remote sensing

MDPI

Article

MSIA-YOLO: A Multi-Scale Semantic Interaction and Alignment Network for Small Object Detection in Low-Altitude UAV Remote Sensing Images

Wen Zhang, Xiaorong Xue *, Bingyan Lu, Yishuo Tian, Jingtong Yang, Xin Zhao and Wancheng Wang

School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China; z13477988185@163.com (W.Z.); lubingyan2332@163.com (B.L.); tian761002@foxmail.com (Y.T.); 13188937083@163.com (J.Y.); 15242191552@163.com (X.Z.); yike8191@163.com (W.W.)

* Correspondence: xr_986@163.com

Highlights

What are the main findings?

- MSIA-YOLO improves small object detection in low-altitude UAV remote sensing images by integrating receptive field aware feature extraction, cascade multi-scale semantic compensation, and task interaction alignment.
- On the VisDrone2019 dataset, MSIA-YOLO achieves ($40.9 \pm 0.2\%$) mAP50 and ($24.5 \pm 0.1\%$) mAP50:95, outperforming the YOLOv11n baseline by 7.7% and 5.1%, respectively. It also achieves competitive performance on the TinyPerson and RSOD datasets, demonstrating its effectiveness and generalization ability across different small object detection scenarios.

What are the implications of the main findings?

- The results indicate that combining high resolution semantic compensation with task interaction alignment is effective for alleviating weak small object features and classification localization inconsistency in UAV remote sensing imagery.
- MSIA-YOLO provides a useful design reference for enhancing fine grained feature representation and multi-scale semantic interaction in complex UAV small object detection scenarios.

Abstract

Small object detection is fundamentally constrained by the lack of discriminative fine-grained features. Although introducing higher resolution detection scales can improve performance, it also amplifies background noise. In addition, the independently decoupled design of conventional detection heads is insufficient to address the persistent challenges of missed detections and false positives for small objects. To this end, we propose MSIA-YOLO, a YOLOv11-based detector with multi-scale semantic interaction and alignment, optimized from three complementary perspectives: feature modeling, high resolution semantic compensation, and task coordinated alignment. First, Receptive Field Attention Convolution (RFAConv) is integrated into the backbone to enhance critical local details, such as edge and texture cues, via receptive field aware attention. Second, to alleviate fine detail attenuation caused by repeated downsampling, we construct a CHSP-P2 small object detection framework with an additional P2 branch. A scale sequence fusion mechanism is further introduced to perform high resolution semantic compensation through cross scale hybrid inputs. Finally, we design a DTIA-Head (Dynamic Task Interaction and Alignment Head), which promotes joint optimization of classification and localization through dynamic task interaction and spatial alignment. Extensive experiments on the

Check for updates

Academic Editor: Mohammad Awrangjeb

Received: 12 May 2026
Revised: 16 June 2026
Accepted: 2 July 2026
Published: 5 July 2026

Copyright: © 2026 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.

public datasets VisDrone, TinyPerson, and RSOD show that, compared with the YOLOv11n baseline, MSIA-YOLO improves mAP50 by 7.7%, 10.3%, and 1.0%, respectively, while also outperforming several advanced detectors. These results demonstrate the effectiveness and generalization capability of the proposed method in small object, dense object, and complex scene object detection scenarios.

Keywords: unmanned aerial vehicle (UAV); receptive field; multi-scale feature fusion; task aligned detection head; small object detection

1. Introduction

As unmanned aerial vehicle (UAV) technology and onboard payload systems continue to advance, object detection based on low-altitude UAV remote sensing images has been widely applied in urban governance, traffic management, environmental monitoring, power line inspection, and emergency rescue [1]. However, compared with ground-based monitoring, UAVs usually operate at higher altitudes and under more variable imaging angles, causing targets such as pedestrians and vehicles to occupy only a small number of pixels in captured images. In addition, UAV remote sensing images usually cover large areas, in which objects are randomly distributed and often appear in dense layouts with local occlusion. Meanwhile, high-speed UAV motion may introduce jitter and viewpoint shifts, resulting in image blur and reduced resolution, which further increases the difficulty of small object detection.

Existing deep learning approaches to object detection generally fall into two main groups: two-stage frameworks and single-stage frameworks, reflecting two different methodological paradigms. Two-stage object detection methods, represented by R-CNN [2], Faster R-CNN [3], and Cascade R-CNN [4], follow a coarse-to-fine pipeline. They first generate candidate object regions using selective search or a region proposal network, and then refine these regions through classification and localization, which enables more sufficient feature learning and often leads to higher detection accuracy. In contrast, single-stage object detection methods regard object detection as a regression problem. Instead of generating candidate regions, they directly predict object locations and categories from the input image, such as the SSD series [5,6], the YOLO series [7–12], and the Transformer-based DETR (DEtection TRansformer) family [13–15].

Multi-scale feature fusion plays an important role in improving detection performance. It preserves high-resolution details from low-level features while introducing high-level semantic cues. This helps strengthen the representation of objects at different scales. Feature pyramid architectures provide a representative framework for multi-scale feature fusion and have therefore been widely explored in object detection. FPN [16] constructs a top-down pathway with lateral connections to progressively propagate deep semantic information to shallow high-resolution feature maps. This design alleviates the semantic gap among objects of different scales and has become a classical feature pyramid architecture for multi-scale object detection. AugFPN [17] further addresses semantic inconsistency and information loss across different pyramid levels. It improves the quality of multi-scale feature fusion through strategies such as consistency supervision and residual feature augmentation, enabling the feature maps after lateral connections to maintain more coherent semantic representations. Building on FPN, PANet [18] introduces a bottom-up path augmentation structure, which allows low-level localization information to be more effectively transmitted to high-level feature maps. This enhances information interaction among multi-level features and improves object localization performance. NAS-FPN [19]

employs neural architecture search to automatically explore connection patterns among multi-scale features. It searches a larger architectural space to identify feature pyramid topologies that are more effective than manually designed structures. BiFPN [20] further introduces bidirectional cross-scale connections and a learnable weighted feature fusion mechanism. This enables features at different resolutions to be adaptively fused according to their relative contributions, thereby improving the efficiency of multi-scale feature representation. In addition, EL-YOLO [21] proposes SCAFPN to mitigate inter-level interference during multi-scale feature fusion. It also incorporates long-range contextual modeling to strengthen context-dependent representations for small-scale objects. These studies indicate that multi-scale feature fusion remains crucial for remote sensing object detection. Further improving fusion quality and strengthening the interaction between local details and global context are therefore important directions for enhancing detection performance.

In recent years, many studies have proposed targeted improvements for object detection in complex remote sensing scenarios. Sadeghian Pour et al. [22] proposed AL-YOLO. The method strengthens feature extraction by integrating MobileViTv2 and an inverted residual design into the backbone. In addition, it introduces a C3RFE module to enlarge the effective receptive field, improving robustness in complex environments. Gao et al. [23] proposed YOLOv11_SDI for foreign object inspection in transmission lines, where hierarchical spatial-channel dynamic inference (SDI) is introduced into the detection head and combined with lightweight spatial attention to achieve adaptive multi-scale feature fusion and suppress complex background interference. Yan et al. [24] proposed LS-MambaNet, which integrates large strip convolution with a multi-granularity spatial Mamba block. This architecture efficiently models global long-range contextual information in complex remote sensing scenes with linear computational complexity. It also enhances the feature representation of high-aspect-ratio objects and objects with large-scale variations in remote sensing images. In addition to single-modal feature enhancement, multimodal information fusion has also been explored to improve detection robustness in complex scenarios. C^2Former [25] constructs a calibrated and complementary Transformer. By introducing cross-modal attention and adaptive feature sampling, it alleviates modality miscalibration and inaccurate fusion between RGB and infrared features. This improves object detection performance in complex scenes. CAGTDet [26] further focuses on the weak misalignment problem in UAV RGB-IR images. It uses a cascade alignment-guided Transformer to model translation, scale, and rotation deviations between different modalities, thereby improving the quality of cross-modal feature fusion. However, cross-modal methods mainly focus on multimodal fusion and modality alignment. They cannot directly address the challenges in single-modal low-altitude UAV remote sensing images, such as the small pixel proportion of targets and the attenuation of fine-grained details. Therefore, in single-modal UAV remote sensing images, it remains necessary to further enhance fine-grained feature representation and multi-scale detection capability for small objects.

Small object detection in low-altitude UAV remote sensing has made considerable progress in recent years. Lim et al. [27] proposed a global spatial attention mechanism for the SSD model, which improves small object detection accuracy by strengthening feature representations. However, their evaluation was mainly conducted on a single dataset, without cross-dataset comparison or generalization testing, and thus, the transferability and generalizability of the method under different scenarios and data distributions remain insufficiently validated. Gong et al. [28] proposed AED-YOLO11, which introduces AFDA frequency-domain feature enhancement to better capture complex small object features and adopts DySample dynamic upsampling to improve spatial transformation capability for small object images. Nevertheless, the stacked plugin style design provides limited structural support for extremely small objects, and the high-resolution branch

still suffers from performance bottlenecks under small-scale and complex background conditions. Xie et al. [29] proposed Drone-YOLO for small object detection in UAV imagery. By improving the feature pyramid structure, their method improves the integration and transmission of hierarchical features, thereby reinforcing the representation of small objects. However, contextual semantic modeling around small objects remains insufficient, making it difficult to fully exploit environmental cues for suppressing complex background interference. Cao et al. [30] proposed ASOD-YOLO, an aerial small object detection method based on YOLOv8n. Their method reconstructs the feature fusion network by replacing the original top-down path with a low-level information distribution structure (Low-GD) and introduces FFCBlock to extract global contextual information. However, it still does not address the fine-grained information loss caused by repeated downsampling. Hou et al. [31] proposed αS-YOLO based on YOLOv8, which introduces the C2f_GE module to enhance dependency modeling through global context and efficient channel attention and designs α-SIoU with adaptive angle control to improve localization accuracy for small objects. Yet, the overall performance gains are limited, and the robustness for extremely small and occluded objects remains inadequate. Liu et al. [32] proposed RFHA-YOLO, which introduces an RFA-Block to enhance the perception of local details in small objects through dynamic receptive field modeling. It also employs a hybrid attention mechanism to strengthen the representation of key regions and suppress interference from complex backgrounds. However, this method mainly focuses on the backbone network, with relatively limited attention to feature fusion in the neck. Therefore, there remains room for further optimization under dense small object and complex background conditions. Although these studies have improved the performance of small object detection in UAV remote sensing to some extent, several key challenges remain unresolved.

First, small objects are often densely distributed and are difficult to distinguish from background noise. Existing models usually rely on receptive fields with fixed and limited scales, making it difficult to adaptively incorporate sufficient contextual semantics for different small object types, and thus limiting robustness under large-scale variation and complex backgrounds. Second, small objects occupy very few pixels, and repeated downsampling via strided convolution or pooling can easily weaken their responses, causing the loss of edge and local details. As the network depth increases, deeper features become more semantically expressive but inevitably sacrifice spatial resolution; conversely, relying on shallow high-resolution features makes the representation more susceptible to background interference. This contradiction between semantic representation and detail preservation is particularly severe in small object detection. Third, many existing methods attempt to improve small object detection by adding an additional high-resolution detection branch. However, directly introducing a P2 output without sufficient semantic aggregation often causes the high-resolution branch to respond heavily to background regions, while remaining misaligned with medium- and high-level semantic information, thereby limiting true multi-scale complementarity. Fourth, object detection is inherently a joint optimization problem of classification and localization. Most single-stage detectors adopt a decoupled-head design with separate branches for the two tasks, which can lead to spatial task misalignment in practical scenarios and consequently increase the risks of missed detection and wrong detection.

To overcome these challenges, MSIA-YOLO is developed for small object detection in low-altitude UAV remote sensing imagery by leveraging multi-scale semantic interaction and alignment. The framework targets feature degradation and information loss and further investigates the optimization of multi-scale feature fusion for small object detection. The main contributions are summarized in three aspects. First, to handle complex background interference and substantial scale variation in UAV imagery, we introduce RFAConv into

the backbone. By assigning adaptive weights to different regions within the receptive field, the backbone can focus more effectively on informative details of small objects, thereby improving feature extraction accuracy. Second, to alleviate detail attenuation during small object feature propagation and the insufficient semantics of the high-resolution branch, we propose a CHSP-P2 architecture. This architecture designs a two-stage cascaded scale sequence fusion and hybrid input strategy in the neck. It collaboratively compensates for spatial details, structural information, and deep semantics from different feature levels. This provides the P2 detection scale with semantically enhanced high-resolution features and structurally strengthens small object detection capability. Finally, to further alleviate the insufficient interaction between classification discrimination and boundary localization in conventional decoupled detection heads, we design the DTIA-Head, a dynamic task interaction and alignment detection head. Based on the enhanced multi-scale features, DTIA-Head performs task-aware feature reconstruction and branch-adaptive alignment. This enables the classification and regression branches to more effectively exploit the semantic and spatial information required by each task, further improving spatial localization accuracy and detection capability.

Our contributions are as follows:

(1) We introduce RFAConv into the YOLOv11n backbone to enhance adaptive receptive field-based feature extraction for small UAV targets. By adaptively weighting local receptive field responses, the backbone can more effectively capture weak edge and texture cues of small objects in complex low-altitude remote sensing backgrounds.

(2) We construct a CHSP-P2 high-resolution semantic compensation framework to alleviate detail attenuation and semantic insufficiency during the repeated downsampling of small object features. Rather than simply adding a P2 detection branch or directly reusing existing scale sequence fusion modules, CHSP-P2 is built on scale sequence modeling in the neck. On this basis, we further propose a cascaded hybrid input semantic compensation strategy. This strategy performs cross-stage collaborative fusion among shallow high-resolution features from the backbone, enhanced medium-scale features from the neck, and high-level semantic features. As a result, the P2 branch can receive both fine-grained spatial details and semantically enhanced multi-scale features. This alleviates semantic insufficiency in the high-resolution branch and reduces background noise interference, thereby improving detection stability for distant and densely distributed small objects.

(3) We further design the DTIA-Head to address task conflict in the detection head. It mitigates the inconsistency between classification and localization through dynamic interactive alignment. By combining shared feature refinement with task-oriented feature distribution, the proposed head enhances the representation capability of multi-scale features and strengthens the information interaction between classification and localization branches. This design further improves small object detection accuracy and localization reliability.

(4) Experimental evaluation was conducted on the VisDrone 2019 low-altitude UAV remote sensing dataset, the TinyPerson dense small object dataset, and the benchmark remote sensing dataset RSOD. The results show that the proposed method achieves significant performance gains on VisDrone 2019 and also delivers strong performance on the other two datasets.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods, including the experimental datasets, evaluation metrics, overall architecture of MSIA-YOLO, and key design choices of the proposed method. Section 3 reports the experimental settings and provides a detailed analysis of the results, including ablation experiments, comparative experiments, and visualization analyses. Section 4 discusses the

effectiveness, generalization capability, and limitations of the proposed method. Finally, Section 5 concludes the paper and outlines future research directions.

2. Materials and Methods

2.1. Overall Architecture of MSIA-YOLO

The proposed UAV small object detection method, MSIA-YOLO, is developed based on YOLOv11 [11]. We adopt the YOLOv11n as the baseline because of its compact model size and modest computational cost, which make it well suited to resource-constrained UAV platforms while still preserving competitive detection efficiency. The overall architecture of MSIA-YOLO is shown in Figure 1. MSIA-YOLO mainly consists of three modules: (1) a backbone with dynamic receptive field modeling for hierarchical feature learning; (2) a Cascaded Hybrid Scale Sequence Pyramid with P2 branch (CHSP-P2) for feature fusion and enhancement; (3) a dynamic task interaction and alignment head (DTIA-Head) for final object prediction. The implementation details of these three improvements are presented in the following subsections.

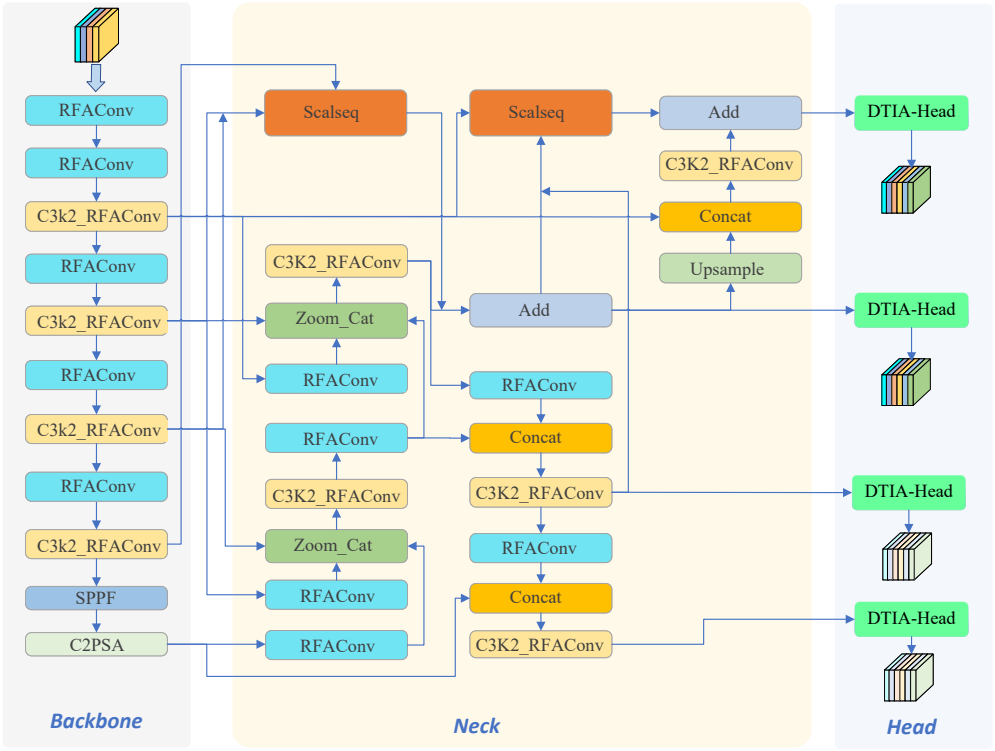


Figure 1. Overall architecture of the proposed MSIA-YOLO.

2.2. RFAConv

Improving object detection performance in low-altitude UAV remote sensing imagery requires a balance between feature representation capability and computational efficiency. Conventional convolution operations are limited by fixed receptive fields and parameter sharing, which makes it difficult to distinguish important information from less informative responses. As a result, feature extraction lacks sufficient flexibility and selectivity, and the ability of convolutional neural networks to model spatially non-uniform structures is constrained. To address these limitations, we introduce RFAConv, a receptive field attention convolution module. RFAConv dynamically modulates convolution kernels through receptive field attention and a dual-branch decoupling design, thereby enhancing the representation of local structural differences and scale variation. The overall process of RFAConv is illustrated in Figure 2.

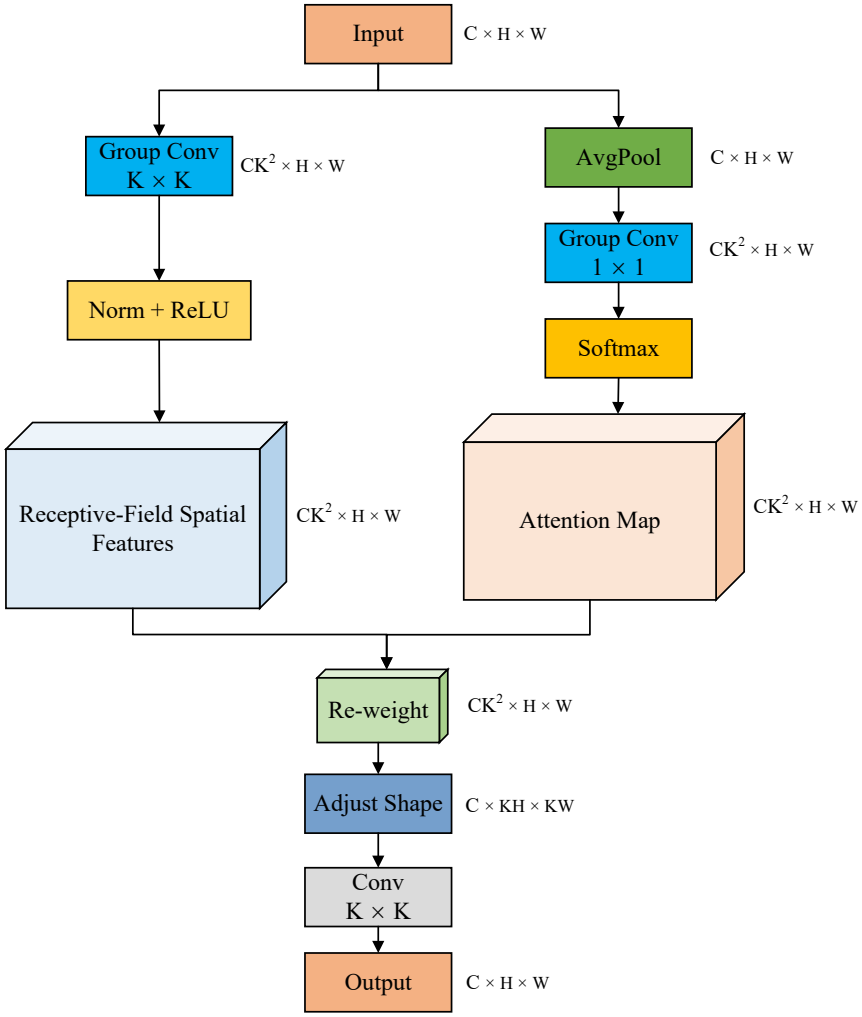


Figure 2. Receptive field attention convolution process.

The parameterization mechanism of RFAConv adopts a dual-branch architecture to decouple global importance estimation from local structural modeling. First, RFAConv efficiently extracts the receptive field spatial feature $\mathbf{F}_{rf}$ from the input feature map through grouped convolution. As shown in Figure 3, the input feature map is divided into multiple non-overlapping local blocks of size $k \times k$ using grouped convolution with a stride of k. Each local block corresponds to an independent receptive field unit, thereby mapping the original spatial feature map into the receptive field spatial domain.

Meanwhile, to maintain computational efficiency, the attention branch applies average pooling to the original input feature map for dimensionality reduction to extract aggregated semantic information from local regions while reducing redundant computation. A grouped 1×1 convolution is then used to establish correlations across different receptive field units, enabling higher-level dependency modeling through cross-receptive field interaction. After normalization, the output of this branch is fed into a Softmax function to generate the receptive field attention map $\mathbf{A}_{rf}$, where each attention vector encodes the importance distribution of features within the corresponding receptive field unit. These attention weights are subsequently applied to the receptive field spatial feature from the spatial feature branch to reweight features from different regions and enhance salient feature responses. Finally, the output is obtained by element-wise multiplication between the attention map and the receptive field feature, as formulated in Equations (1)–(3):

$$\mathbf{A}_{rf} = \text{Softmax}\left(g^{(1\times1)}(\text{AvgPool}(\mathbf{X}))\right) \quad (1)$$

$$\mathbf{F}_{rf} = \mathrm{ReLU}\left(\mathrm{Norm}\left(g^{(k\times k)}(\mathbf{X})\right)\right) \tag{2}$$

$$\mathbf{F} = \mathbf{A}_{rf} \odot \mathbf{F}_{rf} \tag{3}$$

where $g^{(i\times i)}$ represents a grouped convolution with kernel size $i \times i$, k is the receptive field size, and $\mathbf{X}$ is the input feature map. The output feature $\mathbf{F}$ is the final feature map after attention-guided fusion, obtained by multiplying the attention map $\mathbf{A}_{rf}$ with the transformed receptive field spatial feature $\mathbf{F}_{rf}$ [33]. With this design, convolution kernel parameters are no longer uniformly shared across all spatial locations. Instead, they are dynamically modulated by the attention mechanism, allowing the convolution operation to adapt to semantic and structural differences across receptive field regions, and thus enabling spatially adaptive convolution.

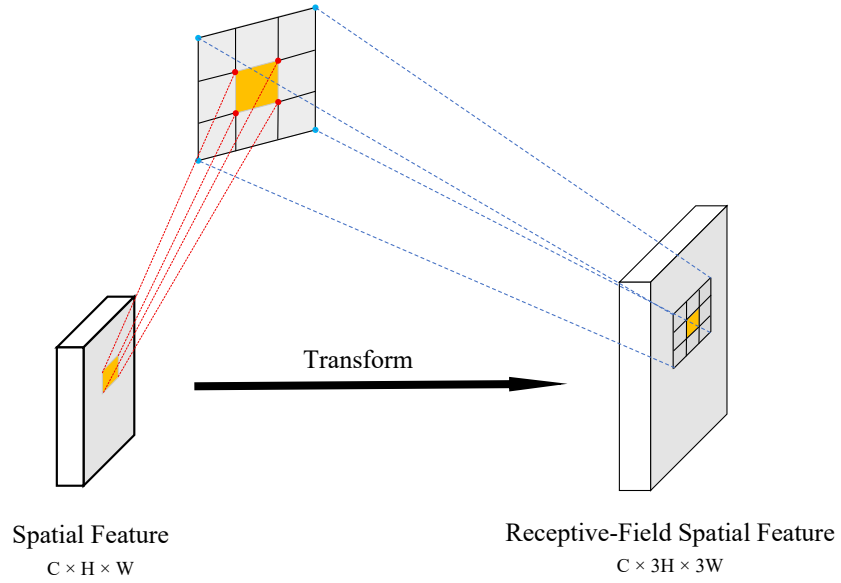


Figure 3. Receptive field spatial feature extraction process.

2.3. CHSP-P2

In low-altitude UAV remote sensing tasks, small objects in remote sensing images usually occupy only a very small number of pixels. After repeated downsampling in the backbone network, key detail representations, such as edges and textures, are prone to information attenuation. Directly adding a high-resolution P2 detection branch can improve spatial resolution to some extent. However, shallow features often contain strong redundant background responses. This significantly reduces the local contrast of small objects and makes shallow high-resolution features more susceptible to noise activation, leading to missed detections and false alarms. To suppress background interference while preserving high-resolution spatial representations, this study moves beyond the limitation of relying only on unidirectional homogeneous feature propagation. We propose a Cascaded Hybrid Scale Sequence Pyramid with P2 Branch, termed CHSP-P2. The core design of this architecture is to introduce a two-stage cascaded scale sequence fusion (ScalSeq) auxiliary branch. Through a hybrid input strategy, CHSP-P2 achieves cross-stage compensation of deep semantic information and improves the representation quality of the high-resolution P2 branch.

Specifically, on the neck side, we draw on the scale modeling idea of ScalSeq in ASF-YOLO, a medical image cell instance segmentation model, which uses 3D convolution to model inter-scale relationships [34]. We embed the scale sequence fusion module into the neck feature fusion process. Furthermore, we design a two stage cascaded scale sequence

fusion strategy with hybrid cross-stage inputs. This strategy collaboratively fuses shallow high resolution features from the backbone, enhanced medium scale structural features from the neck, and high level semantic features. It provides the P2 branch with high resolution representations enhanced by multi-scale semantic compensation and structural information. As a result, the problems of semantic insufficiency and background noise interference caused by directly adding a P2 branch are alleviated.

We retain the PAN-style main path with bidirectional top-down and bottom-up feature flow and introduce a ScalSeq-based scale-aware auxiliary branch to provide targeted enhancement for key scale features. The core of scale sequence fusion is to introduce an explicit scale dimension, organizing multi-scale features into a 4D tensor (scale, channel, height, width). The structure of the scale sequence fusion module is shown in Figure 4. Let S denote the number of feature maps input to the scale sequence fusion module. Taking the second-stage ScalSeq module used to enhance the high-resolution P2 branch in the model neck as an example, the number of input feature levels is $S = 3$. These inputs correspond to the original shallow feature $\mathbf{X}_1$ from the lower stage of the backbone at the P2 scale, the medium-scale feature $\mathbf{X}_2$ enhanced by the first-stage cascade at the enhanced P3 scale, and the strong semantic feature $\mathbf{X}_3$ from the deeper level of the conventional feature pyramid at the P4 scale. Before alignment, these three input features undergo different numbers of downsampling operations. Therefore, they exhibit substantial heterogeneity at the tensor level. For simplicity, the batch dimension is omitted in the following formulations. The initial heterogeneous feature tensors are expressed in Equation (4):

$$\left\{ \mathbf{X}_i \in \mathbb{R}^{C_i \times H_i \times W_i} \mid i \in \{1,\ 2,\ 3\} \right\} \tag{4}$$

where C_i, H_i, and W_i denote the number of channels, height, and width of the i-th input feature, respectively.

Since the subsequent 3D convolution operator requires the input tensors to be fully consistent in both the spatial and channel dimensions, the module first performs an adaptive alignment operation. The spatial resolution of the first high-resolution feature $\mathbf{X}_1$, denoted as $H \times W = H_1 \times W_1$, is used as the target size, and the target number of mapped channels is set to C. The alignment process is formally defined in Equations (5)–(7):

$$\hat{\mathbf{X}}_1 = \begin{cases} \text{Conv2D}_{1\times1}(\mathbf{X}_1), & \text{if } C \neq C_1, \\ \mathbf{X}_1, & \text{if } C = C_1, \end{cases} \in \mathbb{R}^{C\times H\times W} \tag{5}$$

$$\hat{\mathbf{X}}_2 = \mathcal{I}_{\text{nearest}}(\text{Conv2D}_{1\times1}(\mathbf{X}_2)) \in \mathbb{R}^{C\times H\times W} \tag{6}$$

$$\hat{\mathbf{X}}_3 = \mathcal{I}_{\text{nearest}}(\text{Conv2D}_{1\times1}(\mathbf{X}_3)) \in \mathbb{R}^{C\times H\times W} \tag{7}$$

where $\text{Conv2D}_{1\times1}$ denotes a standard two-dimensional convolution with a 1×1 kernel, which is used to unify the channel dimension of the features. $\mathcal{I}_{\text{nearest}}$ denotes the nearest neighbor interpolation operator, which is used to upsample low-resolution feature maps to the target spatial size.

After obtaining the standardized 2D feature streams, the module converts them into a high-dimensional feature volume through dimension expansion and sequential concatenation. Specifically, an unsqueeze operation is first applied to create a discrete scale axis between the channel and spatial axes of each feature. This axis corresponds to the scale dimension S:

$$\tilde{\mathbf{X}}_i = \text{Unsqueeze}(\hat{\mathbf{X}}_i, \text{dim} = 1) \in \mathbb{R}^{C\times1\times H\times W}, \quad i \in \{1,\ 2,\ 3\} \tag{8}$$

Subsequently, the three feature maps located at independent scale coordinates are concatenated along the scale axis in a sequential manner. This constructs a unified multi-scale 4D feature volume, denoted as $\mathbf{F}_{\text{combine}}$:

$$\mathbf{F}_{\text{combine}} = \text{Concat}\big([\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3], \text{dim} = 1\big) \in \mathbb{R}^{C\times S\times H\times W} \tag{9}$$

To perform channel mapping and feature calibration while preserving both the scale and spatial dimensions, the module introduces a 3D convolution layer with a kernel size of $1 \times 1 \times 1$. The output feature $\mathbf{F}_{\text{conv3d}}$ is expressed in Equation (10):

$$\mathbf{F}_{\text{conv3d}} = \text{Conv3D}_{(1,1,1)}(\mathbf{F}_{\text{combine}}) \in \mathbb{R}^{C\times S\times H\times W} \tag{10}$$

The unchanged shape indicates that this operation does not compress the scale dimension. Instead, it performs channel mapping and feature calibration. Subsequently, 3D batch normalization (BN3D) and nonlinear activation are applied to integrate the features, producing the interactive feature $\mathbf{F}_{\text{act}}$, as shown in Equation (11):

$$\mathbf{F}_{\text{act}} = \sigma(\text{BN3D}(\mathbf{F}_{\text{conv3d}})) \in \mathbb{R}^{C\times S\times H\times W} \tag{11}$$

where σ denotes the activation function.

Finally, to feed the high-dimensional feature volume back into the two-dimensional main path of the neck, the module adopts a cross-scale global max pooling layer, denoted as $\text{MaxPool3D}_{(S,1,1)}$. The pooling kernel spans the discrete scale axis, with a coverage size equal to the total number of scales S in the input feature volume. By operating along the scale axis, this operator selects the strongest feature responses from different scale coordinates and achieves scale reduction and dimensional compression:

$$\mathbf{F}_{\text{pool}} = \text{MaxPool3D}_{(3,1,1)}(\mathbf{F}_{\text{act}}) \in \mathbb{R}^{C\times 1\times H\times W} \tag{12}$$

$$\mathbf{Y} = \text{Squeeze}\Big(\mathbf{F}_{\text{pool}}, \text{dim} = 1\Big) \in \mathbb{R}^{C\times H\times W} \tag{13}$$

The squeeze operation is then used to remove the discrete scale axis with length 1, as shown in in Equation (13). The output feature map $\mathbf{Y}$ is restored to a standard 2D feature tensor with shape $C \times H \times W$. This completes feature interaction that integrates multi-scale high-level semantics with low-level spatial details.

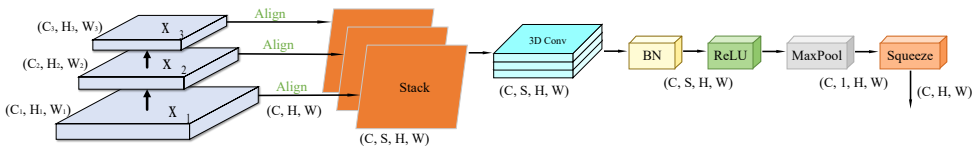


Figure 4. Structure and tensor dimension flow of the scale sequence fusion module. Here, X_1, X_2, and X_3 denote the heterogeneous input feature map from different layers. C, S, H, and W represent the aligned channels, scale coordinates, height, and width of the features, respectively, with the scale dimension instantiated as $S = 3$ in our framework.

In the CHSP neck, we adopt a two-stage cascaded enhancement strategy. In the first stage, scale sequence fusion enables the medium-resolution branch to simultaneously integrate shallow detail features from the backbone and high-level semantic features within a single fusion operation. In the second stage, which targets the high-resolution P2 branch, we further introduce a scale sequence fusion mechanism with cross-stage hybrid feedback. At this stage, the enhanced medium-resolution features are first upsampled and fused with shallow high-resolution features to form the initial representation of the high-resolution

branch. Then, the second-stage ScalSeq unit takes a set of hybrid inputs spanning different stages and semantic levels: the shallow high-resolution base feature provides fine-grained spatial and edge details; the medium-resolution feature enhanced in the first stage provides more structured object shape information; and the high-level semantic branch feature contributes global context and category discrimination cues. Through this design, the high-resolution branch achieves a better balance among background noise suppression, object boundary representation, and category discrimination.

Benefiting from the hybrid cross-stage input design in the neck, the high-resolution P2 branch receives a refined high-resolution representation that has already integrated multi-scale semantics and structural priors, rather than semantically weak shallow features. As a result, the edge and texture details of small objects can be effectively preserved. This collaborative design first performs semantic compensation in the neck and then introduces a high-resolution prediction scale in the head, allowing the resulting high-resolution features to better match the detection requirements of small objects. It also alleviates a structural contradiction in traditional small object detection. Specifically, it helps preserve fine details while maintaining sufficient semantic information.

To verify the effectiveness of the hybrid input strategy and the coupled P2 detection branch, we conduct comparative experiments under the same network framework, including three settings: a standalone P2 branch, pure backbone input (pure), and hybrid input (mixed). In the pure setting, only backbone multi-scale features are used to construct the second-stage scale sequence. In contrast, the mixed setting adopts the hybrid inputs described above, namely, shallow high-resolution features, enhanced medium-scale features, and high-level semantic branch features. The experimental results show that, while maintaining a manageable increase in model complexity and computation, the mixed strategy consistently outperforms the model with only homogeneous backbone inputs in both overall object detection accuracy and small object recall. Moreover, compared with simply adding a standalone P2 branch, introducing a CHSP-enhanced high-resolution small object detection branch significantly improves small object detection performance in UAV remote sensing scenarios with complex backgrounds and long-distance viewpoints. A detailed analysis is provided in the Experimental Section.

2.4. DTIA-Head

Although CHSP-P2 provides the detection head with high-resolution features enhanced by multi-scale semantic compensation, the detection stage still faces inconsistent requirements between classification discrimination and boundary localization. Existing YOLO-based methods generally employ a head architecture that separates classification from bounding-box regression into two distinct branches. The classification branch relies more on high-level semantic representations, whereas the localization branch emphasizes fine-grained spatial geometric information. Sharing features between these two tasks can therefore introduce conflicts, which are more likely to cause wrong detection and missed detection in complex UAV aerial scenes.

Inspired by task-alignment optimization strategies in multi-task learning [35], we design a dynamic task interaction and alignment head (DTIA-Head). This study does not simply reuse an existing task-aligned detection head. Instead, it draws on the idea of task alignment and further considers the distinct requirements of small object detection, where classification relies on semantic discrimination, whereas localization requires spatial boundary alignment. Based on the multi-scale enhanced features output by CHSP-P2, we further construct a collaborative prediction mechanism composed of shared feature refinement, task-interactive decomposition, and branch-adaptive alignment. The structure of DTIA-Head is shown in Figure 5. First, to reduce parameter redundancy caused by inde-

pendent multi-scale detection branches, this module adopts shared convolutional layers in the feature refinement stage. An interactive feature learner is designed to extract and fuse task-related representations through multiple convolutional layers, thereby forming shared interactive features. In the localization branch, these interaction features are coupled with DCNv2 (Deformable Convolution v2) [36] to predict the offsets and modulation masks required for deformable sampling, enabling the convolutional sampling locations to adaptively adjust to object shape and boundary structure. In the classification branch, the interaction features support dynamic feature selection, suppressing noise responses unrelated to category discrimination while emphasizing informative semantic features.

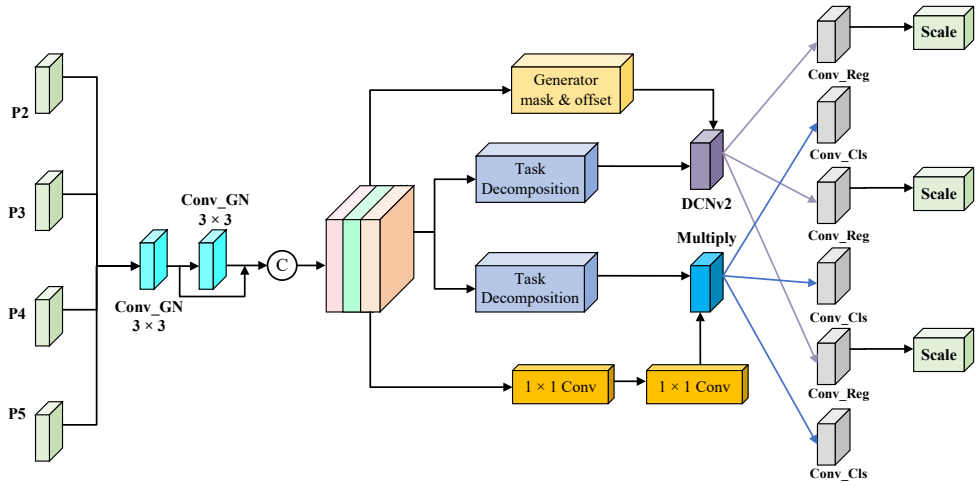


Figure 5. Structure of the DTIA-Head module.

In the detection head, because high-resolution inputs and multi-scale features are processed in parallel, the training batch size is often constrained to a relatively small value. Under this setting, the estimation of batch statistics in batch normalization (BN) can exhibit high variance, which may destabilize the optimization of the classification and localization branches. To reduce dependence on the batch dimension, we replace BN with group normalization (GN) in the shared convolution refinement stage (i.e., Conv-GN) [37]. GN divides features into several groups along the channel dimension and performs normalization by computing the mean and variance within each group. Since this process is independent of batch size, GN maintains stable normalization behavior under different batch size settings. This helps mitigate gradient fluctuations caused by unreliable batch statistics in small batch training, thereby improving convergence speed and training stability. The computation is given in Equations (14)–(16). For an input feature x_i, GN first computes the mean μ_G and standard deviation σ_G within each group, and then applies an affine transformation with learnable parameters γ and β to obtain the output y_i.

$$\mu_G = \frac{G}{C} \sum_{i=1}^{C/G} x_i \tag{14}$$

$$\sigma_G = \sqrt{\frac{G}{C} \sum_{i=1}^{C/G} (x_i - \mu_G)^2 + \epsilon} \tag{15}$$

$$y_i = \gamma \frac{x_i - \mu_G}{\sigma_G} + \beta \tag{16}$$

where C denotes the number of channels and G denotes the number of groups; μ_G and σ_G represent the mean and standard deviation within each group, respectively; γ and β are learnable scaling and shifting parameters; ϵ is a small constant used to ensure numerical

stability; y_i denotes the normalized intermediate feature used to construct the shared interactive feature.

The core decoupling and alignment mechanism of DTIA-Head does not rely on an isolated stacking of operators. Instead, it establishes a hierarchical dual-adaptive weighting mechanism composed of task-aware feature decomposition and spatial-level heterogeneous alignment.

In the first stage of the hierarchical adaptive weighting mechanism, namely channel domain dynamic feature decomposition, two independent task decomposition modules are adopted to separate the cascaded features into task-specific representations. Specifically, after the interaction features are obtained, the task decomposition modules perform controlled decomposition of shared information and dynamically reorganize the interaction features according to contextual cues, thereby generating localization features and classification features, respectively. The structure of the task decomposition module is shown in Figure 6. It first concatenates interaction features from different feature layers to form a cross layer interaction feature $\mathbf{X}^{\text{inter}}$, and then applies global average pooling to obtain its global descriptor $\mathbf{g}^{\text{inter}}$:

$$\mathbf{g}^{inter} = \text{GAP}\left(\mathbf{X}^{inter}\right) \tag{17}$$

Subsequently, the module employs two sets of fully connected mappings with independent parameters. Combined with Sigmoid activation, they generate two decoupled adaptive weight vectors, namely the classification task weight $\boldsymbol{\omega}_{\text{cls}}$ and the localization task weight $\boldsymbol{\omega}_{\text{reg}}$. These weights are used to characterize the contribution of each feature layer to the current task. The corresponding formulations are given in Equations (18) and (19):

$$\boldsymbol{\omega}_{cls} = \sigma\left(l_{c2}^{cls}\left(\delta\left(l_{c1}^{cls}\left(\mathbf{g}^{inter}\right)\right)\right)\right) \tag{18}$$

$$\boldsymbol{\omega}_{reg} = \sigma\left(l_{c2}^{reg}\left(\delta\left(l_{c1}^{reg}\left(\mathbf{g}^{inter}\right)\right)\right)\right) \tag{19}$$

where ℓ_{c1}^{cls} and ℓ_{c2}^{cls} denote the two fully connected mappings in the classification task decomposition module, while ℓ_{c1}^{reg} and ℓ_{c2}^{reg} denote the two fully connected mappings in the regression task decomposition module. δ denotes the nonlinear activation function, and σ represents the Sigmoid function.

This task specific weighting mechanism adaptively recalibrates the responses of different feature components in the shared interactive features according to the different requirements of classification and regression. The classification branch focuses more on enhancing responses with semantic discriminability, whereas the regression branch emphasizes preserving fine grained responses related to boundary structures and spatial locations. In this way, the inconsistency between classification discrimination and boundary localization is alleviated at the feature representation level.

On this basis, task-specific features can be extracted by applying the corresponding weight ω_k to the k-th layer feature component $\mathbf{X}_k^{inter}$:

$$\mathbf{X}_k^{cls} = \omega_k^{cls} \odot \mathbf{X}_k^{inter} \tag{20}$$

$$\mathbf{X}_k^{reg} = \omega_k^{reg} \odot \mathbf{X}_k^{inter} \tag{21}$$

where $k \in \{1, 2, \ldots, N\}$ denotes the different level-wise interactive features generated by the shared convolutional layers.

In the second stage of the hierarchical adaptive weighting mechanism, namely spatial-domain heterogeneous alignment, DTIA-Head further adopts different spatial weight modulation strategies for the classification and regression tasks. For the classification branch, the model predicts the spatial probability weight $\mathbf{P}_{\text{cls}}$ from the shared interactive

features and applies it to the classification feature $\mathbf{X}^{cls}$. This adaptively recalibrates feature responses, suppresses background noise and responses from non-target regions, and highlights effective semantic representations related to category discrimination. For the regression branch, the model predicts the offset and modulation mask, which are then fed into the deformable convolution unit (DCNv2). The offset is used to adjust the sampling positions, while the modulation mask recalibrates the contribution of different sampling points. In this way, the sampling positions of the convolution kernel can adapt to the object structure, enabling alignment with key boundary regions at the feature level. As a result, the classification and regression branches no longer rely on identical spatial responses. Instead, they obtain task specific representations that better satisfy the requirements of semantic discrimination and boundary localization, respectively. Through hierarchical weight modulation at both the channel level and the spatial level, DTIA-Head alleviates the inconsistency between classification confidence and localization quality in terms of feature representation and spatial response. This helps reduce false detections and missed detections of small objects in complex remote sensing backgrounds, thereby improving detection accuracy.

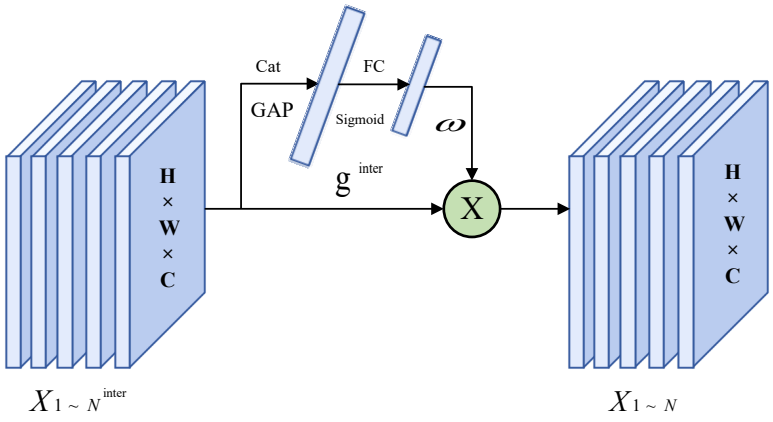


Figure 6. Architecture of the task decomposition module.

3. Results

3.1. Experimental Environments and Implementation Details

Experiments were conducted on a remote server using Python 3.10, PyTorch 2.2.2, and CUDA 11.5. Training and validation were performed on a single NVIDIA RTX 4090 GPU. The input images were preprocessed using the LetterBox strategy and resized to 640 × 640, with a batch size of 8. The model was trained for 300 epochs. Stochastic gradient descent (SGD) was adopted as the optimizer, with an initial learning rate (Lr0) of 0.01 and a final learning rate factor (Lrf) of 0.01. A cosine annealing learning rate schedule was used. To reduce training randomness, the random seed was fixed to 0 in all experiments. For data augmentation, Mosaic augmentation, HSV color space augmentation, and RandAugment were applied. No pretrained weights were used during training, and the same hyperparameter configuration was maintained for training and validation in all experiments. The detailed hyperparameter settings are listed in Table 1.

Considering that the performance differences in some key comparative experiments are relatively small, we further conduct repeated experiments with multiple random seeds for representative models to evaluate the stability of the results. Specifically, for the comparison between MSIA-YOLO and YOLOv11s on VisDrone, and the comparison between MSIA-YOLO and YOLOv11n on RSOD, we retrain the models using different random seeds and report the mean and standard deviation of mAP50.

Table 1. Training hyperparameter settings.

Parameters	Setup
Epochs	300
Batch size	8
Initial learning rate	0.01
Final learning rate factor	0.01
Optimizer	SGD
Momentum	0.937
Random seed	0
Mosaic	1.0
Input image size	640 × 640

In the data preprocessing stage, to satisfy the fixed input size requirement of the network while preserving the original aspect ratio of low-altitude UAV remote sensing images, we adopt the LetterBox strategy for proportional resizing and boundary padding. Let the spatial resolution of the original remote sensing image be $(H_{\text{org}}, W_{\text{org}})$ and the target input size be $(H_{\text{new}}, W_{\text{new}}) = (640, 640)$. The proportional scaling factor r is defined in Equation (22):

$$r = \min\left(\frac{H_{\text{new}}}{H_{\text{org}}}, \frac{W_{\text{new}}}{W_{\text{org}}}\right) \tag{22}$$

The distortion-free intermediate resolution that preserves the aspect ratio, denoted as $(H_{\text{unpad}}, W_{\text{unpad}})$, is expressed as

$$H_{\text{unpad}} = \left[H_{\text{org}} \times r\right], \quad W_{\text{unpad}} = \left[W_{\text{org}} \times r\right] \tag{23}$$

To compensate for the missing pixels required to form the square input, the total padding amounts ΔH and ΔW are defined as

$$\Delta H = H_{\text{new}} - H_{\text{unpad}}, \quad \Delta W = W_{\text{new}} - W_{\text{unpad}} \tag{24}$$

Subsequently, the padding amounts are symmetrically distributed along the top–bottom and left–right directions. A constant gray boundary with a pixel value of 114 is padded around the image. The final input image is therefore obtained with a size of 640 × 640. Meanwhile, the bounding-box annotations are synchronously updated according to the same scaling ratio and padding offsets to ensure consistency between the image and its annotations. This preprocessing strategy satisfies the fixed input size requirement of the network while preserving the original aspect ratio of the image. It also avoids geometric deformation of targets caused by direct stretching.

3.2. Datasets

The VisDrone 2019 dataset [38], released by the Machine Learning and Data Mining Laboratory of Tianjin University, is a UAV aerial image object detection benchmark designed for multi-class object detection from drone-view images. It contains 6471 training images, 548 validation images, and 3190 test images, covering diverse and complex scenarios such as urban roads, residential areas, and campuses. The dataset exhibits typical aerial imaging challenges, including significant viewpoint variation, scale variation, occlusion, and dense object distribution. VisDrone 2019 provides annotations for 10 common object categories, including typical traffic and pedestrian targets such as pedestrians, cars, bicycles, and tricycles. Small objects (with sizes smaller than 32 × 32 pixels) account for nearly 90% of the annotated instances, making this dataset a widely used benchmark dataset for evaluating the robustness and generalization of small object detection methods in complex backgrounds.

The TinyPerson dataset [39] is a public benchmark designed for ultra-small pedestrian detection under long-range imaging and large background conditions. It contains 1610 annotated images with a total of 72,651 bounding-box annotations and is commonly split into 794 training images and 816 test images. In terms of category settings, the dataset defines five labels: ocean, ground, ignore, uncertain, and dense. To follow common evaluation conventions, we treat all small pedestrians labeled as person as a single category. Owing to its extremely small object sizes, large variations in pose and aspect ratio, and strong background interference, TinyPerson is widely used to evaluate model performance in extremely small object recall and dense scene discrimination.

The RSOD dataset [40] is an optical remote sensing image object detection benchmark released by the RSIA-LIESMARS team at Wuhan University. It contains 976 remote sensing images with a total of 6950 object instances, covering four common object categories: aircraft, oiltank, overpass, and playground. The dataset exhibits clear differences in the number of samples and object density across categories, and the images contain typical remote sensing challenges, including complex ground textures, scale variation, and background interference. Accordingly, the dataset is widely utilized to examine the robustness and generalization performance of detection models under complex conditions characterized by scale diversity and dense object distribution. In this paper, the dataset is divided into training, validation, and test subsets with a 7:2:1 split.

3.3. Evaluation Metrics

Common metrics for evaluating object detection performance include Precision, Recall, mAP50, mAP50:95, GFLOPs, and the number of model parameters. The detailed definitions are as follows:

Precision measures the proportion of correctly detected objects among all predictions identified as objects (positive predictions) by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{25}$$

Recall rate refers to the proportion of all genuinely existing targets that are successfully detected by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{26}$$

Here, TP denotes the number of true positives (correctly detected bounding boxes), whereas FN denotes the number of false negatives (missed objects).

mAP50: This metric denotes the mean Average Precision of the model at an IoU threshold of 0.5.

mAP50:95: This metric denotes the mean Average Precision of the model averaged over 10 evenly spaced IoU thresholds from 0.5 to 0.95.

$$\text{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{27}$$

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\text{AP}_i = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 P_i(R)\,\mathrm{d}R \tag{28}$$

AP characterizes the detection quality for one specific class by summarizing the Precision–Recall curve, whereas mAP is obtained by averaging AP values over all categories. In general, higher AP and mAP correspond to better detection performance. The confidence score is defined on the interval from 0 to 1 and expresses the model's certainty regarding a predicted instance.

GFLOPs is commonly used to quantify model complexity, representing the total number of floating-point operations involved in one forward inference, expressed in billions.

3.4. Ablation Experiment and Analysis

3.4.1. Overall Ablation Experiments and Results

To comprehensively validate the effectiveness of the proposed improvements, we conduct ablation experiments on the VisDrone 2019 dataset using YOLOv11n as the baseline, as summarized in Table 2. The results show that the overall performance improves consistently as the proposed modules are progressively integrated. First, RFAConv enhances the representation of fine-grained small object details through dynamic receptive field attention, improving mAP50 and mAP50:95 by 1.9% and 1.2%, respectively. Next, to address the structural bottleneck of detail loss and semantic dilution caused by repeated downsampling, CHSP-P2 strengthens high-resolution representations and extends the small object detection scale. This module improves mAP50 by 2.8% and mAP50:95 by 2.0%, significantly enhancing small object detection capability and multi-scale adaptability. Meanwhile, Precision and Recall increase by 2.3% and 2.5%, respectively, indicating that CHSP-P2 effectively alleviates missed detection caused by insufficient cross-scale semantic compensation. Finally, based on the above improvements, we introduce DTIA-Head to mitigate the conflict between classification and localization and improve localization stability for dense small objects through dynamic task interaction and alignment. DTIA-Head yields consistent gains in Precision (+2.2%), Recall (+2.4%), mAP50 (+3.0%), and mAP50:95 (+1.9%).

It is worth noting that introducing DTIA-Head alone reduces the number of parameters from 2.58 M to 2.25 M, while increasing the computational cost from 6.3 to 8.1 GFLOPs. This phenomenon is mainly attributed to the different sources of parameter count and computational complexity. DTIA-Head adopts shared feature refinement, which reduces the redundant parameters caused by independent convolutional branches at multiple detection scales in the original decoupled head. As a result, the overall parameter count decreases. However, the regression branch integrates deformable convolution (DCNv2) and its spatial offset generator. These dynamic spatial operations need to be performed position by position on the feature maps of each detection scale, especially on high resolution detection layers, which introduces additional computational overhead. More precisely, DTIA-Head is a parameter-efficient task interaction detection head. The increase in mAP50:95 to 24.5% indicates that its improvement in small object localization is significant and structurally necessary. This performance gain reflects a trade off between classification localization alignment capability and computational complexity.

Table 2. Ablation experiment results.

RFAConv	CHSP-P2	DTIA-Head	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	GFLOPs	Params (M)
			33.2	19.4	42.8	34.1	**6.3**	2.58
✓			35.1	20.6	45.6	34.7	6.9	2.69
	✓		37.1	22.1	48.1	35.7	10.9	2.73
		✓	35.5	21.1	46.0	34.9	8.1	**2.25**
✓	✓		37.9	22.6	47.9	37.2	11.6	2.84
✓	✓	✓	**40.9**	**24.5**	**50.1**	**39.6**	17.4	2.64

Note: The detailed comparison between a plain P2 detection branch and the proposed CHSP-P2 is provided in Table 3. Bold values indicate the best performance.

Table 3. Ablation results of CHSP-P2 architecture modeling and input strategies.

Model	Scal Seq	P2	Input Strategy	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	FLOPs (G)	Params (M)
YOLOv11n				33.2	19.4	42.8	34.1	**6.3**	**2.584**
YOLOv11n + ScalSeq	✓			34.1	19.8	45.2	34.5	7.1	2.675
YOLOv11n + P2		✓		35.9	21.5	46.5	35.6	10.4	2.671
YOLOv11n + ScalSeq + P2	✓	✓	Pure Input	36.0	21.4	46.1	**35.9**	10.9	2.731
YOLOv11n + CHSP-P2	✓	✓	Mixed Input	**37.1**	**22.1**	**48.1**	35.7	10.9	2.729

Note: Bold values indicate the best performance.

3.4.2. Ablation Experiments of CHSP-P2 Modules

As shown in Table 3, to explore the optimal feature modeling and input strategy within CHSP-P2, we conduct systematic ablation experiments under a unified training configuration. The results show that when only ScalSeq is introduced for scale sequence modeling, the model achieves mAP50 of 34.1% and mAP50:95 of 19.8%, indicating that scale sequence modeling can improve the semantic consistency of multi-scale features to a certain extent. YOLOv11n + P2 serves as a controlled baseline, in which a standard P2 detection branch is added to YOLOv11n without introducing ScalSeq fusion or the proposed mixed input strategy. The experimental results show that when only the P2 detection branch is introduced, mAP50 and mAP50:95 increase to 35.9% and 21.5%, respectively, indicating that prediction on higher-resolution feature maps can significantly improve small object detection performance. We then investigate the collaborative effect of scale sequence modeling and the small object detection branch. On the basis of simultaneously introducing ScalSeq scale sequence modeling and the high resolution P2 detection branch, we further compare two input strategies, namely pure input and mixed input. It should be emphasized that the P2 branch in CHSP-P2 does not directly receive shallow features in isolation. Instead, it relies on ScalSeq to perform sequential modeling of multi-scale features and then constructs a high resolution semantic compensation path through different input streams. Under the pure input setting, only homogeneous multi-scale features from the backbone are used as the scale sequence inputs. The model achieves mAP50 of 36.0% and mAP50:95 of 21.4%. Compared with the standard P2 branch scheme, the gain is limited, and mAP50:95 even decreases slightly. This indicates that relying only on homogeneous backbone features cannot provide sufficient and stable semantic compensation for the high-resolution branch. In contrast, based on the coupling of ScalSeq and the P2 branch, the mixed input strategy further integrates backbone features with semantically enhanced features from the neck. This enables the P2 branch to obtain shallow high-resolution details, medium-scale structural information, and deep semantic constraints simultaneously. The experimental results show that the mixed input strategy achieves the best overall performance, with mAP50 increasing to 37.1%, mAP50:95 increasing to 22.1%, and Precision reaching 48.1%. These results indicate that the performance improvement in CHSP-P2 does not come merely from adding an extra P2 detection branch. Instead, it results from the collaborative effect among ScalSeq scale sequence modeling, high-resolution P2 prediction, and the mixed input cross-stage semantic compensation strategy.

In terms of computational cost, the pure input and mixed input schemes both have 10.9 GFLOPs, while their parameter counts are 2.731 M and 2.729 M, respectively. This is because the difference between the two input strategies mainly lies in the source and semantic composition of the input features. Both schemes perform feature fusion under the coupled framework of ScalSeq and the P2 branch. Before entering scale sequence modeling, the input features are aligned to the same channel number and spatial resolution, and the subsequent structures remain unchanged. Therefore, the main operators, feature

map sizes, and computational paths are almost identical, resulting in the same GFLOPs. This indicates that the performance gain of the mixed input scheme mainly comes from a more reasonable feature input and fusion strategy rather than additional network depth, width, or computational scale. The collaborative design of semantic compensation and high resolution prediction provides a useful technical perspective for scale feature modeling and information feedback in UAV small object detection.

3.4.3. Comparison Experiments of Different Detection Heads

To verify the effectiveness of the proposed DTIA-Head, we compare it with several detection head structures, including LQEHead, SEAMHead, and DyHead. The results are shown in Table 4. The baseline model adopts the original YOLOv11n detection head, with mAP50 and mAP50:95 of 33.2% and 19.4%, respectively. After introducing LQEHead, the overall improvement is limited. SEAMHead does not bring a stable performance gain on this dataset. DyHead achieves higher detection accuracy through dynamic feature interaction, with mAP50 and mAP50:95 reaching 34.6% and 20.3%, respectively, but it also introduces a higher parameter count. In contrast, the proposed DTIA-Head achieves the best overall performance, with mAP50 and mAP50:95 reaching 35.5% and 21.1%, respectively. Compared with the baseline model, these results correspond to improvements of 2.3% and 1.7%. Meanwhile, DTIA-Head reduces the parameter count from 2.58 M to 2.25 M, indicating that it achieves good parameter efficiency while improving detection accuracy. It therefore provides a more effective balance between detection performance and efficiency.

Table 4. Comparison results of different detection heads.

Methods	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	FLOPs (G)	Params (M)
Baseline	33.2	19.4	42.8	34.1	6.3	2.58
LQEHead	33.6	19.5	44.2	33.3	6.5	2.59
SEAMHead	32.8	19.0	43.4	33.2	**5.8**	2.49
DyHead	34.6	20.3	45.6	33.5	7.6	3.11
DTIA-Head	**35.5**	**21.1**	**46.0**	**34.9**	8.1	**2.25**

Note: Bold values indicate the best performance.

3.5. Comparison Experiment

3.5.1. Comparative Experiment on VisDrone 2019 Dataset

(1) **Overall Comparison Results:** To validate the superiority of the proposed method, we compare MSIA-YOLO with current state-of-the-art models on the VisDrone 2019 dataset. The results are reported in Table 5.

Compared with several high-performing mainstream object detection methods listed in the table, MSIA-YOLO achieves a more competitive overall performance on VisDrone 2019, with Precision and Recall reaching 50.1% and 39.6%, respectively, and mAP50 and mAP50:95 reaching 40.9% and 24.5%, respectively. Compared with the baseline YOLOv11n, MSIA-YOLO improves mAP50 by 7.7% and mAP50:95 by 5.1%, while Precision and Recall improve by 7.3% and 5.5%, respectively. Although the computational cost of MSIA-YOLO increases from 6.3 G to 17.4 G and the FPS decreases from 174.3 to 103.6, the model still maintains a high real-time inference speed. Compared with the classical detectors SSD and Faster R-CNN, MSIA-YOLO improves mAP50 by 17.0% and 7.7%, respectively, and improves mAP50:95 by 14.3% and 7.5%, respectively. Compared with the computationally expensive RT-DETR model, the proposed method trails by 6.3% in mAP50 and 4.5% in mAP50:95. However, its parameter count is substantially reduced from 32.8 M to 2.64 M, making it more suitable for UAV application scenarios. Compared with AED-YOLO11 and

Drone-YOLO, which are designed for UAV aerial image detection, MSIA-YOLO improves mAP50 by 2.4% and 1.8%, respectively, while using only 2.64 M parameters.Compared with recent remote sensing small object detection methods, including LYA-YOLO-N, OSD-YOLOv10, SSCW-YOLO, and SRTSOD-YOLO-n, MSIA-YOLO improves mAP50 by 1.1%, 7.5%, 3.1%, and 4.6%, respectively. It also improves mAP50:95 by 0.8%, 5.4%, 2.2%, and 2.7%, respectively. Among these methods, LYA-YOLO-N achieves a higher FPS and a lower parameter count, but both its mAP50 and mAP50:95 are lower than those of MSIA-YOLO. SSCW-YOLO has a parameter count close to that of the proposed method, whereas MSIA-YOLO still outperforms it by 3.1% in mAP50 and 2.2% in mAP50:95. In addition, TPH-YOLOv5, as a representative method for UAV scene detection, achieves an mAP50 of 41.7%, which is slightly higher than the 40.9% obtained by MSIA-YOLO. However, TPH-YOLOv5 is based on the larger YOLOv5l framework, with 109 G FLOPs and 45.4 M parameters, and its FPS is only 33.1. In contrast, MSIA-YOLO requires only 17.4 G FLOPs and 2.64 M parameters, while achieving an FPS of 103.6. Moreover, its mAP50:95 reaches 24.5%, which is higher than the 24.1% of TPH-YOLOv5.Compared with a range of YOLO-series models, the proposed method shows more prominent performance in both mAP50 and mAP50:95. It exceeds the latest YOLOv13n by 9.8% and 6.6%, respectively. Although the FPS of MSIA-YOLO is 103.6, which is lower than that of some n-series models, it still satisfies the requirement for real-time inference. More importantly, it substantially improves small object detection accuracy in complex UAV scenarios, demonstrating a favorable accuracy efficiency trade-off.

Table 5. Comparison experiments on the VisDrone2019 dataset.

Model	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	FLOPs (G)	Params (M)	FPS
SSD [6]	23.9	10.2	21.0	35.5	87.9	24.5	–
Faster R-CNN [3]	33.2	17.0	45.3	33.8	206.7	41.2	–
YOLOv5n	32.9	19.1	44.5	33.2	7.2	2.51	171.5
YOLOv8n	33.0	19.1	43.9	33.0	6.8	2.69	168.2
YOLOv9t	33.9	19.7	45.5	33.2	7.6	1.97	113
YOLOv10n	33.4	19.6	44.6	32.8	6.5	2.27	220.6
YOLOv11n	33.2	19.4	42.8	34.1	6.3	2.58	174.3
YOLOv11s	39.8 ± 0.1	24.0 ± 0.2	51.7	39.0	21.3	9.42	115.8
YOLOv12n	33.2	19.4	44.3	33.0	6.3	2.56	151.2
YOLOv13n	31.1	17.9	40.1	31.7	6.2	2.45	189
RT-DETR-l	47.2	29.0	62.1	45.1	108	32.8	35.7
AED-YOLO11 [28]	38.5	22.7	48.8	37.5	10.8	1.9	52.44
Drone-YOLO [29]	39.1	24.0	–	–	39.4	8.3	60
LYA-YOLO-N [41]	39.8	23.7	–	–	8.1	0.99	195
TPH-YOLOv5 [42]	41.7	24.1	52.5	40.5	109	45.4	33.1
OSD-YOLOv10 [43]	33.4	19.1	43.9	32.5	7.9	1.6	136
SSCW-YOLO [44]	37.8	22.3	46.9	38.5	8.7	2.73	115.1
SRTSOD-YOLO-n [45]	36.3	21.8	–	–	7.4	3.5	147
MSIA-YOLO	40.9 ± 0.2	24.5 ± 0.1	50.1	39.6	17.4	2.64	103.6

Note: Results marked with "±" are reported as mean ± standard deviation over three independent runs with different random seeds.

It should be noted that MSIA-YOLO achieves significant accuracy gains at the cost of increased computational complexity. However, the complexity increase in the proposed method does not result from unbounded structural stacking. To further verify its effectiveness under comparable complexity conditions, MSIA-YOLO is separately compared with YOLO s-series models in the following analysis. This enables a fairer evaluation of the trade-off among accuracy, parameter count, and computational complexity. Overall, these results clearly demonstrate that MSIA-YOLO has stronger adaptability and robustness

in complex backgrounds and UAV small object detection scenarios. While maintaining relatively controllable model complexity and real-time inference speed, it achieves better localization accuracy and overall detection performance.

(2) **Comparison with Similar Complexity Models:** To further verify that the performance improvement in MSIA-YOLO does not simply result from increased computational complexity, we compare MSIA-YOLO with YOLO series models of comparable complexity on the VisDrone2019 dataset. The results are shown in Table 6. As can be observed, YOLOv8s, YOLOv11s, YOLOv12s, and YOLOv13s all have higher FLOPs and parameter counts than MSIA-YOLO, which requires only 17.4 GFLOPs and 2.64 M parameters. In terms of detection accuracy, MSIA-YOLO achieves an mAP50 of ($40.9 \pm 0.2\%$) and an mAP50:95 of ($24.5 \pm 0.1\%$), both of which are higher than those of the comparable complexity models. These results indicate that MSIA-YOLO achieves better detection performance than the *s*-series models while using fewer computational resources and parameters. This demonstrates its advantage in balancing accuracy improvement and controllable computational cost. Therefore, the performance gain of MSIA-YOLO mainly comes from the effective design of multi-scale semantic compensation and task interaction alignment structures, rather than merely from increased computational complexity.

Table 6. Comparison with similar complexity YOLO models on the VisDrone2019 dataset.

Model	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	FLOPs (G)	Params (M)
YOLOv8s	39.4 ± 0.3	23.6 ± 0.1	50.7	38.1	28.5	11.1
YOLOv11s	39.8 ± 0.1	24.0 ± 0.2	**51.7**	39.0	21.3	9.42
YOLOv12s	39.6	23.9	50.2	38.8	21.2	9.23
YOLOv13s	38.3	22.7	48.8	37.7	20.7	9.0
MSIA-YOLO	**40.9 ± 0.2**	**24.5 ± 0.1**	50.1	**39.6**	**17.4**	**2.64**

Note: Results marked with "±" are reported as mean ± standard deviation over three independent runs with different random seeds. Bold values indicate the best performance.

3.5.2. Comparative Experiment on TinyPerson Dataset

To validate the effectiveness of the proposed method in extremely small object scenarios, we further compare MSIA-YOLO with multiple mainstream object detection methods on the TinyPerson dataset. The results are shown in Table 7.

Compared with the baseline YOLOv11n, MSIA-YOLO achieves a stronger overall performance on TinyPerson, with Precision improved by 5.5%, Recall improved by 9.6%, and mAP50 and mAP50:95 improved by 10.3% and 3.64%, respectively. These results indicate that the proposed method can effectively alleviate missed detection of extremely small objects caused by weak visual details and severe background interference, thereby improving both detection accuracy and stability. Furthermore, compared with multiple YOLO s-series models, MSIA-YOLO maintains a smaller parameter count and lower computational complexity while achieving 43.8% Precision, 33.1% Recall, 30.4% mAP50, and 10.1% mAP50:95. This demonstrates that the proposed method achieves higher detection accuracy while keeping the model scale controllable. Compared with the latest YOLOv13 model, MSIA-YOLO achieves a better balance between detection accuracy and computational efficiency. These comparative results demonstrate that MSIA-YOLO provides stronger detection capability and more stable overall performance in extremely small and densely distributed object scenarios such as TinyPerson, further validating its superiority in small object detection tasks.

Table 7. Comparison results on the TinyPerson dataset.

Model	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	GFLOPs	Params (M)
YOLOv5n	22.1	7.89	38.0	23.2	7.2	2.51
YOLOv6n	19.9	6.96	37.4	21.3	11.7	4.23
YOLOv8n	23.3	8.41	**44.3**	23.3	8.1	3.01
YOLOv9t	20.8	7.43	38.2	23.1	7.6	**1.97**
YOLOv10n	20.0	7.26	36.8	22.0	6.5	2.26
YOLOv11n	20.1	6.46	38.3	23.5	6.3	2.58
YOLOv11s	23.3	7.54	42.3	24.6	21.3	9.41
YOLOv12n	20.6	6.85	36.2	23.4	6.3	2.56
YOLOv12s	24.4	8.93	42.2	25.4	21.2	9.23
YOLOv13n	20.3	7.19	36.4	22.5	**6.2**	2.45
YOLOv13s	21.3	7.84	41.2	22.3	20.7	9.0
MSIA-YOLO	**30.4**	**10.1**	43.8	**33.1**	17.4	2.64

Note: Bold values indicate the best performance.

3.5.3. Comparative Experiment on RSOD Dataset

To validate the generalization capability of the proposed method on conventional remote sensing datasets, we further compare MSIA-YOLO with mainstream YOLO-based object detection methods on the RSOD dataset. The results are shown in Table 8.

Table 8. Comparison results on the RSOD dataset.

Model	mAP50 (%)	mAP50:95 (%)	P (%)	R (%)	GFLOPs	Params (M)
YOLOv5n	89.5	60.6	93.5	81.6	7.2	2.51
YOLOv6n	92.2	61.8	90.2	88.2	11.8	4.24
YOLOv8n	93.9	61.5	93.8	87.5	8.2	3.01
YOLOv10n	85.9	58.9	78.4	85.6	6.5	**2.27**
YOLOv11n	93.3 ± 0.1	61.6 ± 0.1	90.0	89.0	6.3	2.58
YOLOv11s	92.0	**64.5**	**94.1**	85.0	21.3	9.41
YOLOv12n	92.0	60.1	92.7	85.6	6.3	2.56
YOLOv13n	**98.3**	58.1	92.7	84.4	**6.2**	2.45
MSIA-YOLO	94.5 ± 0.2	63.4 ± 0.3	92.5	**93.1**	17.4	2.64

Note: Results marked with "±" are reported as mean ± standard deviation over three independent runs with different random seeds. Bold values indicate the best performance.

Compared with the baseline model, MSIA-YOLO improves Precision by 2.5% and Recall by 4.1%, while mAP50 and mAP50:95 increase by 1.0% and 2.2%, respectively. These results indicate that the proposed method has stronger fine-grained localization capability for small objects and better multi-scale generalization in remote sensing scenes. Further comparison shows that MSIA-YOLO has a clear advantage in mAP50:95 over most lightweight YOLO n-series models. Compared with YOLOv11s, whose mAP50:95 reaches 64.5% and is slightly higher than the 63.7% of MSIA-YOLO, YOLOv11s requires substantially more parameters and computation. This suggests that MSIA-YOLO achieves more favorable small object detection performance with a more controllable model scale. Overall, the RSOD comparison results demonstrate that MSIA-YOLO is effective not only in UAV scenarios but also in traditional remote sensing small object detection tasks, showing strong generalization capability and competitive performance.

3.6. Detection Visualization Results

To further validate the effectiveness and generalization capability of MSIA-YOLO in UAV remote sensing object detection tasks, we conduct visualization experiments on three representative datasets: VisDrone 2019, TinyPerson, and RSOD. Several representative

detection results are presented for qualitative comparison. In the visualization results, yellow circles denote missed detection, and red circles denote wrong detection. The first column shows the original images, the second column shows the prediction results of YOLOv11n, and the third column shows the prediction results of MSIA-YOLO.

3.6.1. Visualization of VisDrone 2019 Detection Results

As shown in Figure 7, MSIA-YOLO demonstrates strong performance on the VisDrone 2019 dataset, with more stable detection results and stronger background suppression capability. In the distant small object scenario in (a), the small scale of distant pedestrians causes multiple cases of missed detection for the baseline model, whereas MSIA-YOLO successfully detects small objects with average sizes below 20×20 pixels. Group (b) presents a challenging case involving extremely small objects. The detection results of YOLOv11n show missed detections of several extremely small targets in the upper-left and top regions. In contrast, MSIA-YOLO effectively addresses these omissions. Moreover, for a heavily occluded car in the upper-right region of this scene, the proposed model also achieves successful detection. In the dim and dense scene in (c), YOLOv11n exhibits obvious missed detection for densely distributed pedestrians and fails to recognize electric vehicles under low illumination. In contrast, MSIA-YOLO shows significantly improved small object detection capability in this scenario, with a substantial reduction in yellow missed detection markers. It also correctly detects overlapping pedestrians and electric vehicles, demonstrating stronger robustness. Similarly, in the dense pedestrian scene shown in group (d), the detection results of MSIA-YOLO show that the tiny and densely distributed pedestrians in the farthest region are detected with more complete and denser bounding boxes. Their confidence scores are also generally improved. In dense regions, higher confidence indicates that the model extracts target features more reliably. This greatly reduces the risk that true targets are filtered out due to low confidence, thereby alleviating missed detections. In the occluded and complex scene shown in group (e), MSIA-YOLO demonstrates better discrimination between occluded targets and background interference regions. Its ability to detect distant and occluded small objects is substantially improved. The comparison between the two models in the occlusion scene shown in group (f) further indicates that YOLOv11n is prone to false detections when targets are severely occluded by slender objects, such as railings, traffic lights, and fences. In contrast, MSIA-YOLO shows stronger occlusion-robust discrimination, allowing occluded targets to be detected with higher confidence.

3.6.2. Visualization of TinyPerson Detection Results

In Figure 8, scene (a) corresponds to a case with highly complex background textures. YOLOv11n shows multiple missed detection cases (yellow markers), indicating insufficient response stability for densely distributed small objects. In contrast, MSIA-YOLO significantly reduces missed detection and also suppresses wrong detection in background interference regions. In the distant and sparse small object scene in (b), MSIA-YOLO successfully detects a large number of small objects with sizes below 16 pixels, producing more complete detection results. This indicates improved sensitivity and stability for very small objects.For the low contrast sea surface scene in (c), MSIA-YOLO also performs better. YOLOv11n is easily affected by wave patterns and specular reflections and fails to detect these small objects with a low signal-to-noise ratio, whereas MSIA-YOLO detects more object instances and reduces missed detection regions. This demonstrates that MSIA-YOLO has stronger capability in enhancing weak small object features and discriminating low contrast targets from background interference.

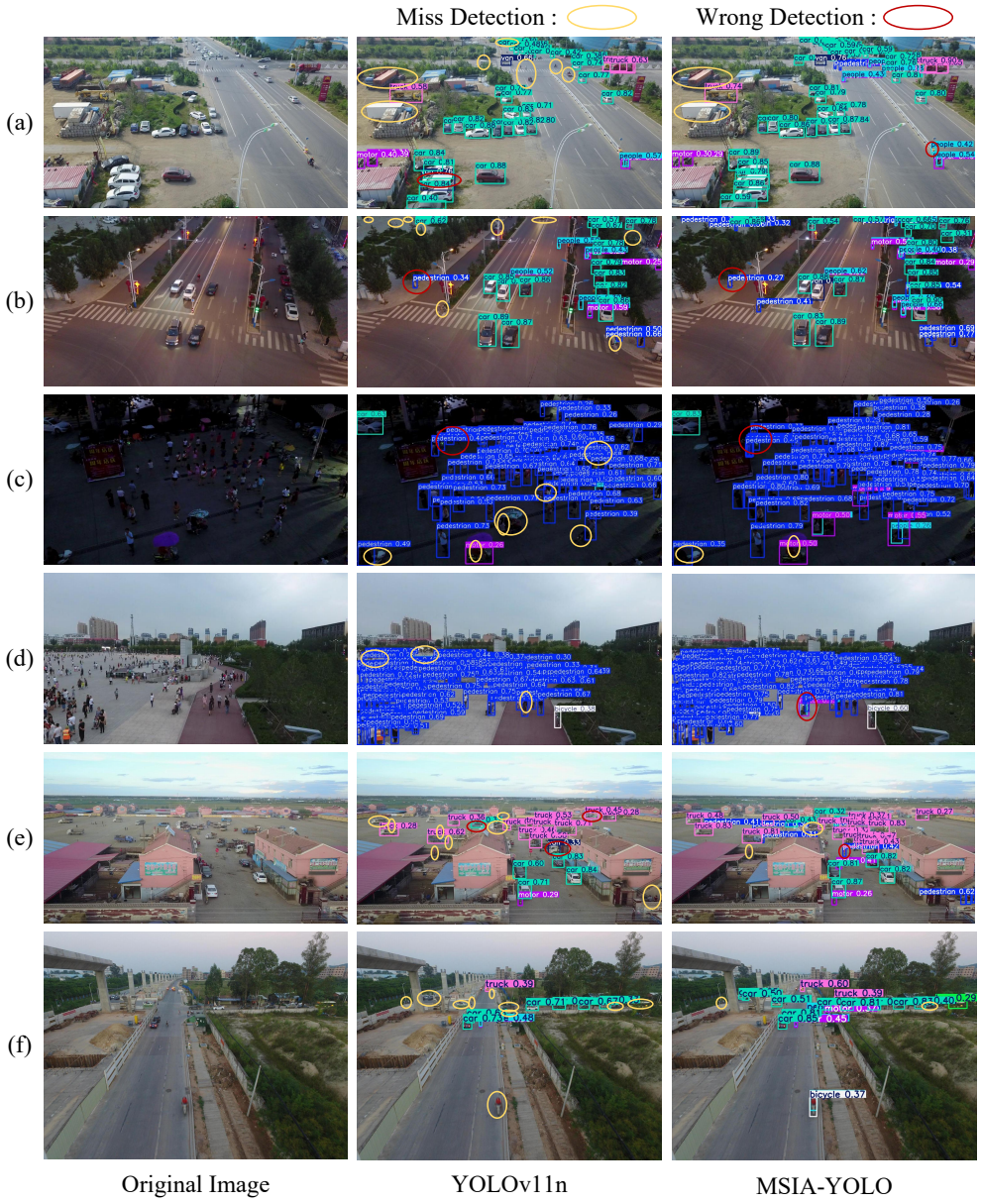


Figure 7. Representative visualization results of MSIA-YOLO on the VisDrone 2019 dataset. In each group, the first image shows the original image, the second image shows the detection result of YOLOv11n, and the third image shows the detection result of MSIA-YOLO. Yellow markers indicate missed detections, and red markers indicate false detections. (**a**) Distant small object scene; (**b**) extremely small object scene; (**c**) dim and dense scene; (**d**) dense pedestrian scene; (**e**) occlusion and complex scene; (**f**) occlusion scene.

3.6.3. Visualization of RSOD Detection Results

Figure 9 shows the detection results of MSIA-YOLO on the RSOD dataset. As a relatively standardized remote sensing dataset with clear object semantics, RSOD usually allows mainstream detectors to achieve relatively high overall object detection accuracy. However, wrong detection and missed detection of small objects still occur. As shown in (a) and (b), MSIA-YOLO provides more complete coverage of object instances under the same conditions, with fewer wrong detection and missed detection cases. In the structurally complex overpass scene in (c), both models can detect the target. However, the prediction boxes of YOLOv11n are generally larger and include more irrelevant regions. In contrast, MSIA-YOLO produces higher confidence predictions with tighter and more convergent bounding boxes, and the predicted boundaries align more closely with the

main object structure and the ground truth (GT) boxes. This indicates that the proposed method achieves better localization quality, further improving the reliability and usability of detection results.

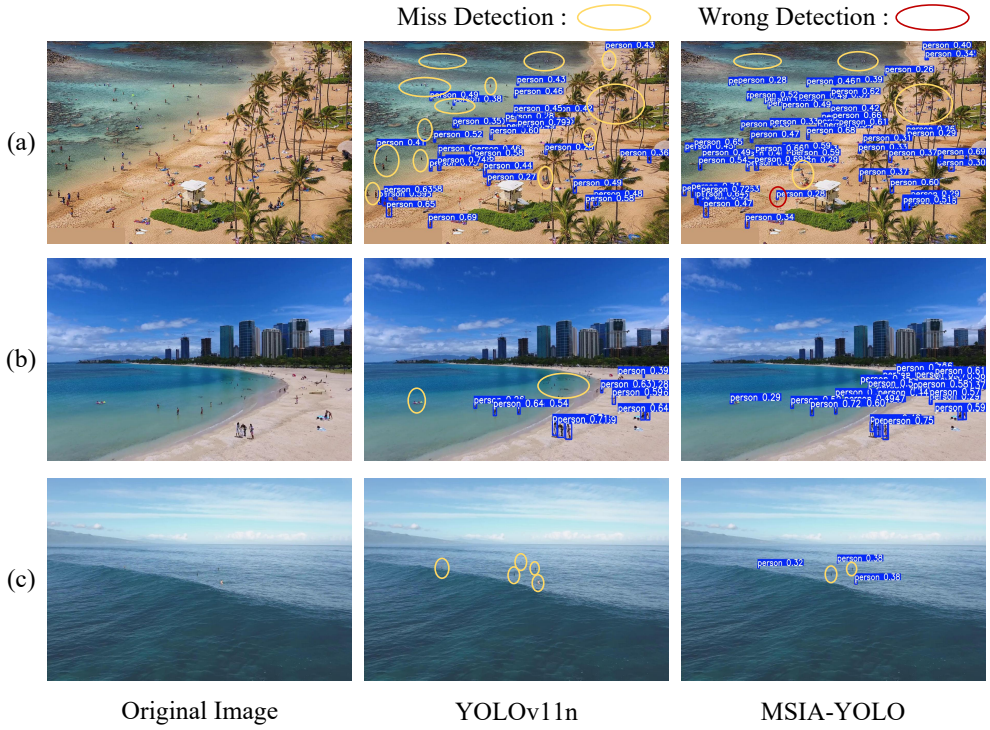


Figure 8. Representative visualization results of MSIA-YOLO on the TinyPerson dataset. (**a**) Highly complex background texture scene; (**b**) distant sparse small object scene; (**c**) low contrast sea surface scene.

The comprehensive visualization comparisons show that MSIA-YOLO has clear advantages in three aspects. First, it provides stable detection performance for distant small objects and low saliency small objects. Second, it shows stronger adaptability to occlusion and complex scene interference, maintaining more consistent responses to small objects. Third, it has better discrimination capability for densely distributed objects and can effectively suppress wrong detection and missed detection. The results on these three datasets collectively demonstrate the practical value and generalization potential of the proposed method for UAV remote sensing small object detection tasks.

3.6.4. Analysis of Failure Cases on the VisDrone Dataset

Although MSIA-YOLO achieves favorable performance in small object detection, it still has certain limitations under extremely complex imaging conditions. Figure 10 shows typical failure cases of MSIA-YOLO under nighttime low-light conditions, strong shadows, and long distance motion blur. The yellow ellipses indicate missed detection regions, while the red ellipses indicate false detection regions. As shown in (a) and (b), under nighttime low-light conditions, the overall signal-to-noise ratio of the image decreases. Vehicle shadows, road edges, and local occlusions further weaken the discriminative differences between small objects and the background. This leads to missed detections of distant small objects, missed detections of nearby occluded targets, or confusion between object categories. (c) presents a failure case under motion blur. Because the object contours become blurred and diffused, the distinguishable features of small objects, such as distant pedestrians and vehicles, are further weakened, resulting in missed detections and false detections.

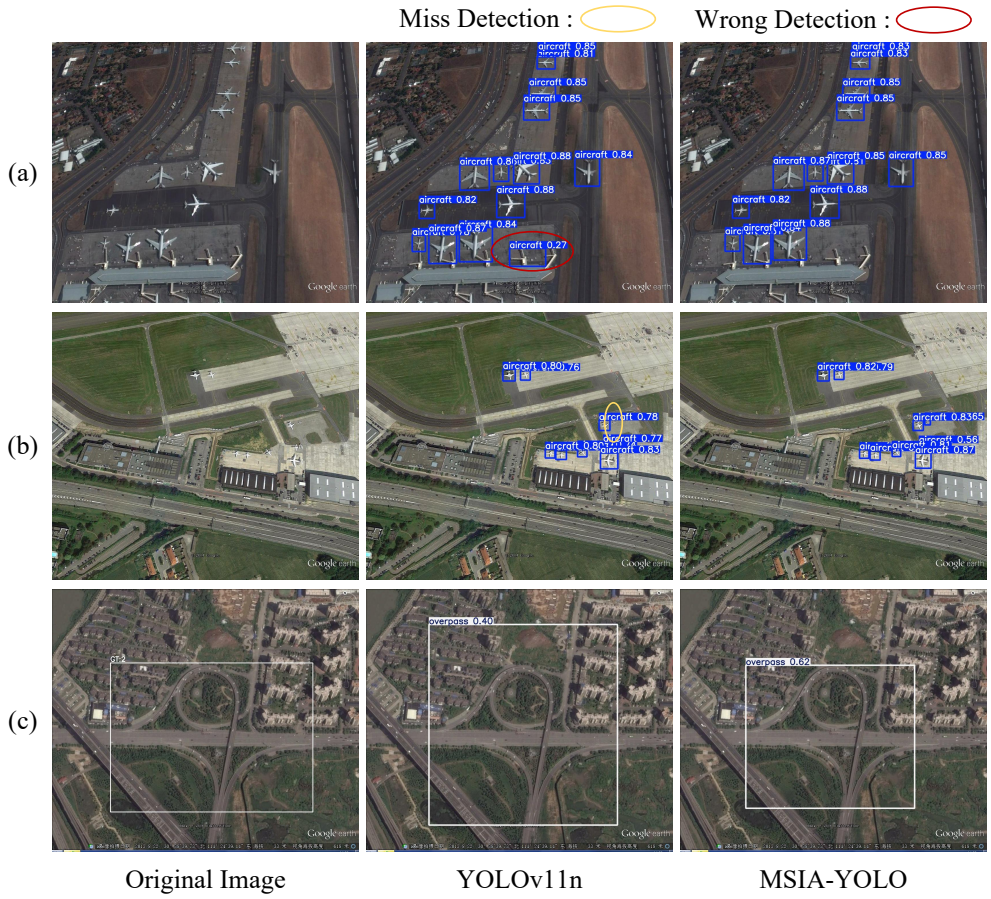


Figure 9. Representative visualization results of MSIA-YOLO on the RSOD dataset. (**a**) Complex background scene; (**b**) dense small object scene; (**c**) structurally complex overpass scene.

These results indicate that the model still has room for improvement in representing weak-texture small objects under degraded conditions, such as extremely low illumination, severe occlusion, strong reflection, and motion blur. Future work may consider introducing low-light image enhancement, deblurring-based data augmentation, and multimodal fusion strategies to further improve the robustness of the model in complex physical environments.

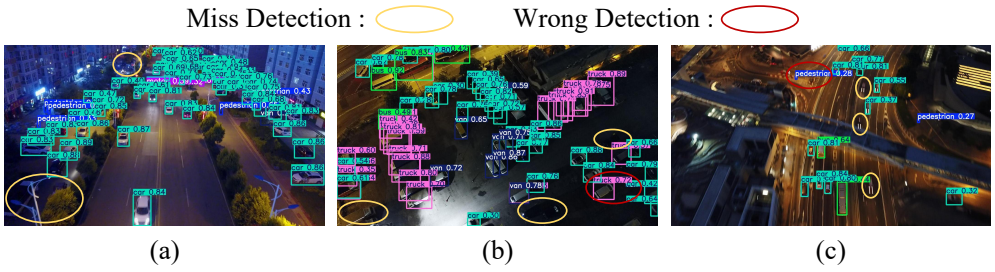


Figure 10. Failure case analysis of MSIA-YOLO under complex degraded scenarios. (**a**) Nighttime occlusion scene; (**b**) nighttime low illumination scene; (**c**) motion blur scene.

3.6.5. Feature Visualization and Module Synergy Analysis

To further analyze the synergistic effect among the proposed modules, we conduct feature visualization experiments, as shown in Figure 11. The first column presents the original UAV remote sensing images. Compared with the original images, the second column shows that after RFAConv processing, the local spatial receptive fields of tiny vehicles and pedestrians are effectively activated. The heatmap energy initially concentrates toward the target regions, indicating that the edge and texture contours of small objects

are preserved and enhanced. The third column shows the feature responses after ScalSeq based semantic compensation. After receiving the features extracted by RFAConv, ScalSeq further integrates local details with deep semantic information, making the responses more concentrated in target-related regions. More continuous high response distributions are formed around road vehicles, distant small objects, and densely distributed pedestrians. Meanwhile, the large vegetation and tree background regions on the left side of (a) and the right side of (b) show overall low responses, indicating that interference from complex natural backgrounds is suppressed to some extent. It should be noted that certain grid-like activations still appear in the open lane region near the center of (a). These responses can be attributed to spatial projections of deep low-resolution semantic features during cross-scale sequence fusion and upsampling. However, their overall intensity is lower than that of the target regions, and they do not form concentrated high responses comparable to real targets. This indicates that ScalSeq can preserve the discriminative distinction between target and background regions while introducing deep semantic information. The last column presents the DCNv2-based regression alignment responses in DTIA-Head. DCNv2 is embedded in the regression branch of DTIA-Head to enhance spatial alignment during bounding-box localization. The visualization results are shown in a response form oriented toward bounding-box regression. As can be observed, DCNv2 produces clear localization related responses around the detected small objects. These responses are mainly distributed inside the target boxes and around their boundary neighborhoods, while grid like activations are suppressed. This suggests that the regression branch can perform adaptive spatial alignment according to target locations and contours. In particular, in the distant dense pedestrian region of (b), multiple adjacent small objects still generate relatively independent local responses.

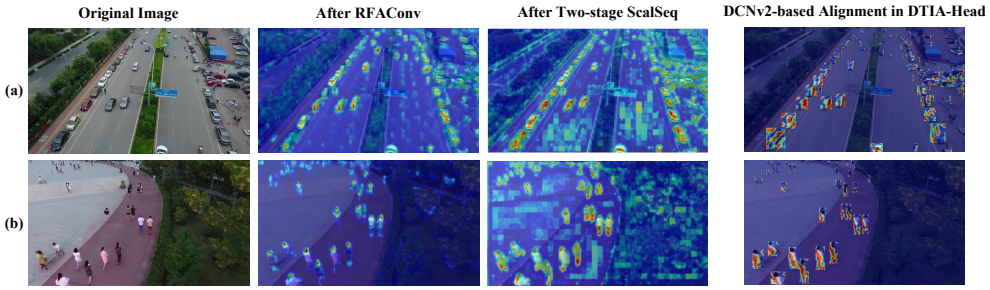


Figure 11. Feature visualization of MSIA-YOLO. The last column visualizes the DCNv2-based regression alignment response within DTIA-Head. (**a**) Road scene with dense vehicles; (**b**) distant scene with dense pedestrians.

Overall, RFAConv, ScalSeq in CHSP-P2, and DCNv2 in DTIA-Head do not function independently. RFAConv provides a clearer basis of local edges and texture details. ScalSeq then performs cross-scale fusion between these local details and deep semantic information and feeds the enhanced representation to the P2 branch. DCNv2 further uses the target-related features obtained from the preceding modules to achieve more precise bounding-box regression alignment. Together, these three components form a collaborative mechanism from detail enhancement to semantic compensation and then to spatial alignment. This enables MSIA-YOLO to better preserve small object contours, suppress background interference, and improve the localization accuracy of dense small objects in complex UAV remote sensing scenes.

4. Discussion

The experimental results demonstrate that MSIA-YOLO effectively improves small object detection performance in UAV remote sensing images through enhanced local detail

representation, multi-scale semantic compensation, and task-aware prediction. Compared with the baseline YOLOv11n model, the proposed method achieves clear improvements in both mAP50 and mAP50:95 on the VisDrone2019 dataset. This indicates that the introduced structural improvements are better suited to the characteristics of low-altitude UAV imagery, including small target scales, complex backgrounds, and dense object distributions. In addition, comparisons with YOLO series models and recent UAV/remote sensing object detection methods show that MSIA-YOLO achieves competitive detection accuracy while maintaining real-time inference capability, demonstrating a favorable accuracy efficiency trade-off.

The performance improvement in MSIA-YOLO mainly benefits from the synergy among its modules. RFAConv improves local feature representation through receptive field-level feature enhancement, which helps preserve edge, texture, and local structural information of small objects. For small objects with a low pixel proportion, shallow features contain rich spatial details but are easily affected by background noise. In contrast, deep features provide stronger semantic information but suffer from insufficient spatial resolution. CHSP-P2 addresses this issue through cross-scale semantic compensation, enabling the high-resolution branch to receive more sufficient semantic support. Finally, DTIA-Head further improves the collaboration between classification and localization through task decomposition, task interaction, and spatial alignment. This allows the detection head to obtain task-specific feature representations that are more suitable for classification discrimination and boundary regression.

From the perspective of accuracy and efficiency, the design of MSIA-YOLO does not simply pursue model lightweighting. Instead, it aims to enhance small-object feature representation and localization reliability while maintaining real-time inference capability. The experimental results show that introducing RFAConv, CHSP-P2, and DTIA-Head effectively improves small object detection performance in complex UAV remote sensing scenarios, although it also introduces a certain computational cost. This suggests that sufficient local detail modeling, multi-scale semantic compensation, and task alignment are necessary for improving detection accuracy in low-altitude UAV small object detection. Compared with merely reducing the number of parameters or computational cost, the proposed method places greater emphasis on achieving a reasonable balance among detection accuracy, model complexity, and inference efficiency.

This study still has certain limitations. In extremely dense scenes, severely occluded scenes, and cases where targets occupy only a very small number of pixels, the model may still produce missed detections or false detections. In addition, this study mainly focuses on single modal low-altitude UAV remote sensing images, while multimodal information, such as infrared imagery, may further improve detection robustness in complex environments. Future work will further optimize the computational complexity of the model, conduct deployment tests on real edge devices, and explore more robust feature representation methods for extremely small and densely distributed targets.

5. Conclusions

To address the key challenges of small object detection in low-altitude UAV remote sensing images, this study proposes MSIA-YOLO, a detection model based on multi-scale semantic interaction and alignment. The proposed method conducts systematic optimization from three aspects: feature modeling, multi-scale feature fusion, and task-coordinated alignment. MSIA-YOLO integrates RFAConv into the backbone network. By introducing receptive field-level attention, it overcomes the limitation of shared convolution and adaptively modulates local structural differences, thereby enhancing the effective representation of small object details, edges, and texture cues. In the neck, a CHSP-P2 small object detec-

tion framework with a P2 branch is constructed. Through scale sequence modeling and cross-stage hybrid feedback, semantic compensation is provided for the high-resolution branch. This alleviates the information loss of small objects caused by repeated downsampling and improves detection stability for distant and densely distributed small objects. DTIA-Head further enhances the collaborative optimization of classification and regression through shared feature refinement, task decomposition, and the dynamic alignment mechanism of DCNv2. It effectively reduces missed detections caused by task conflicts and improves localization quality.

In this study, ablation experiments on the overall framework and the neck module are conducted on the VisDrone dataset to evaluate the contribution of each improvement. In addition, the proposed method is compared with several widely used methods on three datasets. The experimental results verify its effectiveness for extremely small objects and dense scenes, as well as its cross-dataset generalization capability. Future work will further focus on model lightweighting, inference acceleration, and deployment testing on real edge devices, aiming to improve the applicability of the model in practical UAV remote sensing detection scenarios.

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

Funding: This research was funded by the Science and Technology Plan Project of Liaoning Province, grant number 2021JH2/10200023, and the Scientific Research Project of the Education Department of Liaoning Province, grant number LJZZ212410154029.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. VisDrone 2019, TinyPerson, and RSOD can be obtained from their official repositories or corresponding publications.

Acknowledgments: The authors would like to thank the providers of the VisDrone 2019, TinyPerson, and RSOD datasets for making their datasets publicly available.

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

References

1. Li, Q.; Kao, Y.; Zhang, Y.; Xu, P. A review of object detection for UAV aerial images. *J. Graph.* **2024**, *45*, 1145–1164. (In Chinese). Available online: http://www.txxb.com.cn/EN/abstract/abstract2347.shtml (accessed on 15 June 2026).
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv* **2016**, arXiv:1603.00278. [CrossRef]
4. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
5. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. Available online: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html (accessed on 21 November 2025).
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot MultiBox detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]

7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Redmon_YOLO9000_Better_Faster_CVPR_2017_paper.html (accessed on 21 November 2025).
9. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
10. 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475. [CrossRef]
11. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
12. 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]
13. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
14. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2021**, arXiv:2010.04159. [CrossRef]
15. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974. Available online: https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_DETRs_Beat_YOLOs_on_Real-time_Object_Detection_CVPR_2024_paper.html (accessed on 21 November 2025).
16. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
17. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. AugFPN: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604. Available online: https://openaccess.thecvf.com/content_CVPR_2020/html/Guo_AugFPN_Improving_Multi-Scale_Feature_Learning_for_Object_Detection_CVPR_2020_paper.html (accessed on 15 June 2026).
18. 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, 18–25 June 2018; pp. 8759–8768. Available online: https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html (accessed on 15 June 2026).
19. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045. Available online: https://openaccess.thecvf.com/content_CVPR_2019/html/Ghiasi_NAS-FPN_Learning_Scalable_Feature_Pyramid_Architecture_for_Object_Detection_CVPR_2019_paper.html (accessed on 15 June 2026).
20. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790. Available online: https://openaccess.thecvf.com/content_CVPR_2020/html/Tan_EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.html (accessed on 15 June 2026).
21. Xue, C.; Xia, Y.; Wu, M.; Chen, Z.; Cheng, F.; Yun, L. EL-YOLO: An efficient and lightweight low-altitude aerial objects detector for onboard applications. *Expert Syst. Appl.* **2024**, *256*, 124848. [CrossRef]
22. Pour, B.S.; Jozani, H.M.; Shokouhi, S.B. AL-YOLO: Accurate and lightweight vehicle and pedestrian detector in foggy weather. In Proceedings of the 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 27–28 November 2024; pp. 131–136. [CrossRef]
23. Gao, D.; Yin, Y.; Zhang, H.; Li, C.; Wang, B. YOLOv11-based UAV foreign object detection for power transmission lines. *Electronics* **2025**, *14*, 3577. [CrossRef]
24. Yan, L.; He, Z.; Zhang, Z.; Xie, G. LS-MambaNet: Integrating large strip convolution and mamba network for remote sensing object detection. *Remote Sens.* **2025**, *17*, 1721. [CrossRef]
25. Yuan, M.; Wei, X. C2Former: Calibrated and complementary transformer for RGB-infrared object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [CrossRef]
26. Yuan, M.; Shi, X.; Wang, N.; Wang, Y.; Wei, X. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Inf. Fusion* **2024**, *105*, 102246. [CrossRef]

27. Lim, J.-S.; Astrid, M.; Yoon, H.-J.; Lee, S.-I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186. [CrossRef]
28. Gong, X.; Yu, J.; Zhang, H.; Dong, X. AED-YOLO11: A small object detection model based on YOLO11. *Digit. Signal Process.* **2025**, *166*, 105411. [CrossRef]
29. Xie, C.; Wu, J.; Xu, H. An improved YOLOv5-based small object detection algorithm for UAV imagery. *Comput. Eng. Appl.* **2023**, *59*, 198–206. (In Chinese). [CrossRef]
30. Cao, L.; Xu, H.; Xie, G.; Li, Y.; Huang, X.; Chen, H.; Zhu, X. ASOD-YOLO: An improved YOLOv8n-based aerial small object detection algorithm. *Comput. Eng. Sci.* **2026**, 133–145.
31. Hou, W.; Wu, H.; Wu, D.; Shen, Y.; Liu, Z.; Zhang, L.; Li, J. Small object detection method for UAV remote sensing images based on αS-YOLO. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 8984–8994. [CrossRef]
32. Liu, X.; Zheng, Y.; Cai, Y.; Ding, Y.; Li, J.; Kang, W.; Cai, Z. RFHA-YOLO: Dynamic receptive field and adaptive hybrid attention for small-object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–13. [CrossRef]
33. Zhang, X.; Liu, C.; Song, T.; Yang, D.; Ye, Y.; Li, K.; Song, Y. RFAConv: Innovating spatial attention and standard convolutional operation. *arXiv* **2024**, arXiv:2304.03198. [CrossRef]
34. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
35. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499. [CrossRef]
36. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets V2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316. Available online: https://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_Deformable_ConvNets_V2_More_Deformable_Better_Results_CVPR_2019_paper.html (accessed on 21 January 2026).
37. Wu, Y.; He, K. Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. Available online: https://openaccess.thecvf.com/content_ECCV_2018/html/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.html (accessed on 22 January 2026).
38. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [CrossRef] [PubMed]
39. Jiang, N.; Yu, X.; Peng, X.; Gong, Y.; Han, Z. SM+: Refined scale match for tiny person detection. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 1815–1819. [CrossRef]
40. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
41. Wu, P.; Xu, Y.; Ma, Y.; Zhang, Y.; Xu, Y. LYA-YOLO: A lightweight and accurate YOLO model in drone aerial image scenes. *Expert Syst. Appl.* **2026**, *321*, 132166. [CrossRef]
42. 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, QC, Canada, 10–17 October 2021; pp. 2778–2788. Available online: https://openaccess.thecvf.com/content/ICCV2021W/VisDrone/html/Zhu_TPH-YOLOv5_Improved_YOLOv5_Based_on_Transformer_Prediction_Head_for_Object_ICCVW_2021_paper.html (accessed on 4 February 2026).
43. Zhang, Y.; Chen, X.; Sun, S.; You, H.; Wang, Y.; Lin, J.; Wang, J. Vehicle detection in drone aerial views based on lightweight OSD-YOLOv10. *Sci. Rep.* **2025**, *15*, 25155. [CrossRef] [PubMed]
44. He, Z.; She, R.; Tan, B.; Li, J.; Lei, X. SSCW-YOLO: A lightweight and high-precision model for small object detection in UAV scenarios. *Drones* **2026**, *10*, 41. [CrossRef]
45. Xu, Z.; Zhao, H.; Liu, P.; Wang, L.; Zhang, G.; Chai, Y. SRTSOD-YOLO: Stronger real-time small object detection algorithm based on improved YOLO11 for UAV imageries. *Remote Sens.* **2025**, *17*, 3414. [CrossRef]

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.