



ARTICLE

EMW-YOLO: A Detail-Preserving and Multi-Scale Fusion Detector for Remote Sensing Small Object Detection

Heng Wang¹, Shichao Li¹, Long Xu^{2,*}, Chuqiao Wang¹, Yanzhou Feng¹ and Zou Zhou^{1,3,4,*}

¹School of Information and Communication, Guilin University of Electronic Technology, Guilin, China

²School of Life Sciences and Medical Engineering, Guangxi Medical University, Nanning, China

³Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin, China

⁴Guangxi Academy of Artificial Intelligence, Nanning, China

*Corresponding Authors: Long Xu. Email: hahamyt211@outlook.com; Zou Zhou. Email: zhouzou@guet.edu.cn

Received: 03 April 2026; Accepted: 15 June 2026; Published: 13 August 2026

ABSTRACT: The inherent challenges of small objects in remote sensing imagery encompass the degradation of fine-grained spatial details throughout the downsampling stages, semantic inconsistency during multi-level feature fusion, along with unreliable localization caused by noisy samples. To address these issues, this paper proposes an efficient small-object detector termed EMW-YOLO. An Efficient Down-sampling (EDS) module is introduced to preserve fine-grained spatial information and enhance feature representation during feature extraction through spatial rearrangement and cross-dimensional attention. A Multi-Scale Fusion and Enhancement (MSFE) architecture is further developed to improve semantic consistency across feature levels by combining local enhancement with global feature alignment. In addition, the WIoU v3 loss function is also incorporated to suppress outlier samples and enhance the robustness of regression for small objects. Experimental results on the VisDrone2021 dataset demonstrate that EMW-YOLO achieves absolute improvements of 6.2%, 6.9%, and 7.9% over YOLOv8n in precision, recall, and $mAP_{0.5}$, respectively. Further evaluation on the NWPU VHR-10 and RSOD datasets confirms that the proposed method exhibits strong generalization capabilities.

KEYWORDS: Small object detection; attention mechanism; multi-scale fusion; remote sensing images; feature representation

1 Introduction

The advent of Earth observation technology has precipitated a paradigm shift in the realm of intelligent sensing systems, with high-resolution remote sensing imagery emerging as the cornerstone of contemporary systems [1], playing an irreplaceable role in critical domains such as strategic military reconnaissance [1], traffic monitoring [2] and precision agricultural yield estimation [3]. However, despite the exponential growth in data acquisition capabilities, the core bottleneck constraining advancement in this field remains its accurate extraction of high-value information from vast image datasets. The root cause lies in the inherent minuscule scale of objects of interest in remote sensing scenarios [4,5], which makes them highly susceptible to detail loss during feature extraction. Moreover, these minute visual features are frequently obscured by severe illumination variations and complex background noise [6]. This inherently low “signal-to-noise ratio” characteristic exhibited by such data gives rise to three key challenges:

First, the severe degradation of weak structural information emerges during the downsampling stage of feature extraction. To address the issues of high computational complexity and limited receptive field, existing mainstream networks (e.g., ResNet [7] and the YOLO [8] series) generally employ large-stride convolutions or pooling operations. However, such operations on small objects in remote sensing images can disrupt sub-pixel structures, causing significant feature loss in shallow layers [9]. In recent years, although detail-preserving downsampling methods such as SPD-Conv [10] or SliceSamp [11] have been proposed to alleviate this problem, they incur extremely high memory access overhead or merely assign a single static convolution kernel to each channel, making it difficult to capture variable local textures against complex remote sensing backgrounds. Therefore, designing an efficient downsampling mechanism that can preserve spatial geometric structures while possessing strong local feature representation capabilities is a primary prerequisite to prevent small objects from becoming “invisible”.

Second, a contradiction between semantic misalignment and feature submersion exists during the multi-scale feature fusion stage. In the face of scale variations, existing feature pyramid networks (e.g., FPN, PANet, and BiFPN [12]) mostly adopt uniform symmetric fusion strategies. However, in remote sensing scenarios, although shallow features encode rich details, they are accompanied by extensive noise; conversely, deep features possess strong semantics but entirely lose the spatial locations of small objects. Traditional symmetric fusion easily causes intermediate layers that hold key information for small objects to become diluted or overshadowed by the robust semantics of deep layers. Furthermore, the fixed static convolution kernels in conventional pyramids lack geometric adaptability to the arbitrary orientations and complex shapes of remote sensing objects. To this end, there is an urgent need for an asymmetric selective enhancement and fusion strategy to achieve precise protection and dynamic reshaping of critical-scale features.

Finally, the localization regression stage exhibits extreme sensitivity to annotation noise. Bounding box regression loss serves as the decisive factor for localization precision of a model. Most mainstream algorithms utilize static focusing mechanisms such as CIoU or EIou. While these methods perform excellently in natural images, they neglect the particularities specific to remote sensing images: owing to the extremely tiny objects coupled with complex backgrounds, the ground-truth boxes in remote sensing datasets inevitably suffer from human annotation errors and boundary ambiguity [13]. When encountering such “low-quality” annotated samples, static loss functions still impose strong geometric penalties, thereby generating a massive number of harmful gradients. This forces the model to fit incorrect annotations, resulting in a highly unstable training process and severe fluctuations in localization performance. This requires the introduction of a dynamic loss mechanism capable of perceiving the outlieriness of samples and intelligently suppressing the harmful gradients induced by noise.

The study aims to construct a highly precise small-object detection architecture that provides robust detection solutions. To resolve the issues of deficient small-object feature characterization and background disturbance, we present a novel strategy that integrates information fidelity, multi-scale interaction, and dynamic focus regression. The key contributions made by this research are summarized below:

- (1) To address the degradation of feature information during the downsampling process, we design the Efficient DownSampling (EDS) module by coupling spatial rearrangement with cross-dimensional attention, establishing an effective feature extraction paradigm for retaining structural details of small objects.
- (2) To tackle the issue of semantic misalignment during multi-scale fusion, we put forward the Multi-Scale Fusion and Enhancement (MSFE) architecture by integrating adaptive feature enhancement with hierarchical cross-scale fusion to facilitate semantic interaction across different feature levels.

- (3) To resolve the instability in localization regression, we replace the original CIoU loss with WIoU v3 by introducing a dynamic gradient allocation strategy that enables more effective optimization of training samples with different qualities.

The rest of the paper is structured as follows. [Section 2](#) surveys relevant research on small object detection in remote sensing. [Section 3](#) presents EMW-YOLO's architecture and structural modules. [Section 4](#) describes the experimental setup and performance evaluation metrics and analytically compares EMW-YOLO with various models and methods, verifying the generalization performance of the proposed model on multiple datasets. [Section 5](#) summarizes the paper and discusses future developments in remote sensing for object detection.

2 Related Work

2.1 Applications of YOLO in Remote Sensing

As deep learning research paradigms have undergone constant evolution, CNN-based architectures have gradually solidified their pivotal position in object detection methodologies, attaining remarkable advancements in visual domains including remote sensing imagery analysis [14,15]. In recent years, the single-stage detector YOLO has become the dominant approach for remote sensing target detection [8]. In contrast to conventional two-stage methods [16,17], the YOLO family models [18–20] and their improved variants—FFCA-YOLO [21], Hyper-YOLO [22], and TPH-YOLOv5 [23] have effectively mitigated typical challenges in remote sensing scenarios. These challenges include the semantic misalignment of multi-scale objects, missing feature associations in complex backgrounds, and degradation of small object feature information. This has been realized via a range of improvement strategies, including multi-scale feature integration, hypergraph-based computation, and context-aware processing.

Specifically, Hyper-YOLO [22] integrates hypergraph computation into the backbone network to model complex high-order correlations within feature maps, thereby enhancing semantic expression to counteract complex background interference. TPH-YOLOv5 [23] introduces a Transformer Prediction Head (TPH), which effectively mitigates information degradation in the downsampling process via capturing rich global contextual information.

Besides, recent advances in lightweight remote sensing object detection focus on balancing computational constraints with feature representation capability. For instance, M2FE-YOLO employs multi-branch feature enhancement and lightweight aggregation structures to achieve efficient multiscale fusion [24]. Similarly, LMW-YOLO optimizes YOLO architectures with advanced lightweight convolutional mechanisms to reduce parameters while maintaining accuracy for small objects [25]. Furthermore, SEMA-YOLO introduces shallow-layer enhancement and receptive field attention to dynamically capture local details [26]. While these state-of-the-art methods effectively reduce computational overhead, they inevitably face performance bottlenecks caused by sub-pixel structural fragmentation during downsampling and semantic misalignment across complex multiscale levels.

2.2 Efficient Downsampling

The degree of detail preservation during subsampling directly impacts the comprehensive performance of the network, particularly in remote sensing imagery scenarios with dense small objects [23]. SliceSamp [11] enhances local detail retention by combining spatial slicing with depth-separable convolutions while maintaining high computational efficiency. However, its slice-based feature rearrangement approach can lead to fragmented information in complex backgrounds, resulting in insufficient continuity preservation for small object edge structures.

Recently, the Global Dense Local Aggregation (GDLA) module [27] effectively reduces the potential loss in fine-grained semantic details during downsampling by leveraging a combination of parallel convolutionalization, maximum pooling, and channel aggregation mean pooling to capture diverse representations. Furthermore, Strip R-CNN [28] breaks the geometric limitations of traditional square convolutions by implementing cross-striped convolutions for long-range directional correlation capture. Synergistically combining medium kernels, large kernels, and pointwise convolutions enables comprehensive feature extraction across diverse spatial dimensions, directions, and hierarchical levels. Despite these advances in improving subsampling performance, efficiently preserving structural and textural information during subsampling remains a significant challenge.

2.3 Multi-Scale Feature Fusion and Enhancement

Small objects in remote sensing imagery usually account for only a small number of pixels. Effective multi-scale representations are essential, as detection models built on deep learning hinge on backbone architectures to extract high-dimensional semantic features. Traditional feature pyramids generally adopt implicit multi-scale concatenation or addition. To preserve small object structures more precisely, Xiao et al. [29] proposed EdgeFusionNet, which breaks away from traditional implicit fusion limitations by explicitly mining boundary-associated edge information from low-level layers. This method utilizes ScaleFusion blocks to perform bidirectional cross-attention interaction among deep semantic features and shallow edge feature maps. Shi et al. [30] proposed HS-FPN, introducing an adaptive feature selection mechanism, which automatically chooses optimal features for fusion according to their effectiveness at different scales, further enhancing multi-scale representation capabilities. In addition to optimizing feature fusion strategies, pre-fusion feature enhancement operations are equally crucial. Dai et al.'s [31] proposed feature enhancement module employs multi-scale channel attention and an iterative fusion mechanism to more effectively integrate multi-scale features with inconsistent semantic hierarchies.

While the above-mentioned approaches have achieved remarkable advances in feature integration and enhancement, multi-scale fusion may still suffer from information redundancy and semantic misalignment. The integration of high-level semantic features and low-level detail information remains relatively rigid, making it difficult to adaptively address issues across different scenarios.

2.4 Bounding Box Regression Loss Function

Bounding box regression loss typically requires a comprehensive consideration of multiple factors related to the target geometry. The GIoU [32] addresses the zero-gradient issue in non-overlapping scenarios by introducing the minimum bounding rectangle between predicted and ground-truth boxes to measure their overlap. Subsequently, the proposed DIoU [33] incorporates the Euclidean distance between box centers, achieving faster convergence even for overlapping targets. Building upon the DIoU, the complete IoU (CIoU) [34] builds on DIoU by introducing aspect ratio consistency to improve regression stability, while EIoU [35] achieves finer optimization by decomposing width and height errors separately.

Despite being the default loss function in YOLOv8, CIoU exhibits multiple non-negligible drawbacks. A typical scenario arises when the predicted bounding box and its corresponding ground-truth box have identical aspect ratios despite having different dimensions, the aspect ratio component of CIoU fails to deliver effective gradients, which may hinder further model optimization.

3 Method

3.1 Overview

We adopt YOLOv8 as our baseline framework, as it strikes an optimal balance between detection accuracy and computational efficiency and it has been widely utilized as a representative benchmark in recent remote sensing object detection research. The general architectural design adopted in EMW-YOLO is illustrated in Fig. 1. Specifically, the Efficient DownSampling (EDS) module is embedded into the backbone to replace conventional downsampling operations, while the Multi-Scale Fusion and Enhancement (MSFE) architecture is integrated in the multi-scale feature fusion neck to improve cross-scale feature fusion. In addition, the WIoU v3 loss function is incorporated to reduce the adverse impact caused by low-quality samples on the accurate localization of small objects. The details of each component are presented in the following sections.

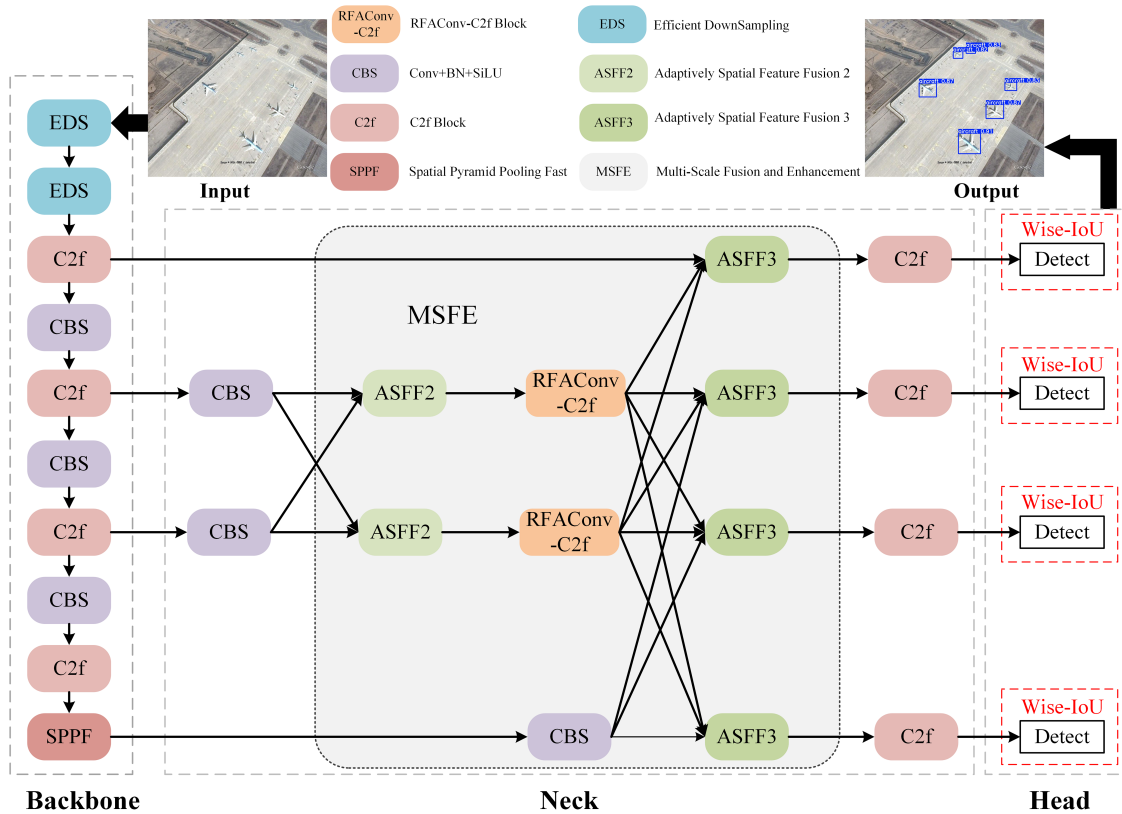


Figure 1: The architecture of the proposed EMW-YOLO framework. Its backbone adopts EDS to extract multi-scale features, while the MSFE neck integrates ASFF and RFACConv-C2f modules to achieve progressive cross-scale feature fusion. The detection head further applies WIoU dynamic focusing to improve localization accuracy.

3.2 EDS: Efficient DownSampling Module

Owing to the complexity of remote sensing scenes, conventional convolutional downsampling modules exhibit limited feature extraction capabilities. Meanwhile, small objects inherently possess weak semantics and are thus more susceptible to information loss during the sampling process. Although existing detail-preserving methods (e.g., SliceSamp [11]) can mitigate this issue, their frequent explicit indexing incurs exorbitant memory access costs. More crucially, this limits its ability to effectively extract fragile edge

contours and variable texture features from tiny targets in intricate remote sensing environments. To mitigate this limitation, we design the proposed EDS module, with its overall architecture depicted in Fig. 2.

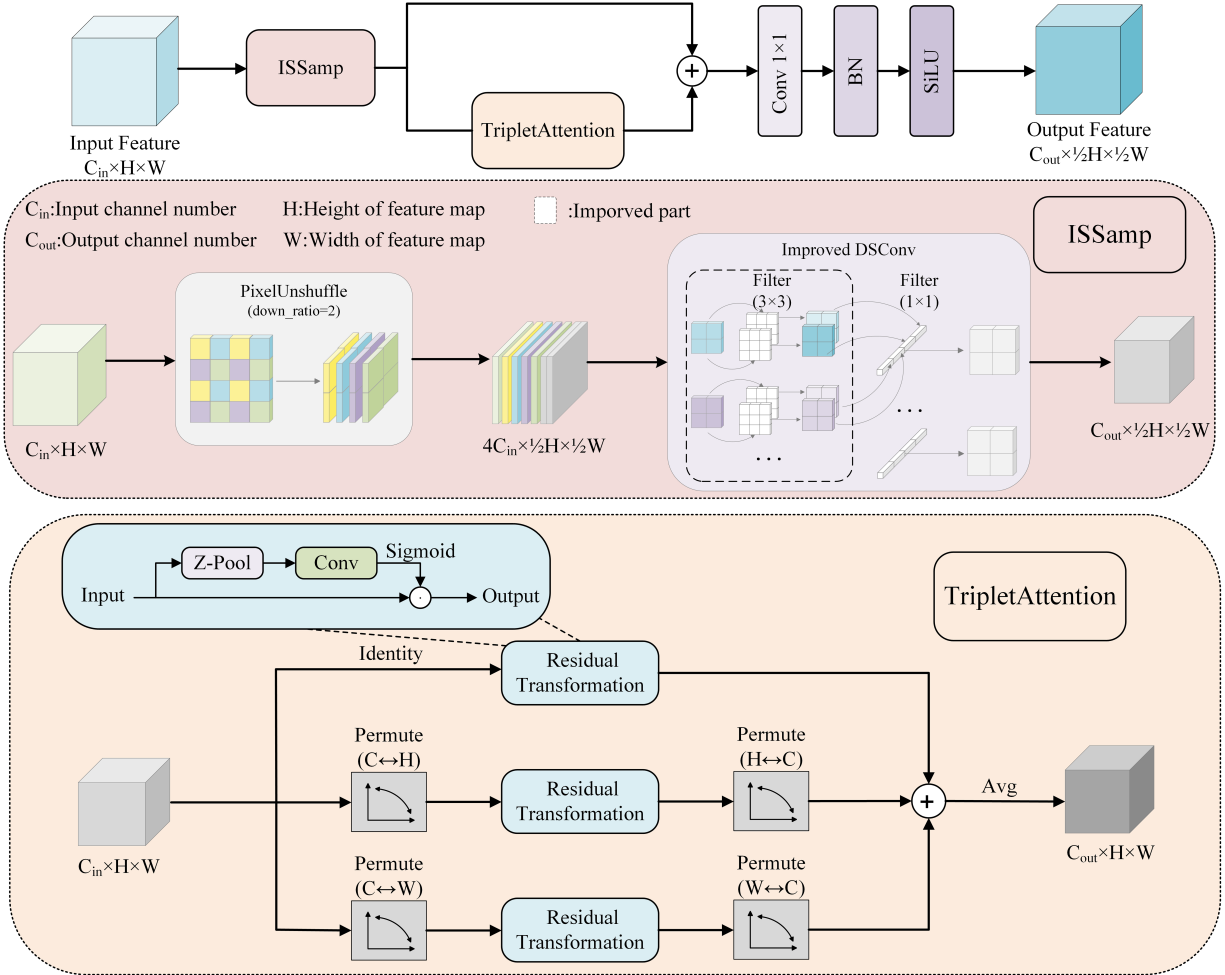


Figure 2: The structure of the proposed efficient downsampling (EDS) module, which simultaneously utilizes ISSamp for spatial downsampling and TripletAttention for cross-dimension interaction to preserve rich feature representations.

First, the PixelUnshuffle operation [36] introduced in EDS replaces conventional downsampling with spatial rearrangement, preserving texture continuity and geometric structures of small objects while avoiding the memory overhead associated with SliceSamp. Second, to accommodate the significant orientation variations and complex backgrounds in remote sensing imagery, multiple depthwise convolution kernels are employed to capture richer local gradient information and strengthen feature representation. Furthermore, a spatial-channel triplet attention mechanism [37] is incorporated to enhance cross-dimensional feature interaction, enabling more effective suppression of background interference and emphasizing discriminative target responses. Finally, these components are integrated within a residual learning framework [7], which facilitates stable optimization and effective fusion of shallow spatial details with attention-enhanced semantic features, thereby improving the preservation of fine-grained information throughout the downsampling process. This is formulated as follows:

$$X_{ps} = \text{PixelUnShuffle}(2, X) \quad (1)$$

$$X_{ISS} = GELU \left(BN_2 \left(W_2^{PW} * GELU \left(BN_1 \left(W_1^{IDW} \odot X_{ps} \right) \right) \right) \right) \quad (2)$$

where $X \in \mathbb{R}^{C \times H \times W}$ denotes the input tensor and $X_{ps} \in \mathbb{R}^{4C \times H/2 \times W/2}$ is obtained after the PixelUnShuffle operation; $W_1^{IDW} \in \mathbb{R}^{4dC \times 4C \times 3 \times 3}$ represents the expanded depthwise convolution kernel, whose output dimension is $4dC$, $W_2^{PW} \in \mathbb{R}^{C_{out} \times 4dC \times 1 \times 1}$ corresponds to the pointwise convolution kernel.

$$\widehat{X}^* = Concat \left(MaxPool_0 \left(\widehat{X} \right), AvgPool_0 \left(\widehat{X} \right) \right) \quad (3)$$

where $MaxPool_0()$ indicates transforming the 0-th dimension to 1, and $Concat(MaxPool_0(\widehat{X}), AvgPool_0(\widehat{X}))$ denotes the Z-pool operation. The input X is rotated 90° counterclockwise about the H -axis to generate tensor $\widehat{X}_1$ with dimensions of $(W \times H \times C)$. Subsequently, The Z-pool operation produces tensor $\widehat{X}_1^*$ with dimensions of $(2 \times H \times C)$. Similarly, a 90° counterclockwise rotation along the W -axis generates tensor $\widehat{X}_2^*$ with a shape of $(2 \times C \times W)$.

$$X_{TA}(\cdot) = \frac{1}{3} \left(\widehat{X}_1^* \sigma(\phi_1(\widehat{X}_1^*)) + \widehat{X}_2^* \sigma(\phi_2(\widehat{X}_2^*)) + X_{ISS} \sigma(\phi_3(\widehat{X}_3)) \right) \quad (4)$$

$$Y = SiLU \left(BN_3 \left(Conv_{1 \times 1} \left(X_{ISS} + X_{TA}(X_{ISS}) \right) \right) \right) \quad (5)$$

where d represents the depth multiplier, while $*$ and $\odot$ indicate the standard convolution operation and the depthwise convolution operations, respectively. The overline ($\widehat{\cdot}$) signifies a 90° clockwise rotation operation in Eq. (4), which is applied to the first two branches for recovering permuted feature maps into their original input dimensions $(C \times H \times W)$ for subsequent aggregation [37], and the final output is Y .

3.3 MSFE: Multi-Scale Fusion and Enhancement Architecture

Unlike the uniform or symmetric feature pyramid construction methods adopted by traditional FPN or BiFPN, MSFE designs an asymmetric selective enhancement strategy based on the prior knowledge that small objects are densely distributed in intermediate feature levels [38]. The architecture of the MSFE module is shown in Fig. 3. In remote sensing images, small objects are highly susceptible to occlusion caused by complex background clutter. For instance, although the shallow feature X_2 , sampled with high fidelity by EDS, contains rich physical details, it inevitably incorporates a massive amount of background noise and computational redundancy. Conversely, while the deep feature X_5 processed by SPPF possesses high-level global semantics, the inherently weak spatial signals of small objects have severely decayed or even been completely lost after multiple downsampling operations. Therefore, the intermediate features X_3 and X_4 extracted by the backbone network achieve a favorable balance between preserving basic object structures and preliminarily filtering out background noise, serving as the critical intermediate features that carry the critical information for remote sensing small object detection.

Traditional uniform fusion is highly prone to causing the strong semantics of deep layers to overshadow the fragile signals associated with small objects in intermediate feature maps. To alleviate this issue, MSFE adopts an asymmetric fusion pathway characterized by “local focus enhancement followed by global semantic alignment,” as shown in Fig. 3. First, for the critical intermediate features X_3 and X_4 , feature X_4 is upsampled by a factor of 2 and feature X_3 is downsampled by a factor of 2 to achieve scale alignment. Subsequently, the adaptive weight mechanism of ASFF2 [39] is utilized to generate the pre-fused features X_3' and X_4' , respectively.

As shown in Fig. 3, to further disentangle object edges and faint textures from complex backgrounds, the RFACnv-C2f module (c) is introduced for deep feature reshaping. As its core component, the RFACnv-Conv_neck (b) replaces the standard Bottleneck (a). Compared with the static receptive fields of traditional

convolutions, RFACnv [40] dynamically generates “region-dependent” receptive field weights. This capability allows the model to effectively accommodate the orientation variations and arbitrary geometric structures exhibited by small objects in remote sensing scenes, thereby overcoming the parameter learning limitations of conventional convolutions to yield the more discriminative enhanced features X_3'' and X_4'' . At the same time, the deep feature X_5 is channel-aligned through convolutional layers to produce the features X_5' .

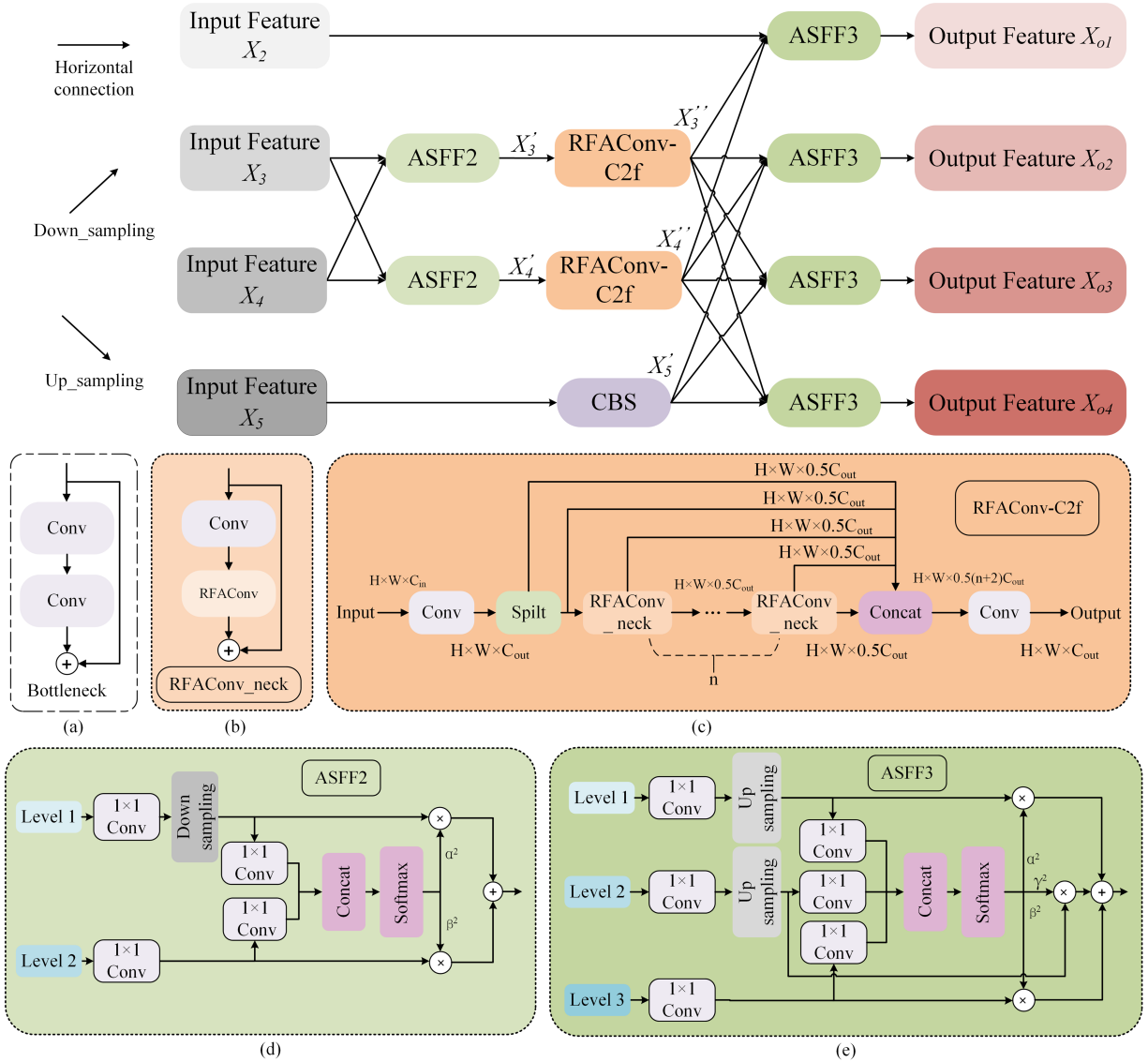


Figure 3: Structure of MSFE architecture (a) Standard Bottleneck in YOLOv8 (b) Improved bottleneck (RFACnv_neck) integrating RFACnv (c) Improved RFACnv-C2f module utilizing RFACnv_neck (d) Structure of ASFF2 (e) Structure of ASFF3.

Finally, to eliminate the semantic gap across multi-scale features and provide global context, a parallel ASFF3 structure is introduced into the module for global dense alignment. As illustrated in Fig. 3, ASFF3 receives the raw detail feature X_2 from the shallow layer, the selectively enhanced intermediate features X_3'' and X_4'' , and the deep semantic feature X_5' .

The ASFF3 module performs three precise steps: First, 1×1 convolutions are utilized to strictly align the channel dimensions of the multi-scale inputs. Second, these aligned features are concatenated along the channel dimension and fed into a softmax activation function to dynamically generate spatial, pixel-wise weight matrices. Finally, these adaptive weights are imposed on their corresponding input feature representations through element-wise multiplication and summation. Through this asymmetric strict alignment and weighted fusion, MSFE effectively circumvents the interference from invalid background features, ultimately outputting a comprehensive feature X_o that integrates high-resolution faint details with robust contextual semantics. The computational pipeline of MSFE can be mathematically described as below:

$$f_{down}^2 = Conv_{3 \times 3} (stride = 2) \quad (6)$$

$$f_{up}^2 = Interpolation_{\times 2} (Conv_{1 \times 1} (stride = 1)) \quad (7)$$

where f_{down}^2 denotes 2 downsampling, f_{up}^2 denotes 2 upsampling, $Interpolation_{\times 2}$ represents two times interpolation

$$X_3'' = RFA \left(ASFF2 \left(X_3, f_{up}^{1/2} (X_4) \right) \right) \quad (8)$$

$$X_4'' = RFA \left(ASFF2 \left(X_4, f_{down}^{1/2} (X_3) \right) \right) \quad (9)$$

$$X_5' = SiLU (BN_3 (Conv_{1 \times 1} (X_5))) \quad (10)$$

$$X_{oi} = ASFF3 (X_i^1, X_i^2, X_i^3), i \in \{1, 2, 3, 4\} \quad (11)$$

where RFA indicates the RFACConv-C2f operation, $ASFF2$ and $ASFF3$ indicating two-input and three-input adaptively spatial feature fusions and X_i^1, X_i^2, X_i^3 refers to the input feature that corresponds to the i -th output X_{oi} .

3.4 WIoU V3: Improved Loss Function

When performing visual detection tasks, a model's spatial localization accuracy is directly governed by its bounding box regression loss. Fig. 4 illustrates the core geometric metric relationships involved in the bounding box regression process, where B_{gt} and B_{pr} represent the areas of the ground-truth box and the inferred bounding region, respectively. Symbols c_w and c_h refer to the width and height dimensions of the minimum enclosing rectangle covering both; the respective widths alongside heights of the inferred bounding box and matched ground-truth box are w_{pr} , h_{pr} , w_{gt} and h_{gt} , respectively; the coordinates of their corresponding center points are $(b_{cx}^{pr}, b_{cy}^{pr})$ and $(b_{cx}^{gt}, b_{cy}^{gt})$, respectively; and $\rho(B_{gt}, B_{pr})$ denotes the straight-line distance separating the centroids of the two boxes.

YOLOv8 utilizes CIoU [34] as the default regression loss. Beyond considering the overlapping area and the central point distance, CIoU incorporates an extra consistency constraint that penalizes discrepancies in bounding box proportions, formulated as follows [34]:

$$L_{CIoU} = 1 - \frac{B_{gt} \cap B_{pr}}{B_{gt} \cup B_{pr}} + \frac{\rho^2(B_{gt}, B_{pr})}{(c_w)^2 + (c_h)^2} + \alpha v \quad (12)$$

$$v = \frac{4}{\pi^2} \left(\tan^{-1} \left(\frac{w_{gt}}{h_{gt}} \right) - \tan^{-1} \left(\frac{w_{pr}}{h_{pr}} \right) \right)^2 \quad (13)$$

$$\alpha = \frac{v}{1 - IoU + v} \quad (14)$$

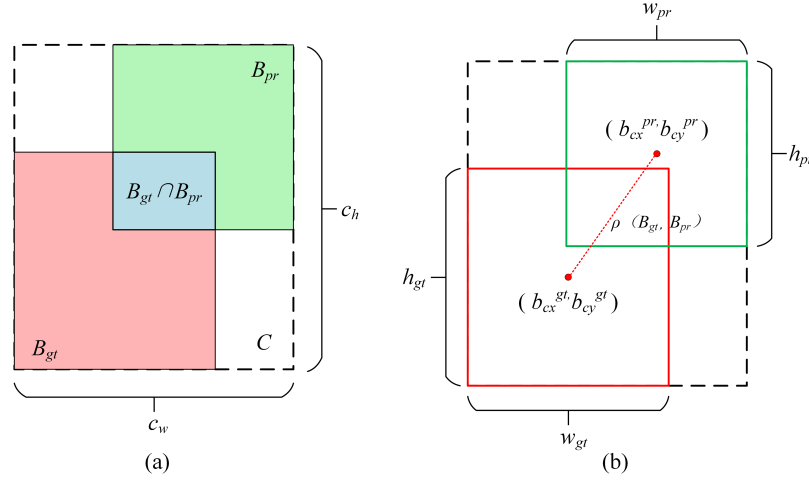


Figure 4: Schematic diagram of basic geometric distance and scale metrics in bounding box regression. (a) Overlap area and minimum enclosing rectangle of the ground-truth and predicted boxes. (b) Center-distance, width, and height metrics of the ground-truth and predicted boxes.

However, directly applying this static focusing mechanism to the recognition of small-scale targets in remote sensing images reveals critical limitations. First, the term v in the aforementioned formula involves complex inverse trigonometric functions, which introduces additional computational overhead during the training phase. More importantly, CIoU lacks a differential perception of sample difficulty and quality.

Considering that small-scale instances in aerial and satellite data take up only a marginal pixel area, they are easily obscured by complex environmental clutter, the ground-truth bounding box annotations in datasets inevitably contain noise and human-induced boundary errors. When the model encounters such “low-quality” small objects with significant annotation deviations, CIoU continues to impose strong geometric penalties. This mechanism generates numerous harmful gradients, forcing the model to fit incorrect annotations. Consequently, it interferes with network convergence on genuine weak features and degrades localization performance.

To break through this bottleneck, this paper discards conventional static loss mechanisms and introduces the WIoU series of loss functions [41]. Therefore, WIoU v3 is adopted because of the small targets’ characteristics in the remote sensing data, which has a non-monotonic focusing mechanism.

To reduce computational redundancy, WIoU v1 first removes the computationally expensive aspect ratio penalty present in CIoU. Instead, it utilizes only geometric properties such as the center point distance illustrated in Fig. 4 to construct an attention-based distance metric R_{WIoU} [41]:

$$L_{Wise-IoUv1} = R_{WIoU} L_{IoU} \quad (15)$$

$$L_{IoU} = 1 - \frac{B_{gt} \cap B_{pr}}{B_{gt} \cup B_{pr}} \quad (16)$$

$$R_{WIoU} = \exp \left(\frac{(b_{cx}^{gt} - b_{cx}^{pr})^2 + (b_{cy}^{gt} - b_{cy}^{pr})^2}{(c_w)^2 + (c_h)^2} \right) \quad (17)$$

However, in complex remote sensing scenarios, extremely low-quality samples (such as faint small objects with severe annotation deviations) generate highly destructive harmful gradients, which the monotonic focusing mechanism is incapable of addressing. Consequently, this paper adopts the advanced WIoU

v3, which innovatively introduces the outlieriness β and a non-monotonic weighting parameter r , specifically designed to adaptively suppress the detrimental gradient signals induced from substandard samples:

$$L_{WIoUv3} = r \times L_{WIoUv1} \quad (18)$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \quad (19)$$

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \in [0, +\infty) \quad (20)$$

here, α and δ are adjustable hyperparameters, L_{IoU}^* is the coefficient of monotonic focusing, and $\overline{L_{IoU}}$ is the average of L_{IoU} . The quality of the prior box is explicitly quantified by β . Specifically, a decrease in β implies enhanced anchor performance, thereby triggering an increased gradient multiplier r ; conversely, when β increases abnormally, the non-monotonic mechanism causes r to decrease rapidly.

Through this dynamic gradient allocation strategy, WIoU v3 can intelligently reduce the weights of harmful gradients induced by annotation errors, preventing the model from undesired fitting to noisy annotations. Such an enhancement allows the optimization objective to place greater emphasis on average-quality samples that genuinely enhance generalization capabilities. By effectively suppressing the influence of outlier samples, this dynamic mechanism improves the regression of bounding frames in complex remote sensing environments.

4 Experimental Results

4.1 Experimental Dataset Description

Throughout this research, minute targets are characterized as instances possessing a spatial area of under 32×32 pixels. The experiments utilize three open-source satellite and UAV image datasets: VisDrone2021 [42], NWPU VHR-10 [14], and RSOD [43]. Their bounding box distributions and scatter distributions are illustrated in Fig. 5.

- (1) VisDrone2021 [42]: This dataset encompasses diverse target classes, including pedestrians, vehicles, and bicycles. As shown in Fig. 5a, the scatter points are densely concentrated in the low-value region, and the bounding-box distributions exhibit heavily nested rectangles. This explicitly reveals a severe scale imbalance consisting primarily of small objects with diverse aspect ratios.
- (2) NWPU VHR-10 [14]: The dataset constitutes a high-resolution remote sensing benchmark dedicated to geospatial target detection, consisting of 715 color images whose resolutions range between 0.5 and 2 m and 85 panchromatic-sharpened color infrared images with 0.08 m. As illustrated in Fig. 5b, the bounding box distributions for NWPU VHR-10 exhibit a broader variance in scale compared to VisDrone2021, though the scatter plot confirms that the dense concentration of targets remains squarely in the small-scale region.
- (3) RSOD [43]: An open-source benchmark dataset for remote sensing target detection released by Wuhan University, consisting of 936 images with a resolution of roughly 1000×1000 pixels. It includes four typical object classes: aircraft, playgrounds, overpasses, and oil tanks. As shown in Fig. 5c, the scatter points for the RSOD dataset display a highly concentrated distribution pattern. Its corresponding bounding box distribution is highly uniform, indicating that its specific remote sensing classes possess highly consistent aspect ratios.

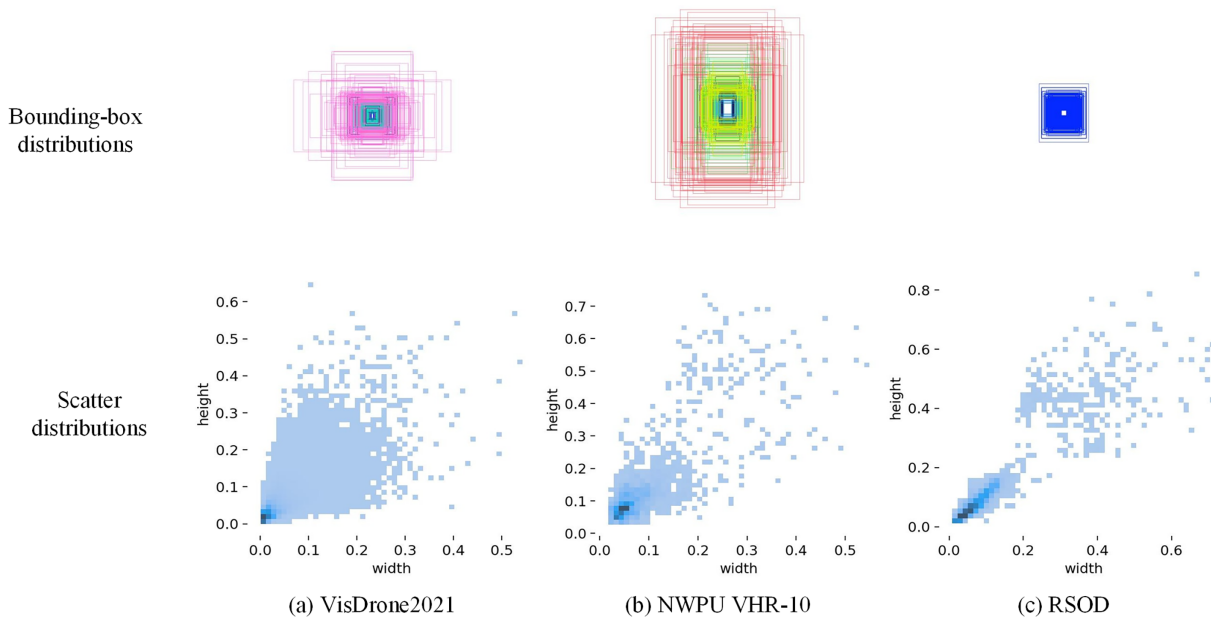


Figure 5: Bounding box and scatter distributions of datasets used in experiments.

4.2 Experimental Setup

All experiments in this work were implemented on a hardware platform fitted with a 12th-generation Intel[®] Core™ i5-12600KF central processing unit and an NVIDIA GeForce RTX 4070 SUPER graphics processing unit. As for the software configuration, it was built on the Windows 11 operating environment, with PyTorch 2.1.2 adopted as the core deep learning framework. The stochastic gradient descent (SGD) optimizer was utilized for model training, and its main hyperparameters were set as follows: the batch size was 16, the input image resolution reached 640×640 , the worker count was set to 4, the initial learning rate was 0.01, the momentum was 0.937, the weight decay was 0.0005, and the total number of training epochs was 300. An early stopping strategy with 50-epoch patience was adopted to avoid overfitting.

To ensure a fair comparison and validate its effectiveness, EMW-YOLO was trained from scratch without specific fine-tuning. Most core hyperparameters, such as momentum, initial learning rate, and weight decay, were strictly aligned with the YOLOv8 default settings to demonstrate that the performance gains stem from structural innovations rather than hyperparameter optimization. Batch size and worker settings were adjusted according to the available GPU resources.

Standard YOLOv8 data augmentation pipelines were applied during training. To preserve the spatial integrity of small objects, no additional dataset-specific augmentations or complex image preprocessing steps were introduced.

The performance of the proposed detector is evaluated using the VisDrone2021 benchmark. Training of the model was performed on 6471 images from the training partition, while its detection performance was validated against 548 images sampled from the validation subset. Generalization experiments were performed using the NWPU VHR-10 and RSOD datasets, with 540 and 748 training images, and 153 and 188 test images, respectively.

4.3 Experimental Evaluation Metrics

This study employs six evaluation metrics to fully assess the detection capability of the EMW-YOLO model, covering precision, recall, F1-Score, mean average precision, parameter scale and floating-point operations (GFLOPs). Specifically, true positives (TP), false positives (FP) and false negatives (FN) serve as the core basis for defining precision and recall:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \quad (21)$$

In remote sensing scenarios, models frequently face a severe trade-off between missing weak small objects and generating false alarms from complex background clutter. The F1-Score is adopted to measure the harmonic average of Precision and Recall:

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \quad (22)$$

A P-R curve can be generated through the calculation of precision and recall at diverse confidence thresholds. The integral area of this curve corresponds to the average precision of a single class. Ultimately, the mean average precision is acquired by averaging all class-based average precision results:

$$AP = \int_0^1 P(R) dR, mAP = \frac{\sum_{i=1}^{N_{cls}} AP}{N_{cls}} \quad (23)$$

mAP can be classified into $mAP_{0.5}$ and $mAP_{0.5:0.95}$ based on varied IoU threshold settings. $mAP_{0.5}$ is the value when IoU is 0.5, while $mAP_{0.5:0.95}$ is calculated by averaging the respective APs across the interval $\text{IoU} \in [0.5, 0.95]$ with 0.05 increments. Params reflects the model's size and storage requirements, while GFLOPs is a key metric for evaluating computational complexity.

Furthermore, based on the COCO evaluation criteria, objects are categorized into small ($Area < 32^2$), medium ($32^2 \leq Area \leq 96^2$), and large ($Area > 96^2$) objects according to their pixel areas, and their corresponding average precision (AP_S, AP_M, AP_L) and average recall (AR_S, AR_M, AR_L) are calculated, respectively. To assess our presented model's capability for small-sized object detection, this study takes AP_S and AR_S as the main evaluation indicators.

4.4 Comparative Experiment

We carried out a wide range of comparative experiments using the VisDrone2021 benchmark dataset in this work. A total of 12 representative target detection algorithms were selected for benchmark evaluation in this experiment, including the Transformer-based detector RT-DETR [44], the two-stage detector Faster R-CNN, the lightweight detector DEIM [45], mainstream YOLO series models (YOLOv5s [18], YOLOv8n/s, YOLOv11n/s [19], and YOLOv12n/s [20]), as well as recent remote sensing-oriented methods FFCA-YOLO [21] and Hyper-YOLO [22]. For a clearer comparison of training dynamics, Fig. 6 presents the convergence curves of six representative YOLO-based models. As can be seen from the training curve in Fig. 6, EMW-YOLO achieves rapid convergence while consistently maintaining superior performance throughout the training process, demonstrating both stable optimization behavior and competitive detection capability.

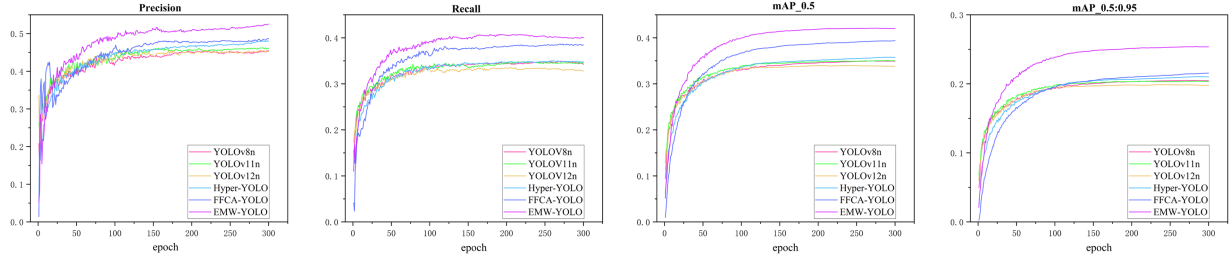


Figure 6: Training process curves of representative models on the VisDrone2021 dataset.

As quantified in Table 1, EMW-YOLO obtains the optimal mAP and F1-Score among all comparison methods. In contrast to the baseline YOLOv8n, EMW-YOLO delivers remarkable enhancements in both accuracy and recall. Crucially, in cross-scale comparisons, EMW-YOLO comprehensively outperforms models like YOLOv11s and YOLOv12s while maintaining nearly a third of their parameter scale. It also maintains a distinct performance edge over specialized remote sensing detectors such as FFCA-YOLO and Hyper-YOLO, validating the efficacy of our structural innovations.

Table 1: Comparison of experimental metrics for YOLO-based algorithms on the VisDrone2021 dataset.

Models	P	R	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params	GFLOPs	F1-Score
Faster R-CNN	–	–	0.309	0.171	41.2M	118.8	–
RT-DETR	0.521	0.379	0.343	0.186	19.9M	57.0	0.44
DEIM	–	–	0.29	0.161	3.7M	7.2	–
YOLOv5s	0.468	0.347	0.345	0.192	7.04M	15.8	0.39
YOLOv8n	0.46	0.331	0.343	0.201	2.69M	6.95	0.38
YOLOv8s	0.499	0.379	0.39	0.237	9.84M	23.6	0.44
YOLOv11n	0.46	0.345	0.351	0.204	2.58M	6.3	0.39
YOLOv11s	0.523	0.398	0.411	0.249	9.43M	21.3	0.44
YOLOv12n	0.449	0.334	0.34	0.199	2.51M	5.8	0.38
YOLOv12s	0.525	0.397	0.412	0.249	9.1M	19.6	0.45
FFCA-YOLO	0.51	0.4	0.414	0.233	3.18M	23.2	0.42
Hyper-YOLO	0.48	0.349	0.358	0.211	3.94M	10.8	0.40
EMW-YOLO	0.522	0.4	0.422	0.255	3.37M	24.44	0.45

These improvements verify the rationality of the model design, which mitigates the loss of fine-grained small-target features via network downsampling and multi-scale feature alignment, combined with WIoU v3 to improve bounding box regression stability in complex backgrounds. This achieves an optimal balance between lightweight parameters and high detection performance.

While EMW-YOLO's GFLOPs (24.44) exceed the baseline YOLOv8n (6.95), this computational expansion is a calculated trade-off for small-object accuracy. This overhead is primarily driven by two structural components: the channel quadrupling (C to 4C) in the EDS module to retain low-level geometric details, and the pixel-wise weight operations within the MSFE neck's asymmetric selective enhancement strategy. Notably, this complexity remains comparable to YOLOv8s (23.6 GFLOPs) and demonstrates the potential for real-time edge deployment [46]. Model compression techniques such as channel pruning will be explored in future work to optimize this overhead.

Furthermore, to validate the effectiveness of WIoU v3, it is compared with CIoU and other mainstream IoU variants (e.g., EIoU, SIoU [47], and Powerful-IoU [48]). Utilizing YOLOv8n as the base model, as presented in Table 2, WIoU v3 and EIoU nearly outperform the other IoU metrics. Although the precision of WIoU v3 is 0.8% lower than that of EIoU, it yields superior performance in recall and mAP. This trade-off aligns with the design philosophy of WIoU v3: its dynamic focusing mechanism prioritizes the gradient contributions of medium-quality samples, which potentially improves recall by detecting more boundary cases, albeit at the expense of a partial loss in precision.

Table 2: Performance comparison among diverse loss functions in terms of IoU.

Loss	P	R	$mAP_{0.5}$	$mAP_{0.5:0.95}$
CIoU	0.46	0.331	0.343	0.201
DIoU	0.462	0.343	0.348	0.204
EIoU	0.462	0.348	0.348	0.203
SIoU	0.45	0.348	0.348	0.204
Powerful-IoU	0.453	0.349	0.349	0.204
WIoU v1	0.443	0.348	0.347	0.204
WIoU v2	0.457	0.344	0.35	0.204
WIoU v3	0.454	0.35	0.353	0.206

4.5 Generalization Experiments

To verify the robustness and cross-domain generalization ability of EMW-YOLO, we carry out experiments on two extra datasets: NWPU VHR-10 and RSOD. The quantitative comparison with multiple SOTA models (e.g., YOLOv11n, RT-DETR) is detailed.

As evidenced in Table 3, EMW-YOLO demonstrates a clear performance lead across most evaluation metrics. Specifically, it achieves a precision of 91.5% and an $mAP_{0.5}$ of 90.0%, representing an absolute improvement of 3.4% and 6.6% over the baseline YOLOv8n, respectively. Notably, in the high-precision localization metric $mAP_{0.5:0.95}$, EMW-YOLO reaches 58.1%, significantly outperforming the latest YOLOv11n and the complexity-heavy Hyper-YOLO. This superior performance is primarily attributed to the EDS module's capability to preserve sub-pixel spatial structures in multi-resolution satellite imagery. Although Hyper-YOLO exhibits a slightly higher recall, EMW-YOLO maintains a much more favorable trade-off between precision and complexity, achieving a superior F1-score of 0.86.

EMW-YOLO achieves the best overall performance, with an $mAP_{0.5}$ of 95.1% and an $mAP_{0.5:0.95}$ of 66.9% on the RSOD dataset. Compared to YOLOv8n, these figures represent substantial gains of 5.6% and 6.3%, respectively. When compared against RT-DETR, which typically excels in global context modeling, EMW-YOLO still provides higher detection accuracy with significantly fewer parameters. The remarkable enhancement in RSOD, particularly for categories with complex shapes and background interference, validates the asymmetric semantic alignment strategy of the MSFE architecture. By dynamically reshaping the receptive field via RFACConv-C2f, EMW-YOLO effectively captures long-range structural dependencies while suppressing harmful gradients from low-quality annotations through the WIoU v3 dynamic focusing mechanism.

In summary, the consistent performance gains across both high-resolution satellite imagery and multi-scale aerial photography indicate that EMW-YOLO generalizes well across different datasets, demonstrating its applicability to complex remote sensing tasks.

Table 3: Comparative results of generalization performance on NWPU VHR-10 and RSOD datasets.

Dataset	Model	P	R	$mAP_{0.5}$	$mAP_{0.5:0.95}$	F1-Score
NWPU VHR-10	RT-DETR	0.79	0.743	0.816	0.509	0.76
	YOLOv8n	0.881	0.78	0.834	0.502	0.82
	YOLOv11n	0.869	0.755	0.825	0.494	0.8
	FFCA-YOLO	0.889	0.816	0.863	0.512	0.85
	Hyper-YOLO	0.896	0.818	0.887	0.545	0.85
	EMW-YOLO	0.915	0.804	0.9	0.581	0.86
RSOD	RT-DETR	0.933	0.909	0.919	0.642	0.92
	YOLOv8n	0.903	0.884	0.895	0.606	0.89
	YOLOv11n	0.907	0.885	0.92	0.615	0.89
	FFCA-YOLO	0.898	0.831	0.891	0.598	0.85
	Hyper-YOLO	0.92	0.89	0.913	0.639	0.91
	EMW-YOLO	0.941	0.919	0.951	0.669	0.93

4.6 Small and Medium Object Detection Experiments

Since conventional mAP metrics fail to intuitively reflect the performance of models on extreme-sized objects, this section analyzes the prediction accuracy between the baseline network and EMW-YOLO on small and medium objects across VisDrone2021, NWPU VHR-10, and RSOD datasets to validate fine-grained feature preservation.

As illustrated in Table 4, EMW-YOLO attains overall more competitive performance relative to the baseline reference model on small as well as intermediate-scale objects. Notably, for the key small-target detection metrics in remote sensing applications, EMW-YOLO achieves significant improvements thanks to its detail-preserving spatial rearrangement and the asymmetric enhancement strategy of MSFE. It also maintains steady performance on medium-sized objects, breaking the scale optimization bottleneck that conventional algorithms struggle with, achieving robust and high-precision detection in complex scenarios.

Table 4: Comparison of small- and medium-scale object detection performance between YOLOv8 and EMW-YOLO on different datasets.

Datasets	Class	AP_S	AR_S	AP_M	AR_M
VisDrone2021	YOLOv8	0.123	0.211	0.335	0.489
	EMW-YOLO	0.192	0.310	0.374	0.535
NWPU VHR-10	YOLOv8	0.246	0.434	0.486	0.580
	EMW-YOLO	0.389	0.576	0.582	0.668
RSOD	YOLOv8	0.395	0.501	0.593	0.659
	EMW-YOLO	0.496	0.590	0.641	0.699

4.7 Ablation Experiment

The VisDrone2021 dataset was adopted to carry out a set of ablation experiments using YOLOv8n as the baseline to verify the effectiveness of the proposed modules. As shown in Table 5, each module's impact on the metrics is demonstrated, with ✓ indicating its application.

Table 5: Ablation study of different components in EMW-YOLO.

Baseline	EDS	MSFE	WIoU v3	P	R	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params	GFLOPs
✓				0.46	0.331	0.343	0.201	2.69M	6.95
✓	✓			0.501	0.383	0.394	0.239	2.78M	12.72
✓		✓		0.468	0.356	0.364	0.218	3.28M	18.68
✓			✓	0.454	0.35	0.353	0.206	2.69M	6.95
✓	✓		✓	0.505	0.382	0.394	0.237	2.78M	12.72
✓		✓	✓	0.462	0.361	0.369	0.22	3.28M	18.68
✓	✓	✓		0.518	0.4	0.417	0.254	3.37M	24.44
✓	✓	✓	✓	0.522	0.4	0.422	0.255	3.37M	24.44

As presented in Table 5, the sequential integration of EDS, MSFE, and WIoU v3 yields steady metric improvements. Using EDS on its own makes the biggest improvement, raising $mAP_{0.5}$ from 34.3% to 39.4%. This shows that detail-preserving spatial rearrangement and cross-dimensional attention protect sub-pixel geometric structures from downsampling. MSFE improves $mAP_{0.5}$ by 2.1%, showing that its asymmetric alignment expands receptive fields and eliminates cross-scale semantic gaps. WIoU v3 provides bounding box regression stability across different configurations with no additional parameters. The combination of EDS and MSFE shows a strong positive synergy. The high-fidelity spatial details preserved by EDS provide a richer and cleaner information base for MSFE’s multi-scale fusion, resulting in an optimal performance of 42.2% $mAP_{0.5}$ for EMW-YOLO—a 7.9% total improvement over the baseline.

Table 6 analyzes how the depth coefficient d in the EDS module affects model performance. Multiple ablation experiments with different d values are implemented based on the YOLOv8n model, where “-” denotes the baseline model. Optimal performance was achieved when d was set to 2. As d was further increased, model performance began to decline despite the continuous rise in parameters and complexity. This phenomenon is mainly ascribed to the expanding number of channels requiring a higher compression ratio in per-pixel convolution, leading to the loss of critical information. Therefore, we selected a depth coefficient of $d = 2$ as the configuration for EDS.

Table 6: Ablation experiments at different EDS depths on VisDrone2021.

d	P	R	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params	GFLOPs
–	0.46	0.331	0.343	0.201	2.69M	6.95
1	0.484	0.372	0.381	0.23	2.75M	10.7
2	0.501	0.383	0.394	0.239	2.78M	12.7
3	0.495	0.368	0.378	0.225	2.82M	14.7

4.8 Qualitative Analysis

To show how well the proposed modules and the network work, the qualitative analysis will be separated into three parts: feature response visualization of the EDS module, MSFE structure detection and EMW-YOLO performance in difficult remote sensing situations.

To evaluate the detail-preserving capability of the EDS module, the Grad-CAM [49] feature responses shown in Fig. 7 are compared. Warm colors indicate strong activation, blue indicates suppression. In the first and third columns the EDS-equipped network achieves sharper, boundary-conforming focus when stacked

against the baseline YOLOv8. This performance difference is clearly reflected in the highly complex scenarios outlined by the red boxes. The baseline model's attention is diffused or lost in these scenarios, whereas the EDS-enhanced network successfully focuses on the fragile targets. The proposed downsampling module is thus superior at preserving geometric details in complex backgrounds.

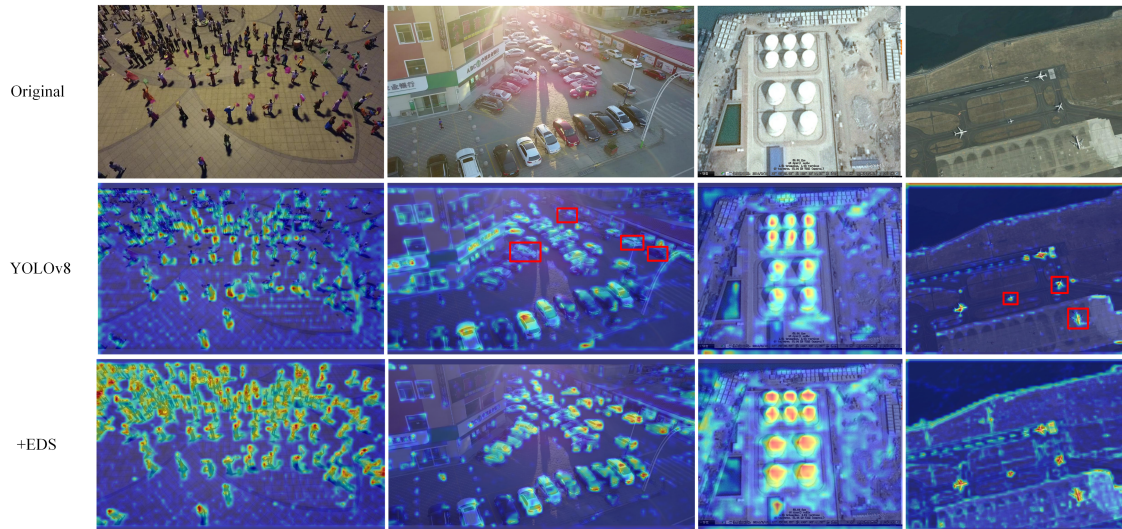


Figure 7: Comparative visualization of feature heatmaps between the baseline YOLOv8 model and the model integrated with our devised EDS unit. Warmer and deeper colors indicate higher feature activation and stronger network attention to the target regions. Red boxes highlight typical regions where the baseline model exhibits weak or diffuse attention compared to the EDS module.

Fig. 8 illustrates the multiscale detection capability of the MSFE architecture. In typical multi-scale scenes like traffic intersections and airport aprons, the baseline YOLOv8 suffers from semantic misalignment, leading to noticeable missed detections of occluded or densely embedded targets. Upon integrating MSFE, these missed objects are successfully recalled. This improvement confirms the efficacy of our asymmetric semantic alignment strategy: dynamically scaling receptive fields via RFACnv-C2f and strictly aligning cross-scale features through cascaded ASFF effectively suppresses semantic noise from deep layers while protecting fragile small-object signals.

Fig. 9 evaluates the engineering applicability of the models under severe imaging conditions. Confronted with glare, deep shadows, and extreme scale variations, YOLOv8 suffers from severe false alarms and missed detections. Conversely, EMW-YOLO exhibits exceptional robustness. By organically integrating the detail-preserving EDS module, the semantically aligned MSFE neck, and the dynamically focused WIoU v3 loss, our framework accurately localizes and classifies targets within complex remote sensing environments.

While EMW-YOLO is generally very robust, there are some situations where it does not perform well. To assess these, we analyzed the failure cases (missed detections) highlighted by yellow circles in the bottom rows of Figs. 8 and 9. In Fig. 8 (2nd column, bottom right), an object at the image edge isn't detected because the target loses its geometric structure. The convolutional receptive field also lags in surrounding context at image boundaries. In Fig. 9 (1st column, bottom left), a vehicle in shadow is missed. In such cases, the object blends with the background as there's almost zero visual gradient. Even with the detail-preserving EDS module, it's hard to extract effective features when the photon information is heavily degraded.

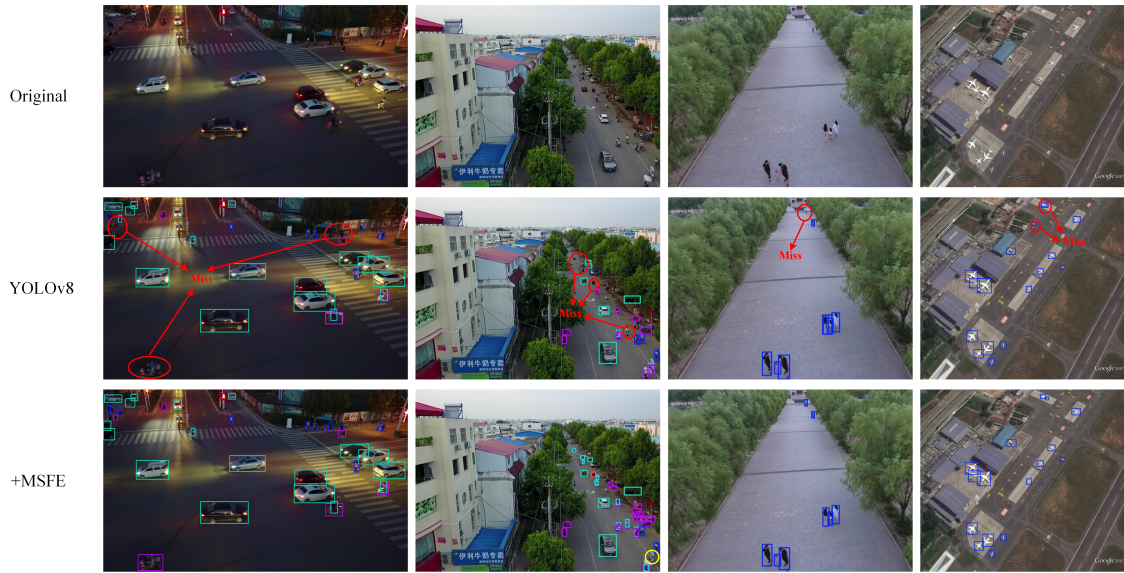


Figure 8: Comparison of detection results between YOLOv8 and the MSFE-enhanced network in handling complex multi-scale objects. Red circles and arrows indicate obvious missed detections of tiny or occluded objects by the baseline YOLOv8.



Figure 9: Comprehensive detection performance comparison between YOLOv8 and EMW-YOLO in extremely complex scenarios. These diagrams include four major complex scenes: large shooting angles (the first column), dense distribution (the second column), low illumination (the third column) and overexposure (the fourth column).

5 Conclusion

Given the challenges inherent in remote sensing imagery, such as easy degradation of small-object details throughout the downsampling stages, semantic inconsistency in multi-scale features, and unstable bounding box regression, this work develops an efficient small object detector called EMW-YOLO improved from YOLOv8n. By integrating an efficient down-sampling strategy, a multi-scale feature enhancement architecture, and the WIoU v3 loss function, the proposed method effectively improves feature representation and regression robustness for small objects. It can be concluded from experimental data that

EMW-YOLO performs considerably better than the baseline model on the VisDrone2021, NWPU VHR-10, and RSOD datasets, achieving mean average precisions of 42.2%, 90.0%, and 95.1%, respectively. To mitigate the computational overhead (24.44 GFLOPs) introduced by these structural enhancements, future research will focus on model quantization and actual deployment on UAV payload platforms via TensorRT acceleration to rigorously validate its real-time inference capabilities in practical engineering applications.

Acknowledgement: The authors gratefully acknowledge the platform support provided by Guangxi Academy of Artificial Intelligence.

Funding Statement: This research was supported by the “Zhi Guo” Action Plan (Guangxi Science and Technology Achievement Transformation Plan, Grant No. Gui Ke ZG2600640082).

Author Contributions: The authors declare their respective contributions to this manuscript as follows: conceptualization and study design: Heng Wang, Long Xu, Zou Zhou; data gathering and model implementation: Heng Wang, Shichao Li, Chuqiao Wang, Yanzhou Feng; analysis and explanation of experimental results: Heng Wang, Shichao Li, Long Xu, Zou Zhou; original draft preparation: Heng Wang; review and editing: Long Xu, Zou Zhou. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The public datasets used in this work, including VisDrone2021, NWPU VHR-10, and RSOD, are accessible via their official published channels. The customized codes and models constructed in this study are available from the corresponding authors upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Han W, Chen J, Wang L, Feng R, Li F, Wu L, et al. Methods for small, weak object detection in optical high-resolution remote sensing images: a survey of advances and challenges. *IEEE Geosci Remote Sens Mag.* 2021;9(4):8–34. doi:10.1109/mgrs.2020.3041450.
2. Zhao X, Yang Z, Zhao H. DCS-YOLOv8: a lightweight context-aware network for small object detection in UAV remote sensing imagery. *Remote Sens.* 2025;17(17):2989. doi:10.3390/rs17172989.
3. Liu J, Xiang J, Jin Y, Liu R, Yan J, Wang L. Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: a survey. *Remote Sens.* 2021;13(21):4387. doi:10.3390/rs13214387.
4. Zhang H, Sun W, Sun C, He R, Zhang Y. HSP-YOLOv8: UAV aerial photography small target detection algorithm. *Drones.* 2024;8(9):453. doi:10.3390/drones8090453.
5. Shi T, Gong J, Hu J, Zhi X, Zhang W, Zhang Y, et al. Feature-enhanced CenterNet for small object detection in remote sensing images. *Remote Sens.* 2022;14(21):5488. doi:10.3390/rs14215488.
6. Fayaz M, Nam J, Dang LM, Song HK, Moon H. Land-cover classification using deep learning with high-resolution remote-sensing imagery. *Appl Sci.* 2024;14(5):1844. doi:10.3390/app14051844.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/cvpr.2016.90.
8. 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)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/cvpr.2016.91.
9. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects. In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8231–40. doi:10.1109/iccv.2019.00832.

10. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Machine learning and knowledge discovery in databases. Cham, Switzerland: Springer; 2023. p. 443–59. doi:10.1007/978-3-031-26409-2_27.
11. He L, Wang M. SliceSamp: a promising downsampling alternative for retaining information in a neural network. Appl Sci. 2023;13(21):11657. doi:10.3390/app132111657.
12. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10778–87. doi:10.1109/cvpr42600.2020.01079.
13. Xu C, Wang J, Yang W, Yu H, Yu L, Xia GS. Detecting tiny objects in aerial images: a normalized Wasserstein distance and a new benchmark. ISPRS J Photogramm Remote Sens. 2022;190(9):79–93. doi:10.1016/j.isprsjprs.2022.06.002.
14. Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans Geosci Remote Sens. 2016;54(12):7405–15. doi:10.1109/TGRS.2016.2601622.
15. Hua W, Chen Q. A survey of small object detection based on deep learning in aerial images. Artif Intell Rev. 2025;58(6):162. doi:10.1007/s10462-025-11150-9.
16. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
17. 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; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6154–62. doi:10.1109/cvpr.2018.00644.
18. Khanam R, Hussain M. What is YOLOv5: a deep look into the internal features of the popular object detector. arXiv:2305.09972. 2023.
19. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2506.08581. 2025.
20. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
21. Zhang Y, Ye M, Zhu G, Liu Y, Guo P, Yan J. FFCA-YOLO for small object detection in remote sens images. IEEE Trans Geosci Remote Sens. 2024;62:1–15. doi:10.1109/tgrs.2024.3363057.
22. Feng Y, Huang J, Du S, Ying S, Yong JH, Li Y, et al. Hyper-YOLO: when visual object detection meets hypergraph computation. IEEE Trans Pattern Anal Mach Intell. 2025;47(4):2388–401. doi:10.1109/tpami.2024.3524377.
23. 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 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17; Montreal, BC, Canada. p. 2778–88. doi:10.1109/ICCVW54120.2021.00312.
24. Wu Q, You X, Huang W, Sun L, Xu Y, Wang X. M2FE-YOLO: multibranch and multilevel feature enhancement network for remote sensing object detection. IEEE Trans Geosci Remote Sens. 2025;63:5646119. doi:10.1109/TGRS.2025.3612212.
25. Qiu Y, Lin Z. Lightweight model LMW-YOLO for small object detection in remote sensing images. Sci Rep. 2026;16(1):11644. doi:10.1038/s41598-026-45055-6.
26. Wu Z, Zhen H, Zhang X, Bai X, Li X. SEMA-YOLO: lightweight small object detection in remote sensing image via shallow-layer enhancement and multi-scale adaptation. Remote Sens. 2025;17(11):1917. doi:10.3390/rs17111917.
27. Liu X, Zhou S, Ma J, Sun Y, Zhang J, Zuo H. DFAS-YOLO: dual feature-aware sampling for small-object detection in remote sensing images. Remote Sens. 2025;17(20):3476. doi:10.3390/rs17203476.
28. Yuan X, Zheng Z, Li Y, Liu X, Liu L, Li X, et al. Strip R-CNN: large strip convolution for remote sensing object detection. arXiv:2501.03775. 2025.
29. Xiao L, Li H, Yao S, Hou C. EdgeFusionNet: edge information-guided small object detection for remote sensing images. Pattern Recognit. 2026;179(6):113621. doi:10.1016/j.patcog.2026.113621.
30. Shi Z, Hu J, Ren J, Ye H, Yuan X, Ouyang Y, et al. HS-FPN: high frequency and spatial perception FPN for tiny object detection. AAAI. 2025;39(7):6896–904. doi:10.1609/aaai.v39i7.32740.

31. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA. p. 3559–68. doi:10.1109/wacv48630.2021.00360.
32. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 658–66. doi:10.1109/CVPR.2019.00075.
33. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. arXiv:1911.08287. 2019.
34. Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans Cybern.* 2022;52(8):8574–86. doi:10.1109/tcyb.2021.3095305.
35. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing.* 2022;506(9):146–57. doi:10.1016/j.neucom.2022.07.042.
36. Shi W, Caballero J, Huszar F, Totz J, Aitken AP, Bishop R, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 1874–83. doi:10.1109/cvpr.2016.207.
37. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA. p. 3138–47. doi:10.1109/wacv48630.2021.00318.
38. Lin TY, 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); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/CVPR.2017.106.
39. Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection. arXiv:1911.09516. 2019.
40. Zhang X, Liu C, Yang D, Song T, Ye Y, Li K, et al. RFACConv: innovating spatial attention and standard convolutional operation. arXiv:2304.03119. 2023.
41. Tong Z, Chen Y, Xu Z, Yu R. Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv:2301.10051. 2023.
42. Cao Y, He Z, Wang L, Wang W, Yuan Y, Zhang D, et al. VisDrone-DET2021: the vision meets drone object detection challenge results. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17; Montreal, BC, Canada. p. 2847–54. doi:10.1109/iccvw54120.2021.00319.
43. 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(5):2486–98. doi:10.1109/tgrs.2016.2645610.
44. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 16965–74. doi:10.1109/CVPR52733.2024.01605.
45. Huang S, Lu Z, Cun X, Yu Y, Zhou X, Shen X. DEIM: DETR with improved matching for fast convergence. arXiv:2412.04234. 2025.
46. Rey L, Bernardos AM, Dobrzycki AD, Carramiñana D, Bergesio L, Besada JA, et al. A performance analysis of you only look once models for deployment on constrained computational edge devices in drone applications. *Electronics.* 2025;14(3):638. doi:10.3390/electronics14030638.
47. Gevorgyan Z. SIoU loss: more powerful learning for bounding box regression. arXiv:2205.12740. 2022.
48. Liu C, Wang K, Li Q, Zhao F, Zhao K, Ma H. Powerful-IoU: more straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. *Neural Netw.* 2024;170(2):276–84. doi:10.1016/j.neunet.2023.11.041.
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis.* 2020;128(2):336–59. doi:10.1007/s11263-019-01228-7.