, including inadequate response to fine-grained shallow features, suboptimal fusion strategies, and low sensitivity of detection heads to small targets, particularly in UAV and remote sensing scenarios. To address these gaps, we propose a unified architecture that integrates fine-grained feature preservation, attention-guided fusion, and a high-resolution small object detection head to improve small object detection accuracy.

3. Method

The YOLOv11 [17] architecture mostly consists of three components: the Backbone, Neck, and Head. Compared to YOLOv8, YOLOv11 replaces the original C2f module with a more efficient C3k2 module, a Cross Stage Partial block with a kernel size of 2, thereby improving feature extraction and feature map partitioning capabilities. To further enhance detection performance under complex backgrounds and occlusion conditions, YOLOv11 introduces a novel C2PSA module, which integrates convolutional blocks with parallel spatial attention. This design strengthens the model's focus on critical regions by employing dual spatial attention branches.

We adopt the lightweight YOLOv11s as the baseline and propose FCA-YOLO, a framework designed to address the challenges of small object detection in UAV images. The overall architecture of FCA-YOLO is illustrated in Figure 1.

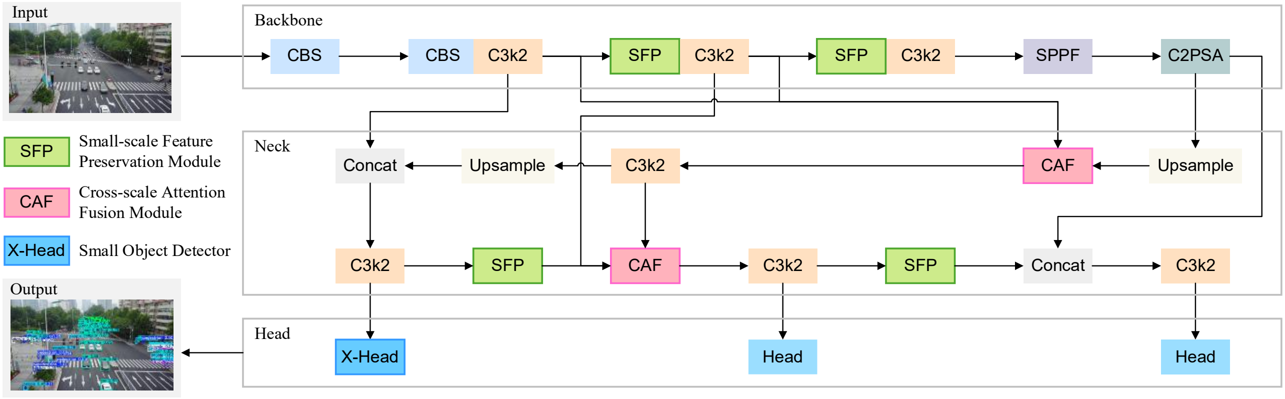

Figure 1. The architecture of the FCA-YOLO network.

First, a Small-scale Feature Preservation (SFP) module is designed and integrated into the key downsampling stages. This module uses an inverted bottleneck residual structure and an expand-and-squeeze convolution strategy to construct a new downsampling block. This design enables the retention of essential fine-grained details for small objects. Second, a three-branch Cross-scale Attention Fusion (CAF) module is proposed to preserve high-frequency edge information and align multi-scale features. By incorporating spatial and channel attention mechanisms, it enhances inter-feature interactions, reinforces salient information, and suppresses redundancy, thereby enriching the overall feature representation. Finally, in the proposed fine-grained high-resolution structure, the resolution of the detection head is increased to mitigate the loss of target information at low resolutions, the backbone network is compressed to reduce the computational burden, and skip connection is introduced to fuse fine-grained features, effectively compensating for the information loss caused by structural simplification and further enhancing the representational capacity of the high-resolution detection head.

3.1. Small-scale key feature preservation module

Although standard convolutional operations are effective in downsampling feature maps and expanding the receptive field, they are often inadequate for small object detection. This shortcoming stems from the fact that small objects typically occupy only a few pixels in the original image, making them highly sensitive to resolution loss. With each downsampling step, conventional convolutions reduce spatial resolution by half, which may result in the loss of critical fine-grained features. Consequently, small objects can become indistinct or disappear during successive downsampling stages.

Therefore, to better preserve small-scale features, the SFP module is designed, as illustrated in Figure 2. The module is inspired by the inverted bottleneck residual architecture from CMUNeXt [27] and incorporates Ghost convolutions [28] to enhance the model's capacity to model and retain essential features of small objects. The input and output feature maps are denoted X and Y, and are assumed to have dimensions of $X \in \mathbb{R}^{B\times C\times H\times W}$ and $Y \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}}$, respectively. Here, B represents the batch size, C is the number of channels, and H and W correspond to the height and width of the feature maps, respectively. X is processed by a depthwise convolution to generate a feature map F_1, which enhances intra-channel spatial awareness while suppressing inter-channel redundancy, thereby improving the extraction of key fine-grained features such as edges and textures of small objects. This process is described in (1).

$$F_1 = DWConv_{k\times k} * X. \tag{1}$$

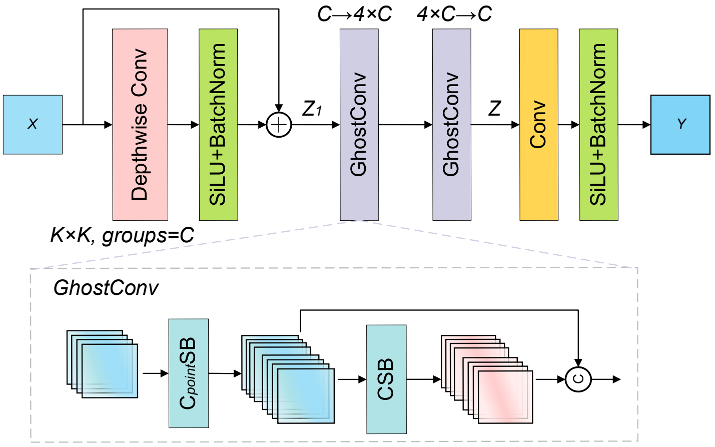

Figure 2. Structure of the SFP module.

The normalized feature map for channel c, denoted $\hat{F}_1^{(c)}$, is computed by (2).

$$\hat{F}_1^{(c)} = \gamma_c \cdot \frac{F_1^{(c)} - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}} + \beta_c, \mathrm{c} = 1, 2, \ldots, \mathrm{C}, \tag{2}$$

where μ_c and σ_c^2 represent the mean and variance of the feature map across spatial dimensions, respectively. ε is a constant that prevents division by zero. γ_c and β_c are learnable affine transformation parameters that restore the normalized feature. Thereafter the nonlinear characteristics are preserved through SiLU in (3).

$$F_2^{(c)} = \hat{F}_1^{(c)} \cdot \frac{1}{1+e^{-\hat{F}_1^{(c)}}}, \mathrm{c} = 1, 2, \ldots, \mathrm{C}. \tag{3}$$

Then, the $F_2^{(c)}$ of each channel is integrated, denoted F_2. Z_1, which is obtained by combining

F_2 with the residual connection from X. Subsequently, the SFP module applies two consecutive Ghost convolution layers for channel expansion and compression, thereby enhancing the semantic expressiveness of small object regions. The channel expansion operation is described in (4). $W_{1\times1}^{C\to4C}$ denotes the primary 1 × 1 convolution that expands the channel from C to $4C$, and $f_{cheap}^{(C)}$ represents a cheap convolution. Similarly, the second Ghost convolution in (5) projects the expanded features back to the original dimensionality, with $[:C]$ indicating channel truncation to the first C channels.

$$Z_2 = G_C^{(1)}(Z_1) = Concat\left(W_{1\times1}^{C\to4C} * Z_1, f_{cheap}^{(C)}\right)[:4C], \tag{4}$$

$$Z = G_C^{(2)}(Z_2) = Concat\left(W_{1\times1}^{4C\to C} * Z_2, f_{cheap}^{(C)}\right)[:C]. \tag{5}$$

Subsequently, the enhanced feature map Z is processed by a convolutional operation with a kernel size of 3 and stride of 2, followed by batch normalization (BN) and a SiLU activation, to obtain the downsampled feature map Y.

Overall, the SFP module refines spatial features through depthwise separable convolutions, enhances channel-wise representation using lightweight Ghost convolutions, and promotes stable information flow via residual connections. This design significantly strengthens the model's ability to capture and retain critical features of small objects, making it especially effective for detecting dense, small-scale, and occluded targets in complex scenes.

3.2. Cross-scale attention fusion module

Shallow features retain rich spatial details, while deep features provide strong semantic representations. Most mainstream detection models adopt simple operations such as concatenation or element-wise addition for multi-scale feature fusion. Although effective to some extent, these approaches often lead to misalignment and redundancy, which weaken the model's ability to detect small objects that rely on fine-grained details. To address these limitations, we propose a Cross-scale Attention Fusion (CAF) module, as illustrated in Figure 3.

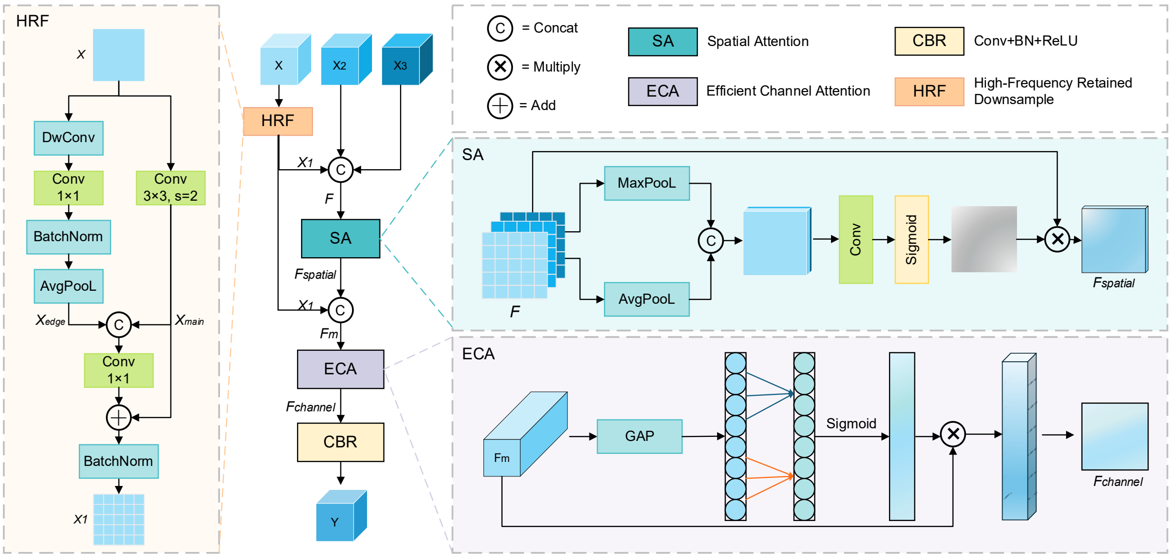

Figure 3. Structure of the CAF module.

This module integrates spatial and channel attention mechanisms to align and adaptively fuse features from different layers. By mitigating scale inconsistencies and enhancing the focus on object-relevant regions, CAF improves small object detection accuracy. To achieve scale alignment, the shallow feature maps are first reshaped. However, directly applying standard convolution may lead to the loss of fine-grained details, while pooling operations tend to blur or eliminate high-frequency components such as edges and contours. To address this issue, we design a High-Frequency Retained (HFR) downsampling module that preserves edge-level high-frequency information during spatial downsampling. This alleviates the degradation of fine details typically caused by conventional downsampling operations.

The HFR module consists of two branches. The main branch applies a standard 3 × 3 convolution with a stride of 2 to the input feature map *X*, producing the downsampled output X_{main}. The auxiliary branch serves as an edge-aware compensation path. It employs depthwise convolution to capture local spatial variations, followed by channel fusion and spatial resizing to generate an edge response map X_{edge}, which complements the main feature stream. As shown in (6), the outputs from both branches are concatenated along the channel dimension and fused via a 1 × 1 convolution. Then, a residual connection adds the fused features to X_{main}, followed by BN to stabilize training.

$$X_1 = BN\left(X_{main} + Conv_{1*1}\left([X_{main}; X_{edge}]\right)\right). \tag{6}$$

The output X_1 from the HFR module, along with two additional features X_2 and X_3 from different levels, forms a three-branch input. These features are concatenated along the channel dimension to form an aggregated feature *F*. A spatial attention mechanism is then applied to emphasize the locations of small objects. To capture informative spatial cues, average pooling and max pooling are applied along the channel dimension. The average pooled feature map F_{avg} captures the global spatial response, as defined in (7).

$$F_{avg}(1, h, w) = \frac{1}{C}\sum_{c=1}^{C} F_{c,h,w}. \tag{7}$$

The max pooled map F_{max}, as defined in (8), preserves the maximum response values at each spatial location, thereby focusing on the most salient features.

$$F_{max}(1, h, w) = \max_{1 \leq c \leq C} F_{c,h,w}. \tag{8}$$

Then, average and max pooled maps are concatenated along the channel axis to form a 2-channel spatial descriptor, as shown in (9).

$$F_{cat}(2, h, w) = Concat\left[F_{avg}, F_{max}\right]. \tag{9}$$

To generate the spatial attention map, the concatenated feature $F_{cat} \in \mathbb{R}^{2 \times H \times W}$ is processed by a convolutional layer followed by a sigmoid activation function, as defined in (10).

$$M_{spatial}(1, h, w) = \sigma\left(Conv_{7\times7}(F_{cat})\right). \tag{10}$$

This operation generates a single-channel spatial attention map $M_{spatial}$ that assigns importance weights to different spatial locations. It is subsequently applied to the input feature F via element-wise multiplication to produce the spatially enhanced feature $F_{spatial}$. Next, $F_{spatial}$ is concatenated with X_1 to form an intermediate feature F_m, which retains detailed information from shallow layers and helps the attention mechanism better localize target areas. To further model inter-channel dependencies, we incorporate the Efficient Channel Attention (ECA) [29] mechanism. As shown in (11), global average pooling is applied to each channel of the feature map F_m, resulting in a channel descriptor $Z \in \mathbb{R}^C$, where each element Z_c represent the mean activation of each channel over all spatial positions.

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_m^{(c)}(i,j)\,, c = 1,2, \dots, C. \tag{11}$$

A 1D convolution followed by a sigmoid activation is applied to Z to generate the learned channel attention weights W. These weights are multiplied with each channel of the input feature map F_m, resulting in the final fused output $F_{channel}$.

The CAF module effectively captures contextual and salient information from multi-path features, enhancing the representation of object contours and edges, particularly under complex backgrounds. By performing attention-guided cross-scale fusion, it improves the preservation and discrimination of critical features of small objects during detection.

3.3. Fine-grained high-resolution (FHR) structure

In traditional YOLO architecture, three detection heads are typically employed to generate feature maps with resolutions of 80 × 80, 40 × 40, and 20 × 20, respectively, enabling multi-scale object detection, as shown in Figure 4(a).

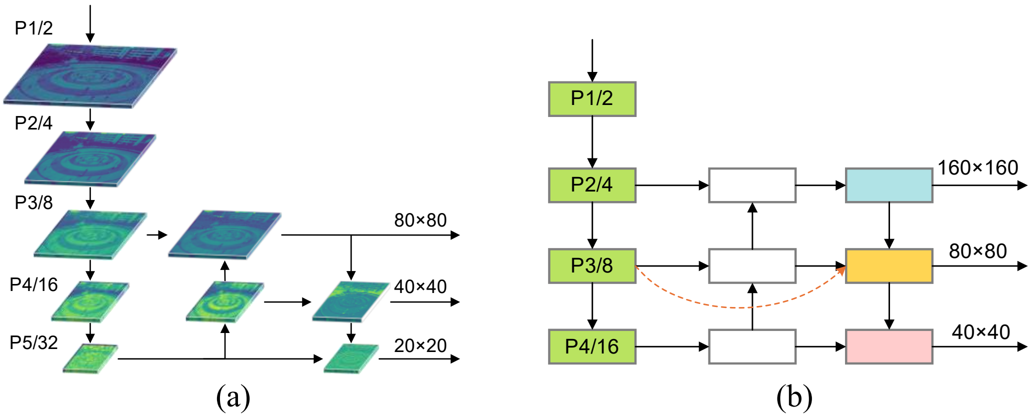

Figure 4. Structure of the detection heads. (a) YOLOv11 detection head and (b) FCA-YOLO detection head.

While this strategy improves adaptability to objects of varying sizes, it remains suboptimal for small object detection due to the limited spatial resolution and loss of fine-grained details in deeper layers. To overcome this limitation, we introduce a high-resolution detection head tailored for small object detection, operating on a 160 × 160 feature map. This detection head is built upon the P2 layer output from the backbone, which retains rich edge, texture, and positional information. These shallow, high-resolution features enhance the model's capability to accurately detect small and densely packed objects.

Experimental results indicate that incorporating the high-resolution detection head yields a 3.06% improvement in mAP@0.5, validating its effectiveness in enhancing small object detection. However, this gain comes at a cost: The model parameters increase by 11.58%, resulting in higher computational overhead and inference latency. Further analysis shows that the original 20 × 20 detection head depends on deep semantic features. While effective for high-level understanding, it exhibits limited sensitivity to fine-grained structural cues, which are essential for accurately detecting small-scale targets. To address this, we remove the original P5 detection head and introduce a high-resolution detection head to enhance the detection of small objects. To complement this structural adjustment, we further design a feature aggregation strategy based on skip connections. As illustrated in Figure 4(b), features from P3 are directly fused into the 80 × 80 detection branch, providing additional fine-grained details and mitigating the information loss caused by repeated downsampling.

This structural optimization enhances the model’s capacity to detect small-scale objects while suppressing interference from less relevant feature scales. Moreover, this strikes a balance between fine-grained detail preservation and semantic richness, making the model more robust in complex detection scenes.

4. Experiments

4.1. Dataset

We conduct experiments on the VisDrone2019 dataset [30] collected by the AISKYEYE team from the Machine Learning and Data Mining Lab at Tianjin University using various UAV platforms. The dataset contains a total of 8,629 images, with 6,471 for training, 548 for validation, and 1,610 for testing. All images are finely annotated and span 10 object categories across scenes involving varying weather conditions, lighting environments, and camera viewpoints.

The 10 annotated object categories in the VisDrone2019 dataset include pedestrian, people, car, bicycle, van, truck, tricycle, awning-tricycle, bus, and motor. With nearly 70% of the labeled instances measuring less than 32 × 32 pixels, the dataset is heavily biased toward small objects. Moreover, it features densely populated scenes and numerous targets that are partially or fully occluded, making it particularly suitable for evaluating methods that aim to preserve spatial details and accurately detect small objects in complex environments.

4.2. Evaluation metrics

To evaluate the accuracy of small object detection, we adopt Precision (P), Recall (R), and mean Average Precision (mAP) as primary evaluation metrics. These provide a comprehensive evaluation of both detection quality and coverage. The metrics are computed as follows:

$$P = \frac{TP}{TP+FP}, \tag{12}$$

$$R = \frac{TP}{TP+FN}, \tag{13}$$

$$AP = \int_0^1 p(r)dr, \tag{14}$$

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i. \tag{15}$$

In these equations, *TP* represents true positives that are correctly predicted, *FP* represents false positives that are incorrectly predicted as positives, and *FN* represents false negatives that are incorrectly predicted as negatives. *K* represents the number of classes.

4.3. Experimental setup

The detailed experimental configuration is summarized in Table 1. During training, all input images are resized to 640 × 640 pixels. The training process is conducted for a total of 300 epochs, with a batch size of 8. The Stochastic Gradient Descent (SGD) optimizer is employed, with an initial learning rate of 0.01, a momentum coefficient of 0.937, and a weight decay of 0.0005.

Table 1. Hardware and software configuration.

Options	Configuration
GPU	NVIDIA GeForce RTX 4090D
CPU	AMD EPYC 9754 128-Core
RAM	755Gi
Operating System	Ubuntu 22.04.1
Python Version	3.10.16
PyTorch Version	2.2.1
CUDA Version	12.1

4.4. Comparison with the baseline model

To evaluate the performance of the proposed model, Figure 5 presents the visualization experimental results of three images from the VisDrone2019 test set, which are representative, including scenes with dense and overlapping small objects and scenes from a distant perspective. For example, the first group contains numerous small objects with frequent overlaps. YOLOv11 fails to detect certain targets, such as bicycles overlapping with pedestrians. In contrast, FCA-YOLO successfully detects bicycles and pedestrians under occlusion, demonstrating enhanced robustness to overlapping instances. The second and third groups depict scenes with densely distributed targets. Moreover, while both models exhibit comparable performance in detecting medium and large objects at closer distances, FCA-YOLO achieves higher accuracy in detecting small-scale objects at greater distances.

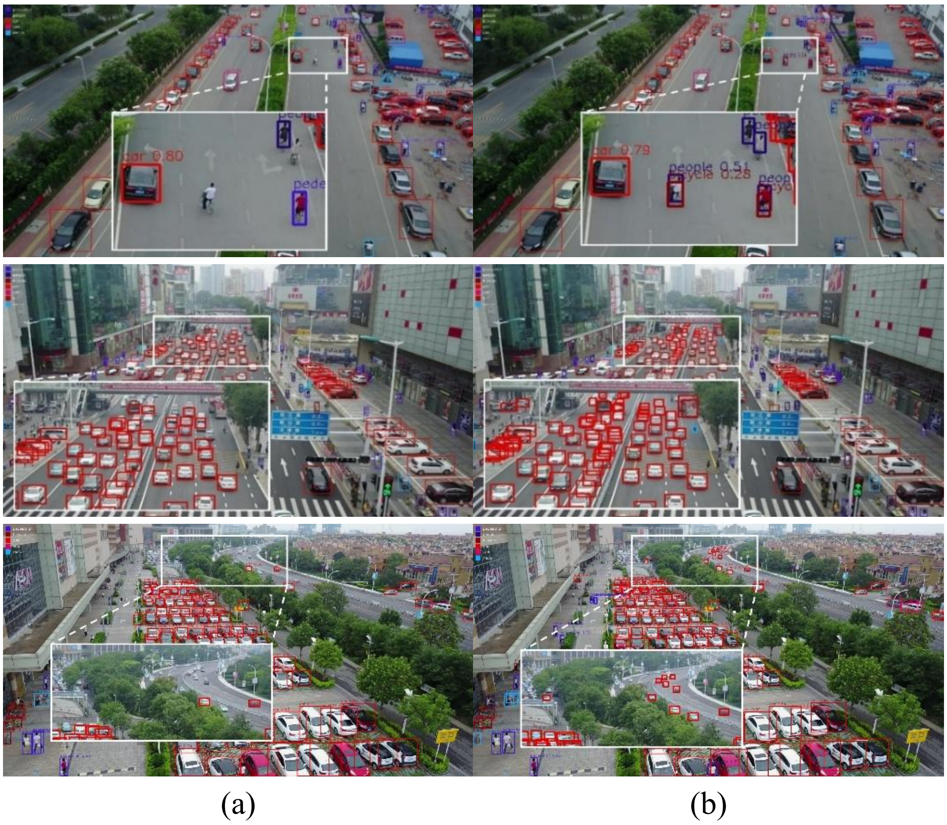

Figure 5. Visualization results on the test set: (a) YOLOv11s; and (b) FCA-YOLO.

These results underscore the effectiveness of FCA-YOLO in handling small object detection, particularly in complex and cluttered aerial scenes.

4.5. Comparison with other models

To assess the effectiveness of the proposed model, we compare FCA-YOLO with several representative models from the YOLO series first. Table 2 presents the evaluation results on the VisDrone2019 dataset, where the best and second-best results are highlighted in bold and underlined, respectively.

Table 2. Comparison with YOLO series object detection algorithms.

Model	P(%)	R(%)	mAP@0.5(%)	mAP@0.5:0.95(%)	Params(M)	Model Size (MB)
YOLOv3-tiny [31]	37.2	24.1	23.2	12.9	12.1	24
YOLOv5s [32]	42.1	32.6	30.1	17.0	9.12	17.7
YOLOv5l [32]	<u>47.4</u>	<u>36.7</u>	<u>35.0</u>	<u>20.4</u>	53.13	101.9
YOLOv6s [33]	41.2	31.7	29.6	16.9	16.3	31.32
YOLOv8s [34]	44.5	33.5	31.9	18.6	11.14	21.5
YOLOv10s [35]	44.1	32.3	30.6	17.4	<u>8.0</u>	<u>15.7</u>
YOLOv11s [17]	45.4	34.3	32.7	18.8	9.42	18.3
YOLO12s [36]	44.0	33.7	31.9	18.6	9.10	17.8
FCA-YOLO(Ours)	**48.5**	**39**	**37.7**	**21.6**	**6.59**	**13.0**

Compared with previous versions of the YOLO series, FCA-YOLO achieves improvements across all key metrics. Specifically, it achieves improvements of 5.0% in mAP@0.5 and 2.8% in mAP@0.5:0.95 compared to the baseline. Additionally, it attains the highest precision of 48.5% and recall of 39.0% among all compared models. In terms of model complexity, FCA-YOLO contains only 6.59 million parameters with a model size of 13.0 MB, representing approximately a 30% reduction compared to baseline. While YOLOv5l also achieves second-highest precision, it significantly introduces more parameters, resulting in increased model complexity. In contrast, FCA-YOLO achieves a better balance between accuracy and complexity.

To further validate the performance of FCA-YOLO in detecting small objects under complex scenes, we compare it with several representative one-stage and two-stage object detection algorithms. Table 3 presents the AP@0.5 scores for each object category, along with the overall mAP@0.5. The best results are highlighted in bold, and the second-best results are underlined.

Table 3. Accuracy comparison of detection algorithms across categories.

Model	Object category (AP@0.5 (%))										mAP@0.5
	Ped	Ppl	Bike	Car	Van	Truck	Tri	Awn-tr	Bus	Motor	(%)
Faster R-CNN [11]	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
Cascade R-CNN [37]	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2
DDOD [38]	21.9	11.9	7.4	55.3	31.4	25.4	14.5	8.6	37.2	19.8	23.3
RetinaNet [39]	13.0	7.9	1.4	45.5	19.9	11.5	6.3	4.2	17.8	11.8	13.9
VFNet [40]	20.6	9.1	6.7	55.3	32.3	25.3	14.7	8.3	39.0	19.2	23.1
MSC-CenterNet [41]	33.7	15.2	12.1	55.2	40.5	34.1	**29.2**	21.6	42.2	27.5	31.1
MSA-YOLO [42]	33.4	17.3	11.2	76.8	41.5	41.4	14.8	18.4	**60.9**	31.0	34.7
MPE-YOLO [43]	34.9	21.8	12.9	77.2	42.1	**42.8**	23.1	20.3	59.5	35.6	37
FCA-YOLO(Ours)	**36.2**	**22.3**	**13.1**	**77.7**	**43.2**	40.4	23.3	**24.4**	60.7	**36.2**	**37.7**

The experimental results indicate that FCA-YOLO delivers the highest performance in most object categories, particularly excelling in small object classes such as Pedestrian, People, Bicycle, Awning-tricycle, and Motor. This advantage is closely related to the architectural design that emphasizes early-stage preservation of high-resolution shallow features and their alignment with deep semantic representations at the detection head. By integrating the SFP module into the down-sampling pathway and introducing a high-resolution detection branch with skip connections, FCA-YOLO mitigates spatial information loss at the source and enhances the sensitivity to fine-grained local cues that are critical for densely distributed and small-scale targets. In contrast, recent YOLO-based adaptations primarily strengthen small-object representation by enriching hierarchical feature propagation through perception-enhanced convolutional blocks and attention-augmented multi-scale aggregation in the backbone and neck. While these strategies improve semantic expressiveness across scales, FCA-YOLO places greater emphasis on directly preserving and reintroducing shallow spatial details into the detection head, resulting in a structurally distinct balance between feature preservation and cross-scale fusion.

For middle-scale categories such as Car and Van, FCA-YOLO remains competitive, indicating that the cross-scale attention fusion mechanism effectively integrates shallow spatial details with deeper semantic information to support balanced multi-scale object modeling. In the Bus category, the model achieves second-highest in AP@0.5, suggesting that the high-resolution feature branch, while favoring local refinement, can maintain a degree of holistic structural representation. However, for categories with higher structural diversity and complex spatial layouts, such as Truck and Tricycle, accurate recognition relies more heavily on global context aggregation and long-range dependencies. Methods such as MSC-CenterNet, which adopt anchor-free keypoint-based regression and explicitly model object geometry at a global level, exhibit advantages in these scenarios. This comparison highlights a deliberate design trade-off in FCA-YOLO, where prioritizing shallow feature preservation and local attention enhances sensitivity for small object but may constrain the effective receptive field for large-scale or structurally irregular objects.

Overall, FCA-YOLO achieves the highest performance among all compared models with mAP@0.5 of 37.7%, demonstrating a favorable balance between fine-grained feature modeling and multi-scale representation capacity in UAV-based remote sensing detection tasks.

4.6. Ablation study and analysis

To evaluate the effectiveness of each proposed component in enhancing small object detection, we conduct a series of ablation experiments on the three key modules: the Small-Scale Key Feature Preservation Module (SFP), the Cross-scale Attention-based Fusion module (CAF), and the Fine-grained High-Resolution structure (FHR). The results are summarized in Table 4.

Table 4. Ablation study results.

SFP	CAF	FHR	P(%)	R(%)	mAP@0.5(%)	mAP@0.5:0.95(%)	Params(M)	Model Size (MB)
×	×	×	45.4	34.3	32.7	18.8	9.42	18.3
√	×	×	46.2	36.0	34.4	20.0	10.38	20.3
×	√	×	45.5	36.0	33.7	19.3	18.99	38.8
×	×	√	46.9	36.6	35.2	20.3	2.88	5.9
×	√	√	47.2	37.6	36.1	20.5	6.16	12.1
√	×	√	48.9	38.0	37.2	21.5	3.32	6.8
√	√	√	48.5	39.0	37.7	21.6	6.59	13.0

Introducing the SFP module into the baseline model leads to a 1.7% increase in mAP@0.5 with negligible additional computational overhead, confirming its effectiveness in preserving critical details of small objects. The inclusion of the CAF module further improves mAP@0.5 to 33.7%, indicating that the cross-scale attention mechanism significantly enhances feature alignment and semantic consistency across levels. The FHR module brings the most notable gains, boosting mAP@0.5 to 35.2% and mAP@0.5:0.95 to 20.3% while maintaining the lowest parameter count and model size. These results highlight the importance of high-resolution feature extraction for capturing fine-grained object structures.

We also visualize the feature responses before and after the introduction of each module, as illustrated in Figure 6. Brighter regions in the figure represent areas where the model's attention is stronger. Figure 6(b) shows the feature map of the baseline model, where the yellow boxes highlight the targets that are missed but can be successfully detected by our proposed model. In Figure 6(c)–(e), the yellow boxes indicate the targets that are detected more after introducing and integrating relevant modules compared to the baseline model. It can be observed that the addition of SFP, CAF, and FHR modules significantly improves the model's ability to detect distant targets and shows stronger responses for small-scale human targets, further validating the effectiveness and adaptability of the proposed methods in complex scenarios.

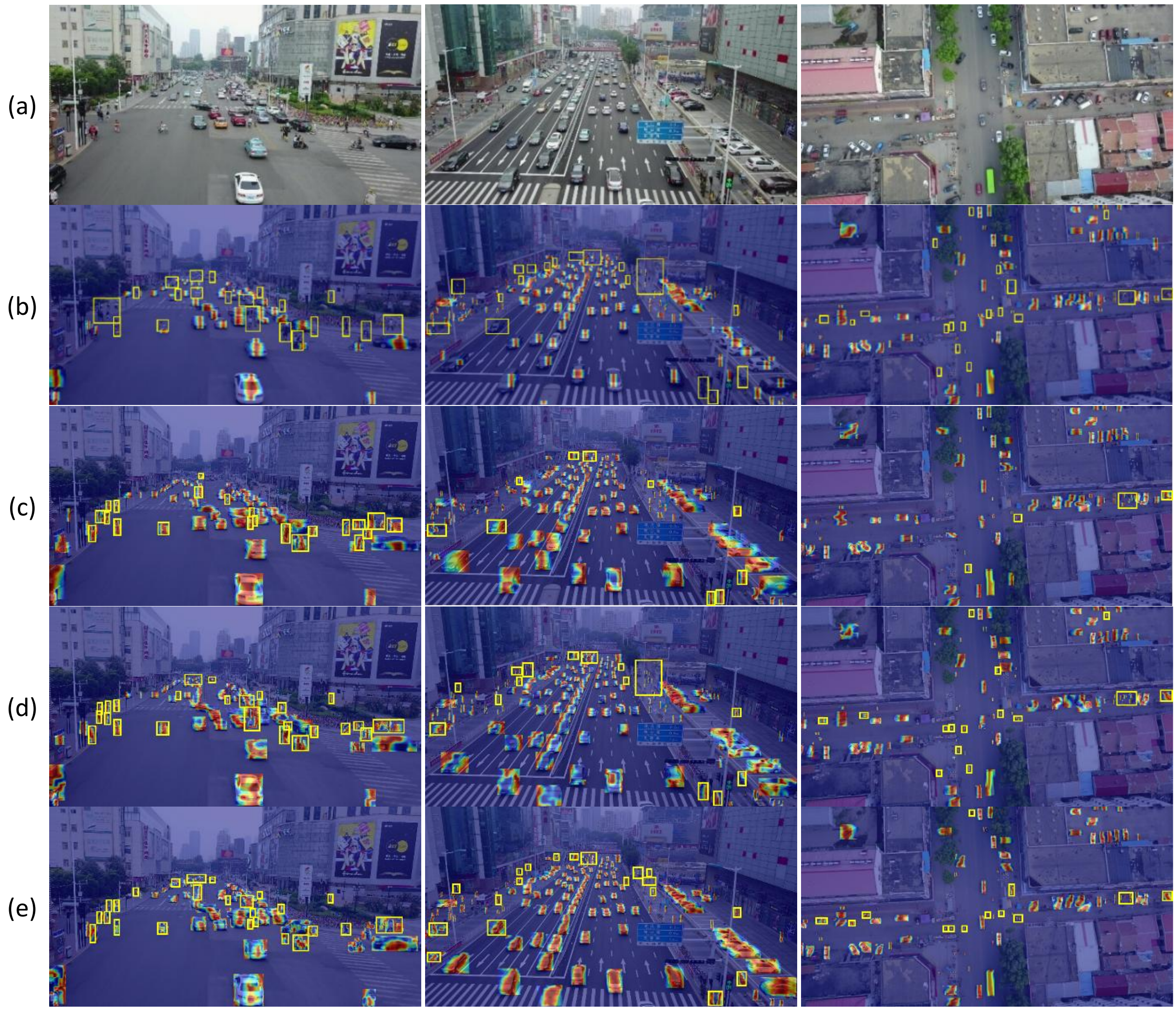

Figure 6. Visualization of heatmaps for each module: (a) Original image; (b) Original feature map of baseline; (c) After SFP; (d) After CAF; and (e) After FHR.

Furthermore, combination experiments demonstrate the synergistic benefits among the proposed modules. Considering that the SFP and CAF modules rely on the high-resolution features and multi-scale feature streams provided by FHR, respectively, we do not test the combination of SFP and CAF in isolation. Instead, we focus on analyzing their integration with FHR.

Experimental results show that CAF+FHR and SFP+FHR configurations outperform the use of FHR alone across all evaluation metrics. This validates the complementarity between detail preservation and fine-grained detection, as well as the enhancement of feature perception provided by the CAF module to FHR. When all three modules are integrated, the model achieves its best performance, confirming the effectiveness and synergy of the proposed components.

4.7. Generalization ability evaluation

To further evaluate the accuracy performance of the proposed FCA-YOLO model, we conduct experiments on two additional aerial image datasets, DOTAv1.0 [44] and VEDAI [45], and compare it with representative detectors, including RT-DETR [46], YOLOv3 [31], YOLOv5s [32], YOLOv8s [34], and YOLOv11s [17]. The results are summarized in Table 5.

Table 5. Results of the accuracy performance on DOTAv1.0 and VEDAI datasets.

Dataset	Model	P(%)	R(%)	mAP@0.5(%)
DOTAv1.0	RT-DETR	76.2	66.4	68.7
	YOLOv3	71.5	54.6	58.7
	YOLOv5s	76.0	60.1	64.5
	YOLOv8s	71.9	64.7	68.6
	YOLOv11s	77	67.4	71.8
	FCA-YOLO(Ours)	77.8	68.1	72.6
VEDAI	RT-DETR	45.5	43.6	45.5
	YOLOv3	71.6	67.8	72.9
	YOLOv5s	79.3	69.5	78.4
	YOLOv8s	77.8	72.0	78.3
	YOLOv11s	81.8	73.4	78.9
	FCA-YOLO(Ours)	84.2	76.7	79.6

On the DOTAv1.0 dataset, which consists of high-resolution images with 15 object categories and exhibits significant variations in object scale, orientation, and scene complexity, FCA-YOLO achieves an mAP@0.5 of 72.6%, surpassing YOLOv11 by 0.8%. On the VEDAI dataset, which focuses on small-vehicle detection under diverse viewpoints, illumination conditions, and complex backgrounds, FCA-YOLO reaches an mAP@0.5 of 79.6%, with precision of 84.2% and recall of 76.7%. These results indicate that FCA-YOLO consistently maintains competitive performance across datasets with different characteristics, supporting its generalization capability and robustness in UAV and remote-sensing environments.

4.8. Efficiency analysis

To further evaluate the computational efficiency and deployment feasibility of the proposed FCA-YOLO framework, we conduct a comprehensive comparison in terms of model size, computational complexity (GFLOPs), and inference speed (FPS), as reported in Table 6. Model size reflects the storage and memory requirements during practical deployment, while GFLOPs measure the theoretical computational cost of a single forward pass. FPS is reported as the average end-to-end inference speed, including preprocessing, network inference, and post-processing.

Table 6. Comparison of model complexity and inference efficiency.

Model	Params(M)	Model Size(MB)	GFLOPs	FPS
YOLOv11s [17]	9.42	18.3	21.3	257.69
YOLOv12s [36]	9.10	17.8	19.3	177.5
FCA-YOLO(Ours)	6.59	13.0	46.7	83.46

As shown in Table 6, FCA-YOLO achieves the smallest model size and the lowest number of parameters among all compared methods, indicating the best performance of its reliance on storage and memory. However, as this design emphasizes improving the detection accuracy and robustness for small and densely distributed targets, the incorporation of multi-branch feature attention fusion and high-resolution feature processing increases the overall computational complexity in terms of GFLOPs. Nevertheless, under the same hardware settings, FCA-YOLO can achieve an inference speed of 83.46 FPS, which satisfies the real-time requirements of typical UAV-based vision tasks.

5. Conclusions

In this paper, we propose FCA-YOLO, a high-precision small object detection model tailored for UAV-view remote sensing images. To mitigate the information loss caused by down-sampling, we design a small-scale feature preservation module that combines inverted bottleneck residual structures with lightweight convolutions, effectively enhancing the retention of small object features. During the feature fusion stage, we introduce a cross-scale aligned three-branch attention fusion module that integrates spatial and channel attention mechanisms; the former focuses on critical regions while the latter enhances salient features and suppresses interference, thereby improving the synergistic representation of shallow and deep features. Finally, a high-resolution detection module with skip connections further strengthens the model's sensitivity to small object details. Experimental results on the VisDrone2019, DOTAv1.0, and VEDAI datasets demonstrate that FCA-YOLO improves detection accuracy for small and dense objects, while maintaining a reduced model parameter size.

Although FCA-YOLO attains promising performance in small object detection tasks, especially in modeling fine-grained information under complex scenarios, it faces several challenges in practical UAV applications. For instance, variations in flight altitude causing viewpoint tilt, drastic changes in object scale, complex backgrounds, and frequent occlusions may affect deployment robustness. Moreover, UAV platforms are often constrained by limited computational resources and power consumption. Our primary research objective of this paper is to enhance the accuracy of small object detection. Regarding algorithm efficiency, we consider only improving the model's dependence on storage and memory. In terms of real-time performance, it needs to meet only the common application requirements, and no higher requirements have been proposed for improving runtime performance indicators such as inference speed and computational complexity. However, in the experimental section, we provide comparison results between our model and the baseline model and the latest model on these efficiency indicators. Based on this, we will include more standardized benchmarking of inference latency and computational cost under different hardware environments and input resolutions, to more comprehensively analyze the trade-off between detection accuracy and efficiency. In addition, model compression and acceleration techniques, such as pruning, quantization, and knowledge distillation, will be explored to support lightweight deployment on resource-constrained UAV platforms without sacrificing detection performance.

Author contributions

The authors confirm contribution to the paper as follows: Conceptualization, Qiming Li and Yonghui Yan; methodology, Qiming Li, Yonghui Yan, and Shaohui Lan; software, Yonghui Yan; validation, Qiming Li and Yonghui Yan; formal analysis, Qiming Li and Yonghui Yan; investigation, Qiming Li, Yonghui Yan, and Shaohui Lan; resources, Qiming Li and Yonghui Yan; data curation, Yonghui Yan; writing—original draft preparation, Qiming Li and Yonghui Yan; writing—review and editing, Qiming Li, Yonghui Yan, and Shaohui Lan; visualization, Yonghui Yan; supervision, Qiming Li; project administration, Qiming Li and Shaohui Lan; funding acquisition, Qiming Li and Shaohui Lan. All authors reviewed the results and approved the final version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the Key R&D and Transformation Projects of Xizang (Tibet) Autonomous Region Science and Technology Program (funder: the Department of Science and Technology of the Xizang (Tibet) Autonomous Region), funding (grant) number: XZ202401ZY0004.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Z. Li, Y. Wang, N. Zhang, Y. Zhang, Z. Zhao, D. Xu, et al., Deep learning-based object detection techniques for remote sensing images: a survey, *Remote Sens.*, **14** (2022), 2385. https://doi.org/10.3390/rs14102385
2. I. Attri, L. K. Awasthi, T. P. Sharma, P. Rathee, A review of deep learning techniques used in agriculture, *Ecol. Inf.*, **77** (2023), 102217. https://doi.org/10.1016/j.ecoinf.2023.102217
3. X. W. Ye, T. Jin, C. B. Yun, A review on deep learning-based structural health monitoring of civil infrastructures, *Smart Struct. Syst.*, **24** (2019), 567–585. https://doi.org/10.12989/sss.2019.24.5.567
4. A. Boukerche, Z. Hou, Object detection using deep learning methods in traffic scenarios, *ACM Comput. Surv.*, **54** (2021), 1–35. https://doi.org/10.1145/3434398
5. M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, L. Lu, SSPNet: Scale selection pyramid network for tiny person detection from UAV images, *IEEE Geosci. Remote Sens. Lett.*, **19** (2021), 1–5. https://doi.org/10.1109/LGRS.2021.3103069
6. Y. Cai, D. Du, L. Zhang, L. Wen, W. Wang, Y. Wu, et al., Guided attention network for object detection and counting on drones, *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, 709–717. https://doi.org/10.1145/3394171.3413816
7. Y. Li, Q. Li, J. Pan, Y. Zhou, H. Zhu, H. Wei, C. Liu, SOD-YOLO: Small-object-detection algorithm based on improved YOLOv8 for UAV images, *Remote Sens.*, **16** (2024), 3057. https://doi.org/10.3390/rs16163057

8. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, 886–893.
9. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. https://doi.org/10.1007/BF00994018
10. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, *IEEE Trans. Pattern Anal. Mach. Intell.*, **32** (2010), 1627–1645. https://doi.org/10.1109/TPAMI.2009.167
11. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
12. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., SSD: Single shot multibox detector, *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
13. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826
14. X. Zhou, D. Wang, P. Krähenbühl, Objects as points, *arXiv Preprint*, 2019. https://doi.org/10.48550/arXiv.1904.07850
15. M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and efficient object detection, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 10781–10790. https://doi.org/10.1109/CVPR42600.2020.01079
16. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 779–788. https://doi.org/10.1109/CVPR.2016.91
17. G. Jocher, J. Qiu, *Ultralytics YOLO11*, GitHub, San Francisco, CA, USA, 2024.
18. Z. Liu, Y. Gao, Q. Du, M. Chen, W. Lv, YOLO-extract: Improved YOLOv5 for aircraft object detection in remote sensing images, *IEEE Access*, **11** (2023), 1742–1751. https://doi.org/10.1109/ACCESS.2023.3233964
19. Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, J. Yan, FFCA-YOLO for small object detection in remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–15. https://doi.org/10.1109/TGRS.2024.3363057
20. R. Li, Y. Shen, YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO, *Signal Process.*, **208** (2023), 108962. https://doi.org/10.1016/j.sigpro.2023.108962
21. M. Wang, W. Yang, L. Wang, D. Chen, F. Wei, H. Ke, et al., FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection, *J. Vis. Commun. Image R.*, **90** (2023), 103752. https://doi.org/10.1016/j.jvcir.2023.103752
22. K. Li, Z. Lu, L. Meng, Z. Gao, YOLO-FA: Type-1 fuzzy attention based YOLO detector for vehicle detection, *Expert Syst. Appl.*, **237** (2024), 121209. https://doi.org/10.1016/j.eswa.2023.121209

23. Y. Hui, J. Wang, B. Li, DSAA-YOLO: UAV remote sensing small target recognition algorithm for YOLOv7 based on dense residual super-resolution and anchor frame adaptive regression strategy, *J. King Saud Univ.-Comput. Inf. Sci.*, **36** (2024), 101863. https://doi.org/10.1016/j.jksuci.2023.101863
24. X. Zhu, S. Lyu, X. Wang, Q. Zhao, TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, 2778–2788. https://doi.org/10.1109/iccvw54120.2021.00312
25. J. Gong, J. Zhao, F. Li, H. Zhang, Vehicle detection in thermal images with an improved YOLOv3-tiny, *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2020, 253–256. https://doi.org/10.1109/ICPICS50287.2020.9201995
26. L. Shen, B. Lang, Z. Song, CA-YOLO: Model optimization for remote sensing image object detection, *IEEE Access*, **11** (2023), 64769–64781. https://doi.org/10.1109/ACCESS.2023.3290480
27. F. Tang, J. Ding, Q. Quan, L. Wang, C. Ning, S. K. Zhou, CMUNeXt: An efficient medical image segmentation network based on large kernel and skip fusion, *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, 1–5. https://doi.org/10.1109/ISBI56570.2024.10635609
28. K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, GhostNet: More features from cheap operations, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 1580–1589. https://doi.org/10.1109/cvpr42600.2020.00165
29. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 11534–11542. https://doi.org/10.1109/cvpr42600.2020.01155
30. D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, et al., VisDrone-DET2019: The vision meets drone object detection in image challenge results, *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, 213–226. https://doi.org/10.1109/ICCVW.2019.00030
31. P. Adarsh, P. Rathi, M. Kumar, YOLOv3-Tiny: Object detection and recognition using one stage improved model, *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, 687–694. https://doi.org/10.1109/ICACCS48705.2020.9074315
32. Ultralytics, *YOLOv5: A state-of-the-art real-time object detection system*, 2021. Available from: https://docs.ultralytics.com.
33. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, et al., YOLOv6: A single-stage object detection framework for industrial applications, *arXiv Preprint*, 2022. https://doi.org/10.48550/arXiv.2209.02976
34. G. Jocher, A. Chaurasia, J. Qiu, J. Stoken, *YOLOv8: Ultralytics official implementation*, GitHub repository, 2023. Available from: https://github.com/ultralytics/ultralytics.
35. H. Chen, K. Chen, G. Ding, J. Han, Z. Lin, L. Liu, et al., YOLOv10: Real-time end-to-end object detection, *Adv. Neural Inf. Proc. Syst.*, **37** (2024), 107984–108011. https://doi.org/10.52202/079017-3429
36. Y. Tian, Q. Ye, D. Doermann, YOLOv12: Attention-centric real-time object detectors, *arXiv Preprint*, 2025. https://doi.org/10.48550/arXiv.2502.12524

37. Z. Cai, N. Vasconcelos, Cascade R-CNN: High quality object detection and instance segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 1483–1498. https://doi.org/10.1109/TPAMI.2019.2956516
38. Z. Chen, C. Yang, Q. Li, F. Zhao, Z. J. Zha, F. Wu, Disentangle your dense object detector, *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, 4939–4948. https://doi.org/10.1145/3474085.3475351
39. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, 2980–2988. https://doi.org/10.1109/ICCV.2017.324
40. H. Zhang, Y. Wang, F. Dayoub, N. Sünderhauf, VarifocalNet: An IoU-aware dense object detector, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 8514–8523. https://doi.org/10.1109/cvpr46437.2021.00841
41. Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, VisDrone-DET2021: The vision meets drone object detection challenge results, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, 2847–2854. https://doi.org/10.1109/iccvw54120.2021.00319
42. G. T. Mao, T. M. Deng, N. J. Yun, Object detection in UAV images based on multi-scale split attention, *Acta Aeronauticaet Astronaut. Sin.*, **44** (2023), 326738.
43. J. Su, Y. Qin, Z. Jia, B. Liang, MPE-YOLO: Enhanced small target detection in aerial imaging, *Sci. Rep.*, **14** (2024), 17799. https://doi.org/10.21203/rs.3.rs-3998400/v1
44. G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, et al., DOTA: A large-scale dataset for object detection in aerial images, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 3974–3983. https://doi.org/10.1109/cvpr.2018.00418
45. S. Razakarivony, F. Jurie, Vehicle detection in aerial imagery: a small target detection benchmark, *J. Vis. Commun. Image Represent.*, **34** (2016), 187–203. https://doi.org/10.1016/j.jvcir.2015.11.002
46. Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, et al., Detrs beat YOLOs on real-time object detection, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, 16965–16974. https://doi.org/10.1109/cvpr52733.2024.01605