scientific reports

Check for updates

OPEN

# Improved YOLOv9-based remote sensing image detection method

Ming Chen, Chunping Wang✉, Ying Yu, Yirui Li, Yifan Li & Xianglong Chen

**Object detection in remote sensing imagery presents significant challenges due to complex backgrounds, the prevalence of small objects, and high instance density, all of which hinder both detection accuracy and computational efficiency. To address these issues, we propose an enhanced version of the YOLOv9 architecture specifically designed for remote sensing image analysis. Our model incorporates several key innovations: a multi-scale feature integration module (C3) that jointly captures fine-grained details and high-level semantic information; a channel attention mechanism (Squeeze-and-Excitation module) that adaptively highlights informative features while suppressing irrelevant background regions; an additional detection head (P2) aimed at improving small object recognition; and the Generalized Intersection over Union (GIoU) loss for more accurate bounding box regression and faster training convergence. Extensive experiments on the SIMD dataset demonstrate that our model achieves state-of-the-art performance, with 86.6% mAP@0.5 and 71.5% mAP@0.5–0.95, while operating at 84.0 FPS—significantly outperforming the baseline YOLOv9. Moreover, the model reduces the number of parameters by 21.2%, highlighting its efficiency. These advancements position our model as a highly effective solution for real-world remote sensing applications such as environmental monitoring, urban planning, and military reconnaissance.**

**Keywords** YOLOv9, Remote sensing imagery, Object detection, Attention mechanism, Multi-scale feature learning, GIoU loss

Object detection in remote sensing imagery plays a vital role in a variety of applications, including urban planning, environmental monitoring, and defense reconnaissance[1,2]. Despite its importance, the task remains particularly challenging due to the complex backgrounds, high object density, and frequent presence of small targets typical in aerial and satellite images. These factors often lead to issues such as missed detections, false alarms, and inaccurate localization, especially when general-purpose detection models are applied. The YOLO series has gained widespread adoption in real-time object detection due to its favorable trade-off between efficiency and accuracy[3–7]. As the latest iteration, YOLOv9 introduces advanced designs such as CSPNet and programmable gradient information, achieving state-of-the-art results on standard benchmarks[8]. However, when directly applied to remote sensing imagery, YOLOv9 still exhibits notable limitations, particularly in detecting small and densely distributed objects[9].

The inherent characteristics of remote sensing images—such as large spatial coverage, multi-scale objects, and the low resolution of small targets—pose unique challenges that are not adequately addressed by existing YOLO-based models. Several studies have attempted to adapt YOLO architectures for remote sensing tasks. For example, Lv et al.[10] incorporated a Squeeze-and-Excitation (SE) module to improve feature representation, though at the expense of increased computational complexity. Hou et al.[11] proposed R-YOLO, which integrates a transformer and an attention mechanism for oriented object detection, but its real-time performance remains limited. Li et al.[12] introduced RSI-YOLO with a dedicated small-object detection layer, yet the model still faces difficulties in discriminating similar object categories and generalizing across scenarios.

More recent approaches, such as AF-SSD[13] and RGTGAN[14], have focused on feature fusion and super-resolution techniques, respectively. However, these methods either impose constraints on input sizes or incur high computational costs. Similarly, MPE-YOLO[9] and AGW-YOLOv8[15] have made progress in detecting small objects, but often at the cost of inference speed or under specific environmental assumptions. Despite these efforts, a significant gap remains in developing a detection model that effectively balances accuracy, speed, and model complexity for remote sensing imagery—particularly in handling small and densely packed objects under real-world operational constraints.

To address these challenges, we propose a tailored version of YOLOv9 specifically optimized for remote sensing image detection. The main contributions of this work are as follows:

School of Information and Intelligent Engineering, University of Sanya, Sanya 572022, China. ✉email: chunpwang@163.com

- We introduce a multi-scale feature integration module (C3) into the backbone network to enhance the model's capacity for capturing both low-level details and high-level semantics, thereby improving feature representation across different scales.
- We incorporate a channel attention mechanism (SE module) into the neck network to adaptively recalibrate feature responses, enabling the model to focus on informative regions while suppressing background noise.
- We add a dedicated small-object detection head (P2) that leverages higher-resolution feature maps to improve the detection of small and densely distributed objects.
- We replace the original CIoU loss with the Generalized IoU (GIoU) loss to enhance bounding box regression performance—especially in cases of low overlap—and accelerate model convergence.

Extensive experiments on the SIMD dataset demonstrate that our model achieves state-of-the-art performance, with 86.6% mAP@0.5 and 71.5% mAP@0.5–0.95, while maintaining a real-time inference speed of 84.0 FPS. Additionally, the model reduces the parameter count by 21.2% compared to the original YOLOv9, highlighting its efficiency and practical value.

## Related work

### Evolution of object detection techniques

Object detection methodologies have evolved substantially, transitioning from traditional feature-engineered approaches to modern deep learning-based frameworks. Early techniques predominantly relied on handcrafted features and sliding-window detection paradigms—exemplified by the combination of Haar-like features with AdaBoost for face detection, and Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVM) for pedestrian detection. While effective in constrained scenarios, these methods suffered from high computational costs and limited generalization in complex environments, hindering their practical deployment[16].

The rise of Convolutional Neural Networks (CNNs) marked a turning point, establishing deep learning as the dominant paradigm in object detection. The R-CNN family (R-CNN[17], Fast R-CNN[18], Faster R-CNN[19] introduced a region proposal-based framework combined with deep feature extraction, significantly boosting detection accuracy. Nevertheless, these two-stage detectors incurred considerable computational overhead, limiting their applicability in real-time systems[20]. This motivated the development of single-shot detectors such as SSD (Single Shot MultiBox Detector) and the YOLO (You Only Look Once) series, which process entire images in one forward pass, effectively balancing speed and accuracy while dramatically reducing inference time[21].

As deep learning advanced, subsequent research has focused on further enhancing the robustness and efficiency of detection models. For instance, the integration of multi-scale feature fusion and attention mechanisms has substantially improved model capability in handling complex scenes and objects of varying scales (Guo et al., 2020). More recently, efforts have shifted toward lightweight and efficient architectures—including real-time detectors based on streamlined deep networks—to accommodate the computational constraints of embedded and mobile platforms[22].

In summary, object detection has achieved remarkable progress through continuous deep learning innovations, particularly in harmonizing real-time inference with high accuracy. Still, critical challenges remain in further optimizing computational efficiency and strengthening model performance under complex and varied environmental conditions, pointing to a key direction for future research[23].

### YOLOv9 model

Since its introduction in 2015, the YOLO (You Only Look Once) series has garnered considerable recognition for its efficient single-stage detection framework. YOLOv1 pioneered the reformulation of object detection as a unified regression problem, enabling end-to-end prediction of bounding boxes and class labels through a single network pass[24]. Subsequent iterations, including YOLOv2 and YOLOv3, further refined this paradigm, leading to marked improvements in both detection accuracy and inference speed[25,26].

As a recent successor in this lineage, YOLOv9[27] integrates a suite of advanced techniques to further elevate detection performance. It incorporates a deeper network architecture augmented with efficient feature extraction components such as Cross Stage Partial Network (CSPNet) and Spatial Pyramid Pooling (SPP). These modules are designed to enrich feature representation while maintaining computational efficiency[28,29]. In addition, YOLOv9 employs a refined loss function that enhances bounding box regression accuracy, yielding state-of-the-art results on multiple public benchmarks. These advantages are particularly pronounced in scenarios involving complex backgrounds and small object detection[30].

Despite these advancements, the YOLOv9 model still exhibits certain limitations when applied to specialized domains such as remote sensing imagery. Specifically, its computational overhead remains high when processing large-scale, high-resolution images, and it shows a tendency toward both missed detections and false alarms in densely packed object scenarios. These shortcomings underscore the need for further architectural and optimization efforts—particularly in boosting computational efficiency and detection precision under demanding operational conditions[31,32].

### Improvement methods for the YOLO series models

Based on the YOLOv9 framework, researchers have proposed various enhancement strategies to further advance its detection capabilities. These improvements primarily concentrate on architectural refinement, feature fusion mechanisms, loss function redesign, and innovations in data augmentation techniques.

In terms of network architecture, several studies have introduced lightweight modules to reduce parameter count and computational complexity, thereby accelerating inference. For instance, Shen et al.[33] developed DS-YOLOv8, which integrates deformable convolutions and a self-calibrating attention mechanism to improve

multi-scale and small object detection while preserving real-time performance. Similarly, Liu et al.[34] proposed NRT-YOLO, incorporating a nested residual Transformer module into YOLOv5 to enhance small object recognition in remote sensing imagery.

Regarding feature fusion strategies, Xie et al.[35] introduced YOLO-RS, which employs an adaptive spatial feature fusion structure to better integrate multi-scale feature information, leading to improved detection accuracy in complex backgrounds. Ma et al.[36] presented YOLO-UAV, an optimized variant of YOLOv5 that refines both network structure and feature fusion strategy to boost detection performance in UAV-captured imagery.

Loss function optimization has also been a key focus in enhancing YOLO performance. Su et al.[37] proposed a multi-scale strip convolutional attention mechanism coupled with a multi-scale feature fusion strategy, significantly improving detection across varying object scales in remote sensing contexts. In a similar vein, Peng et al.[38] designed AMFLW-YOLO, a lightweight detection model that incorporates a coordinate attention mechanism and a bidirectional feature pyramid network to strengthen feature extraction and detection precision. Zhang et al.[39] further contributed a lightweight improved YOLOv5 model, which introduces an asymmetric detection head and a novel C3CA module to enhance both speed and accuracy in remote sensing object detection.

While these enhancements yield notable performance gains in specific scenarios, they are not without limitations in practical deployment. Lightweight models, for example, often trade off detection accuracy for speed, while complex feature fusion strategies may improve detection capability at the cost of increased computational overhead, potentially compromising real-time performance. Therefore, a careful balance among these factors must be struck according to the requirements of the target application.

## Improved methods

In deep learning-based object detection, models learn a hierarchy of feature representations, progressing from low-level structures such as edges and textures to high-level semantic concepts that constitute object parts. These internally learned patterns enable the model to perform bounding box localization and class prediction, without relying on explicit extraction of predefined object features. To enhance the performance of YOLOv9 in remote sensing object detection, this paper introduces a set of targeted improvements, including structural optimizations, loss function refinements, and the integration of novel modules. The subsequent sections detail the design, implementation, and functional role of each proposed enhancement within the overall architecture.

## Strategies for improving YOLOv9

This paper presents a series of enhancements to the YOLOv9 architecture, as illustrated in Fig. 1. In the backbone network, the C3 module has been integrated to effectively aggregate multi-scale features spanning from low-level edges and textures to high-level semantic information, thereby significantly improving the overall feature representation capability. Within the neck network, the Squeeze-and-Excitation (SE) module is introduced to enhance the model's attention to salient regions in the input images, which increases its sensitivity to critical information and boosts detection performance. Additionally, a P2 detection head is incorporated, extending the model's feature hierarchy to better capture fine-grained details of small objects, thus ensuring accurate localization and recognition of small-scale targets in complex scenes. Furthermore, the Generalized Intersection over Union (GIoU) loss is adopted as the localization objective, which not only ensures spatial consistency and precision of the predicted bounding boxes but also accelerates the convergence process during training.

## C3 module

The C3 module[40] enhances the original Cross-Stage Partial (CSP) design in YOLOv9 through a partition-and-aggregation strategy. The input feature map is split into two branches: one undergoes a series of convolutional layers for richer feature learning, while the other retains the original information via a shortcut connection. These two branches are subsequently concatenated at the output. This structure offers dual benefits: firstly, it mitigates computational redundancy by processing only a portion of the features through the computationally intensive convolutional stack. Secondly, it enriches the feature representations by integrating shallow, fine-grained details with deeper, more semantic information. In our improved YOLOv9 architecture, the C3 module is extensively deployed across multiple levels of the backbone network, substantially boosting the efficiency of feature extraction.

As illustrated in Fig. 2, the C3 module employs grouped convolutions followed by a weighted fusion mechanism. It segments the input features into independent groups, each processed by separate convolutional paths. The resulting features are then fused at the output stage. This design not only strengthens the network's capacity for modeling complex, nonlinear relationships but also significantly lowers computational overhead. Consequently, the enhanced YOLOv9 model achieves comparable or even superior detection accuracy while operating with greater computational economy.

### *SE module*

The Squeeze-and-Excitation (SE) module is a channel-wise attention mechanism designed to enhance the representational capacity of convolutional neural networks by adaptively recalibrating channel-wise feature responses[41–43]. By explicitly modeling interdependencies among feature channels, the SE module enables the network to emphasize informative features and suppress less useful ones, thereby improving feature discriminability in complex or multi-target scenarios.

In the improved YOLOv9 model proposed in this paper, SE modules are embedded into multiple feature extraction layers. Through channel-wise weighting of feature maps, the network dynamically prioritizes task-

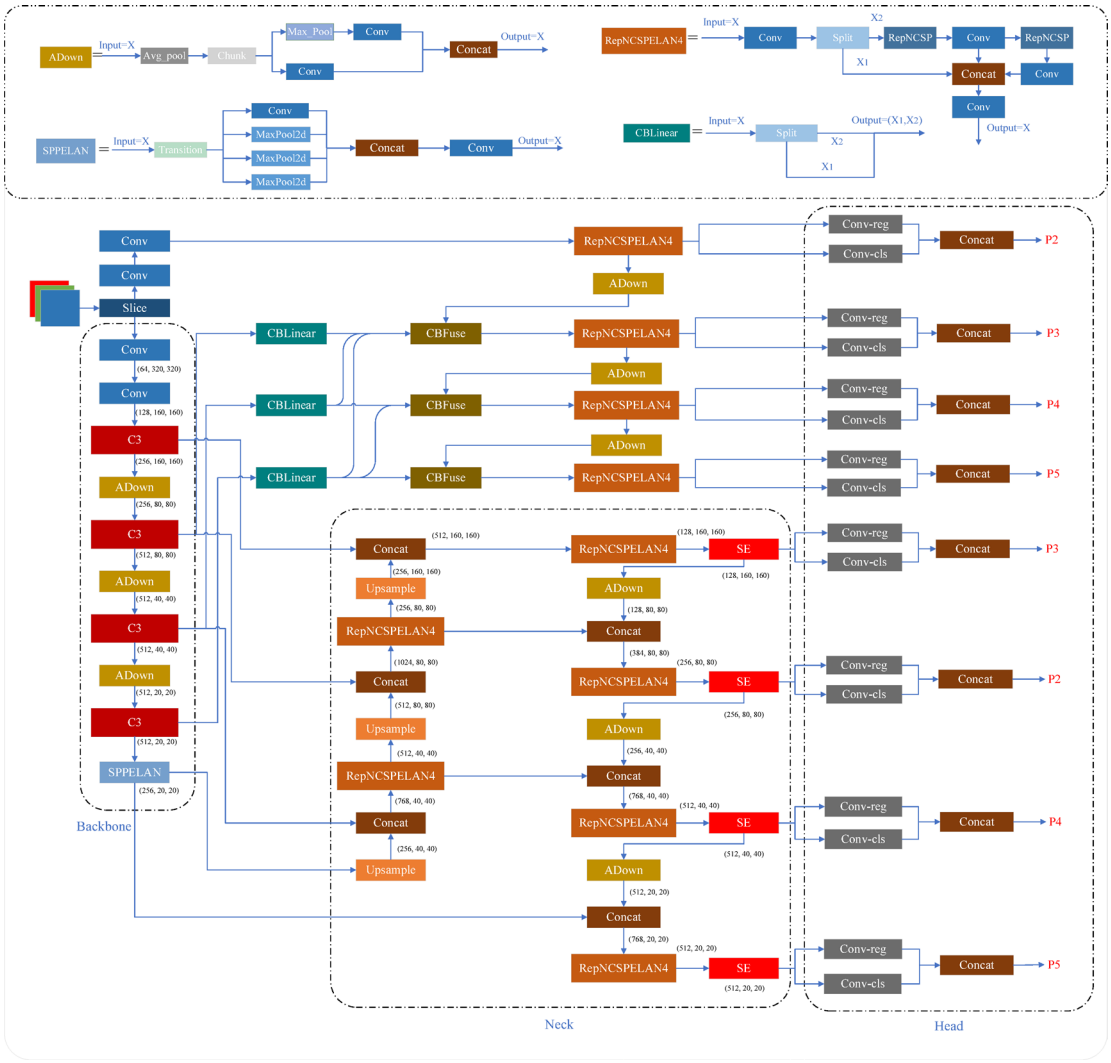


**Fig. 1**. Enhanced YOLOv9 network architecture.

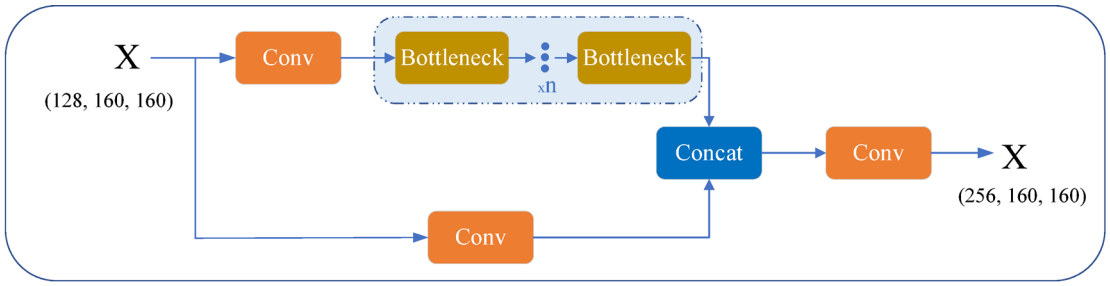


**Fig. 2**. C3 Module.

relevant features during convolution, significantly boosting detection accuracy and robustness—particularly beneficial in dense remote sensing imagery.

As illustrated in Fig. 3, the SE module operates in two sequential steps: Squeeze: Global Average Pooling (GAP) is applied to aggregate spatial information of each feature map into a channel-wise descriptor, yielding a compressed $1 \times 1 \times C$ representation that encapsulates global contextual information. This operation is formulated in Eq. (1):

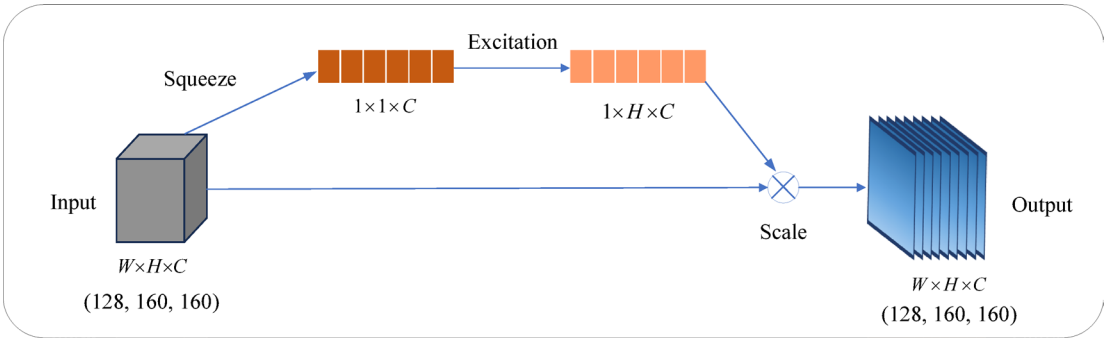


**Fig. 3**. SE Module.

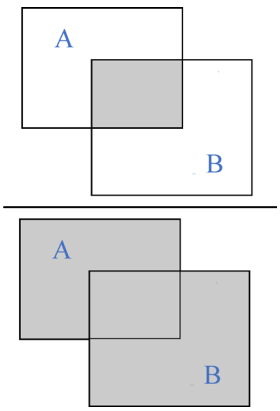


**Fig. 4**. IoU diagram.

$$z_k = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} x_k(i,j), k = 1, 2, \cdots, C, \tag{1}$$

In Eq. (1), $x$ represents the feature map obtained after convolution, $C$ is the number of channels in $x$ and $W \times H$ is the spatial dimension of $x$.

Excitation: The compressed feature descriptor subsequently undergoes a nonlinear transformation via a two-layer bottleneck structure. This process employs two fully connected (FC) layers: the first acts as a dimensionality reduction layer, while the second restores the channel dimensionality to its original size. A ReLU activation function is incorporated between these FC layers to introduce nonlinearity, culminating in a Sigmoid activation that generates normalized gating weights S for each channel. These weights quantify the relative importance of individual channels within the final feature representation, as formalized in Eq. (2):

$$s = \sigma[W_2 \delta(W_1 z)], \tag{2}$$

In Eq. (2), $\delta$represents the nonlinear activation function ReLU, $W_1$and $W_2$are the parameters of the two FC layers, and $\sigma$is the Sigmoid function.

Reweight: The channel-wise weights S derived from the Excitation step are applied to the original feature maps via channel-wise multiplication. This operation adaptively recalibrates the feature responses, effectively enhancing the discriminative power of informative channels while suppressing less relevant ones. The final output of the SE module is thus obtained, as formulated in Eq. (3):

$$\tilde{x}_k = s_k \cdot x_k, k = 1, 2, \cdots, C, \tag{3}$$

In Eq. (3), $s_k$ denotes the weight values, $x_k$ denotes the feature maps, and $\tilde{x}_k$ denotes the reweighted feature maps.

## Small object detection head

Accurately detecting small objects remains a significant challenge in remote sensing imagery. In this work, we define a "small object" quantitatively as one with a bounding box area of less than 32 × 32 pixels in the input image, consistent with established benchmarks in the field. Conventional YOLO models primarily depend on

deeper feature maps (P3, P4, P5) for detection. While these layers capture rich semantic information, their substantially reduced spatial resolution often leads to the loss or severe degradation of subtle pixel-level cues associated with small objects during progressive downsampling, thereby limiting detection performance.

To overcome this limitation, we introduce an additional P2 detection head that utilizes higher-resolution feature maps from earlier network stages. Specifically, the P2 head operates on features from the second stage of the backbone, which retains twice the spatial resolution of the P3 layer. This enhanced resolution is critical for small object detection, as it preserves fine-grained texture and structural details. For an object below the $32 \times 32$-pixel threshold, the P2 feature map provides a more substantial and discriminative feature set—whereas in coarser layers (P3–P5), the same object may span only a few pixels, making reliable detection difficult. As shown in Fig. 1, integrating the P2 detection head into the feature pyramid network enables the model to leverage these preserved high-resolution details, significantly improving the accurate localization and recognition of even the smallest objects in complex remote sensing scenes.

## GIoU loss function

In the YOLO series of models, bounding box regression is generally achieved through the IoU (Intersection over Union) loss function. However, the conventional IoU loss fails to deliver effective gradient signals when the predicted bounding box and the ground truth box do not overlap, thereby impairing model convergence and regression accuracy. To mitigate this limitation, this study employs an enhanced GIoU (Generalized Intersection over Union) loss function. The GIoU loss extends the standard IoU by incorporating the smallest enclosing rectangle that covers both the ground truth and the predicted box, thereby introducing a more informative distance measure. In contrast to traditional IoU, GIoU is capable of supplying meaningful gradient information even in cases of non-overlapping boxes, which accelerates convergence and substantially improves the precision of bounding box regression[44,45]. By integrating the GIoU loss, the refined YOLOv9 model achieves more accurate object localization in high-resolution remote sensing imagery, exhibiting particularly robust performance in complex scenarios. The GIoU metric is formally defined as follows:

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \tag{4}$$

In the given equation, A denotes the area of the predicted bounding box, while B corresponds to the area of the ground truth bounding box.

The schematic illustration of the IoU computation is presented in Fig. 4. Mathematically, IoU is defined as the ratio of the area of overlap between the two bounding boxes to the area of their union.

Figure 5 illustrates the schematic diagram of GIoU. As an enhanced variant of IoU, GIoU is designed to overcome its key limitation by introducing the concept of the minimal enclosing rectangle (C), which mitigates optimization stagnation in non-overlapping cases. Specifically, a penalty term is incorporated by subtracting from the original IoU value—a term that quantifies the proportion of the empty area within the minimal enclosing rectangle not occupied by either bounding box. In contrast to IoU, which is bounded within [0, 1], GIoU exhibits a symmetric value range of [−1, 1]. It attains a maximum of 1 when the two boxes perfectly coincide and approaches a minimum of −1 as they become completely disjoint and infinitely distant. This property renders GIoU a more robust and informative geometric measure. The principal advantage of GIoU lies in its capacity to account not only for the overlapping region but also for non-overlapping areas, thereby enabling effective gradient propagation even when the predicted and ground-truth boxes do not intersect. This leads to a more accurate and stable characterization of spatial alignment, contributing to a more reliable and convergent loss function. The GIoU metric is formally defined as follows:

$$GIoU = IoU - \frac{|C/\ (A \cup B)|}{|C|}, \tag{5}$$

In the equation, A, B, and C denote the areas of the predicted bounding box, the ground truth bounding box, and the smallest enclosing rectangle that contains both, respectively.

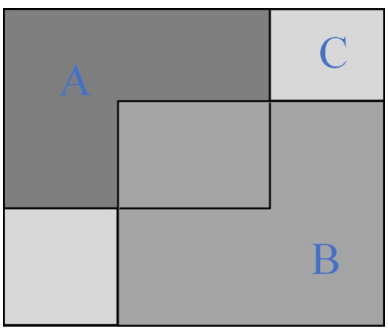


**Fig. 5**. GIoU diagram.

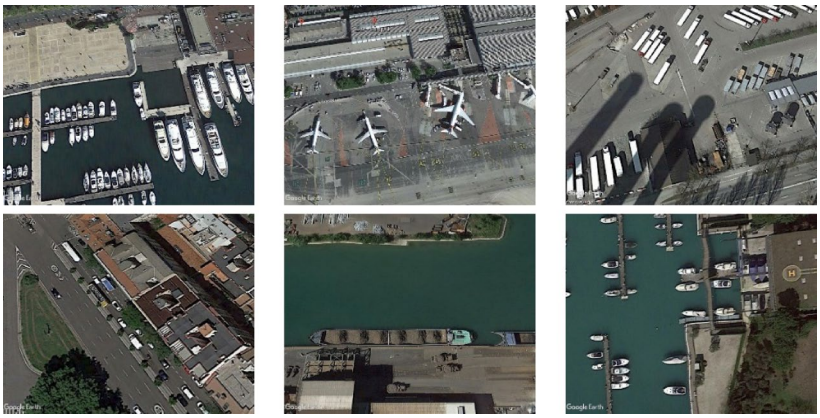

**Fig. 6**. Sample Images from SIMD.

| Algorithms | Precision/% | Recall/% | mAP@0.5/% | mAP@0.5–0.95/% | FPS | GFLOPS | Params (M) |
|---|---|---|---|---|---|---|---|
| SSD | 81.18 | 77.27 | 81.99 | — | — | — | — |
| Faster R-CNN | 64.55 | 81.38 | 76.76 | — | — | — | — |
| YOLOv5s | 82.2 | 78.7 | 83.4 | 66.3 | 294.1 | 15.9 | 7.05 |
| YOLOv5l | 83.6 | 81.5 | 83.6 | 68.2 | 212.7 | 107.9 | 46.18 |
| YOLOv7 | 84.5 | 82.0 | 84.8 | 69.5 | 103.0 | 103.4 | 36.55 |
| YOLOv8s | 80.7 | 81.3 | 83.3 | 68.1 | 227.2 | 28.5 | 11.13 |
| YOLOv8l | 82.7 | 80.4 | 84.9 | 70.1 | 163.9 | 164.9 | 43.61 |
| YOLOv9 | 83.9 | 83.0 | 85.7 | 70.8 | 58.4 | 236.8 | 50.73 |
| Ours | 81.5 | 84.1 | 86.6 | 71.5 | 84.0 | 304.0 | 39.97 |

**Table 1**. Algorithm comparison experimental Results.

## Experimental setup

In this section, we evaluate the performance of the enhanced YOLOv9 model on remote sensing object detection tasks and assess its practical applicability. The experimental setup is detailed below, encompassing the dataset description, implementation environment, hyperparameter configuration, and evaluation metrics.

### Dataset

Experiments in this study are conducted using the SIMD dataset[46], an open-source benchmark specifically designed for evaluating object detection models in remote sensing imagery. Characterized by high-resolution images and a diverse set of object categories, the dataset enables robust assessment of model performance under complex backgrounds and multi-object scenarios.

The SIMD dataset integrates multiple publicly available remote sensing image sources, with all images standardized to a resolution of $1024 \times 768$ pixels. It encompasses a wide range of realistic scenes—including urban, rural, and airport environments—ensuring high representativeness for practical applications. The dataset comprises 15 object categories: Car, Truck, Van, Long Vehicle, Bus, Airliner, Propeller Aircraft, Trainer Aircraft, Chartered Aircraft, Fighter Aircraft, Others, Stair Truck, Pushback Truck, Helicopter, and Boat. In total, 45,096 object instances are annotated, offering both diversity and a substantial level of detection challenge. Example images from the dataset are provided in Fig. 6.

For model training and validation, the dataset is partitioned into training and validation sets at an 8:2 ratio. Specifically, the training set contains 4,000 images, and the validation set contains 1,000 images. This split ensures adequate data for effective model learning and reliable performance evaluation. Furthermore, the relatively balanced distribution of object categories across the dataset helps alleviate potential bias caused by class imbalance during training.

### Experimental environment

In the experimental setup, a high-performance computing platform was employed to ensure efficient training and testing of the improved YOLOv9 model. For the hardware configuration, an NVIDIA RTX 4090 GPU with 24 GB of VRAM was utilized, enabling high-speed processing of large-scale datasets and complex computational tasks. The system was equipped with an AMD EPYC 7T83 central processing unit (CPU), featuring 64 physical cores and 22 virtual CPUs (vCPUs), to deliver robust performance for data preprocessing and other parallel computations. A system memory of 30 GB was allocated to mitigate potential bottlenecks during large-scale data operations. Furthermore, all experimental data were stored on a 50 GB solid-state drive (SSD) to facilitate rapid data access and efficient read/write operations throughout the training and evaluation processes.

| Algorithms Category | SSD | Faster *R*-CNN | YOLOv5s | YOLOv5l | YOLOv7 | YOLOv8s | YOLOv8l | YOLOv9 | Ours |
|---|---|---|---|---|---|---|---|---|---|
| car | 91.2 | 84.5 | 93.4 | 91.7 | 93.7 | 93.7 | 93.7 | 94.0 | 94.8 |
| Truck | 77.1 | 74.8 | 84.9 | 83.7 | 86.0 | 83.9 | 85.7 | 87.2 | 88.8 |
| Van | 83.7 | 77.8 | 84.3 | 82.7 | 84.8 | 83.7 | 84.7 | 85.6 | 87.5 |
| Long Vehicle | 75.3 | 74.2 | 86.8 | 84.4 | 84.8 | 85.4 | 85.5 | 82.0 | 87.2 |
| Bus | 88.9 | 86.5 | 93.6 | 92.6 | 93.3 | 90.8 | 92.8 | 93.8 | 93.8 |
| Airliner | 97.9 | 98.2 | 98.3 | 98.3 | 98.4 | 97.7 | 98.5 | 98.9 | 98.8 |
| Propeller Aircraft | 94.9 | 94.1 | 98.2 | 96.7 | 96.4 | 96.7 | 97.1 | 97.1 | 97.2 |
| Trainer Aircraft | 95.2 | 91.8 | 97.8 | 97.6 | 97.6 | 96.4 | 96.9 | 97.3 | 98.2 |
| Chartered Aircraft | 96.1 | 96.4 | 93.5 | 94.4 | 95.7 | 95.6 | 94.6 | 95.8 | 94.0 |
| Fighter Aircraft | 100.0 | 97.0 | 90.8 | 94.5 | 90.9 | 99.5 | 99.5 | 99.5 | 99.5 |
| Others | 39.2 | 31.5 | 33.6 | 34.4 | 36.6 | 36.8 | 31.9 | 36.2 | 33.3 |
| Stair Truck | 52.5 | 45.1 | 51.1 | 51.5 | 59.3 | 47.5 | 56.4 | 61.2 | 62.1 |
| Pushback Truck | 42.7 | 26.1 | 51.3 | 59.0 | 67.2 | 51.2 | 66.8 | 71.3 | 72.9 |
| Helicopter | 100.0 | 81.8 | 94.5 | 93.9 | 88.3 | 92.2 | 92.1 | 88.0 | 91.6 |
| Boat | 94.6 | 91.0 | 98.5 | 98.2 | 98.9 | 97.9 | 97.9 | 98.4 | 98.4 |

**Table 2**. Comparison of AP results for different Models.

| Number | Experiments | Precision/% | Recall/% | .mAP@0.5–0.95/% | FPS | GFLOPS | Params (M) |
|---|---|---|---|---|---|---|---|
| 1 | YOLOv9 | 83.9 | 83.0 | 70.8 | 58.4 | 236.8 | 50.73 |
| 2 | YOLOv9 + C3 | 83.7 | 81.4 | 70.6 | 97.0 | 263.0 | 49.84 |
| 3 | YOLOv9 + C3 + P2 | 82.9 | 82.4 | 70.9 | 86.9 | 304.0 | 39.93 |
| 4 | YOLOv9 + C3 + P2 + GIou | 81.6 | 83.5 | 70.8 | 85.4 | 304.0 | 39.93 |
| 5 | YOLOv9 + C3 + P2 + GIou + SE | 81.5 | 84.1 | 71.5 | 84.0 | 304.0 | 39.97 |

**Table 3**. Impact of different improvement steps on the performance of the YOLOv9 Model.


**Fig. 7**. Comparison of Heatmap Effects. (**a**) Network attention areas and intensity in the model detection without the SE module; (**b**) Network attention areas and intensity in the model detection with the SE module.

Regarding the software environment, the experiments were conducted on a Windows operating system, with an Ubuntu 20.04 virtual instance deployed for its stability and compatibility. The implementation and training of the model were carried out using the PyTorch 1.11.0 deep learning framework, with Python 3.8 as the primary programming language—a combination widely adopted in the research community for its flexibility and extensive library support. To maximize GPU acceleration, CUDA 11.3 was integrated to harness the parallel computing capabilities of the hardware, substantially boosting the model training efficiency.

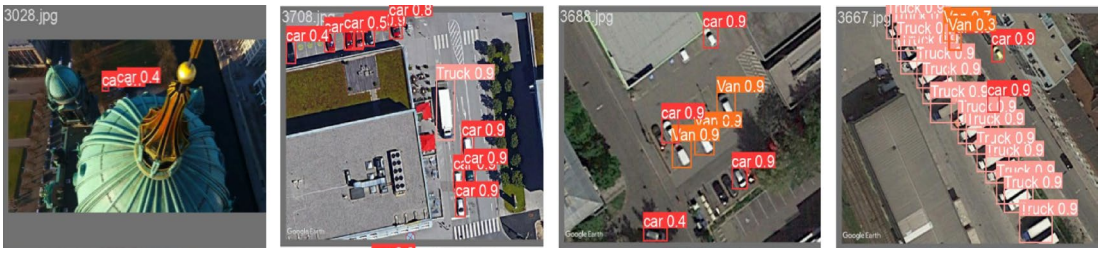


**Fig. 8**. YOLOv5s Algorithm.

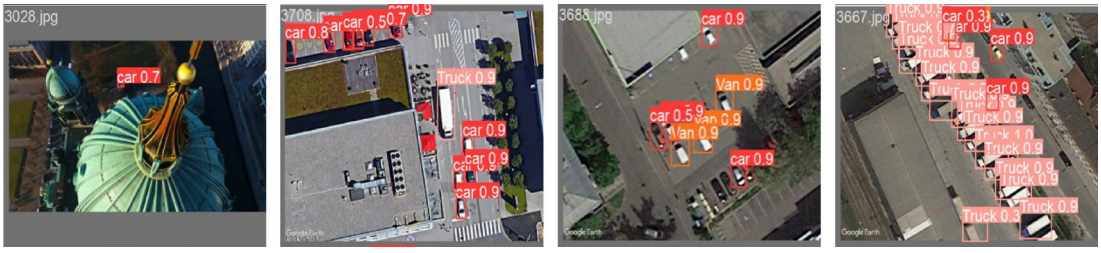


**Fig. 9**. YOLOv5l Algorithm.


**Fig. 10**. YOLOv7 Algorithm.


**Fig. 11**. YOLOv8s Algorithm.


**Fig. 12**. YOLOv8l Algorithm.

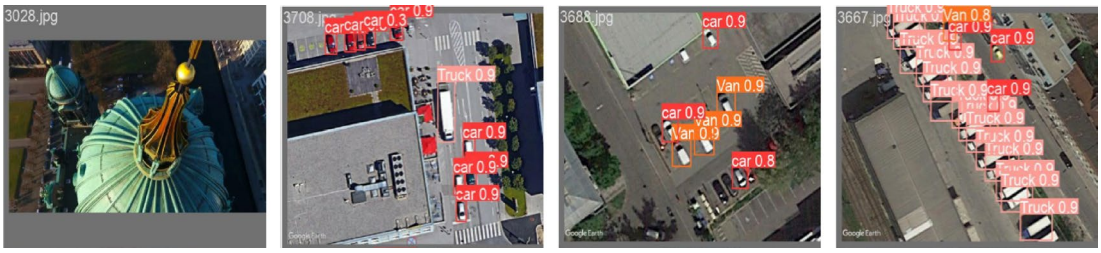


**Fig. 13**. YOLOv9 Algorithm.

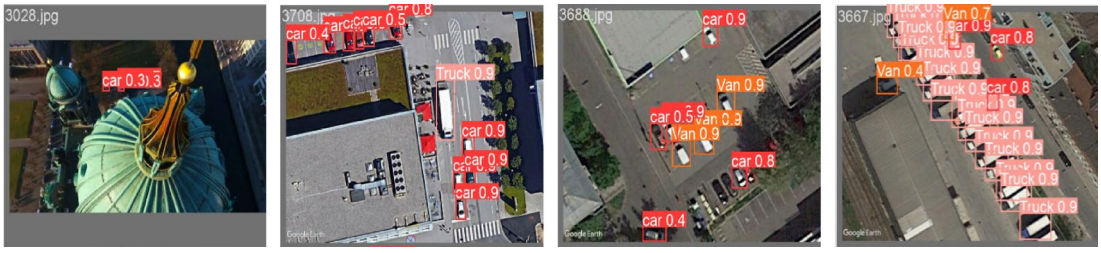


**Fig. 14**. Improved YOLOv9 algorithm proposed in this paper.

### Hyperparameter settings

During model training, the input image resolution was uniformly set to 640 × 640 pixels to achieve a balance between feature extraction capability and computational efficiency. A batch size of 4 was adopted to fully exploit GPU parallel processing, thereby accelerating the training process and promoting convergence. The model was trained over 300 epochs with an initial learning rate of 0.01 to ensure stable convergence during the initial training phase. To enhance data loading efficiency, 8 worker threads were employed for parallel data preprocessing. In terms of optimizer configuration, a momentum factor of 0.937 was applied to expedite convergence and mitigate oscillations, while a weight decay coefficient of 0.0005 was introduced to regularize the model, thereby alleviating overfitting and improving generalization performance. All the aforementioned hyperparameters were meticulously tuned to ensure training stability and maximize the overall model performance.

### Evaluation metrics

In this paper, the evaluation metrics used for the experiments include Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP)[47]. Additionally, the number of Parameters is also considered. The calculation expressions for these metrics are as follows:

$$\text{Precision} = \frac{T\text{p}}{T\text{p} + F\text{p}} \tag{6}$$

$$\text{Recall} = \frac{T\text{p}}{T\text{p} + F\text{N}} \tag{7}$$

$$\text{AP} = \int_0^1 P(R)\text{d}R \tag{8}$$

$$\text{mAP} = \frac{1}{n}\sum_{\text{i}=0}^{n} AP(i) \tag{9}$$

Where $T\text{p}$represents the number of correctly detected defective objects.$F\text{p}$represents the number of incorrectly detected defective objects.$F\text{N}$represents the number of missed defective objects. n represents the number of defect categories.$AP(i)$ denotes the average precision for the i-th object class.

## Experimental results and analysis

In this section, we present a detailed analysis of the experimental results obtained with the enhanced YOLOv9 model, including comprehensive comparisons with other leading object detection frameworks. Through carefully designed ablation studies, the individual contribution of each proposed module to the overall performance is systematically evaluated. Furthermore, visualization techniques such as heatmaps are employed to illustrate the detection outcomes, providing additional evidence for the efficacy of the introduced improvements.

## Model performance comparison

To evaluate the efficacy of the enhanced YOLOv9 model for remote sensing object detection, we conducted a benchmark comparison against several mainstream detectors, namely SSD, Faster R-CNN, YOLOv5, YOLOv7, and YOLOv8. The comparative results, summarized in Table 1, were obtained using multiple evaluation metrics—including Precision, Recall, mAP@0.5, mAP@0.5–0.95, and Frames Per Second (FPS)—to provide a comprehensive assessment of detection accuracy and computational efficiency.

Table 1 indicates that the SSD model attains a precision of 81.18%, a recall of 77.27%, and an mAP@0.5 of 81.99%, reflecting a relatively balanced performance profile. However, its detection accuracy and recall are slightly lower than those of the YOLO series models. Faster R-CNN achieves a higher recall of 81.38%, indicating a strong ability to identify positive instances, yet its precision is comparatively suboptimal at 64.55%, and its mAP@0.5 reaches 76.76%. Overall, its performance is inferior to both the YOLO family and the proposed algorithm. Among the YOLO series, YOLOv7 attains the highest precision of 84.5%, underscoring its effectiveness in minimizing false positives. In contrast, the proposed algorithm excels in recall, achieving 84.1%, which signifies its superior capacity to detect a greater number of true positive samples. YOLOv9 maintains a well-balanced performance with a precision of 83.9% and recall of 83.0%. In terms of mAP metrics, the proposed algorithm outperforms all comparative models, achieving 86.6% for mAP@0.5 and 71.5% for mAP@0.5–0.95, thereby demonstrating a significant advantage in detection accuracy. YOLOv9 follows closely, with mAP@0.5 and mAP@0.5–0.95 values of 85.7% and 70.8%, respectively. Regarding inference speed, YOLOv5s leads with an impressive 294.1 FPS, rendering it highly suitable for real-time applications. The proposed algorithm achieves a moderate FPS of 84.0, which is counterbalanced by its superior accuracy and recall, albeit at a higher computational cost of 304.0 GFLOPS. In terms of model size, YOLOv5s is the most lightweight with only 7.05 M parameters, contributing to its high inference speed. The proposed algorithm incorporates 39.97 M parameters, striking a balance between model complexity and performance, whereas YOLOv9, with the largest parameter count of 50.73 M, still delivers excellent mAP@0.5 and recall. In summary, while SSD and Faster R-CNN deliver moderate detection performance, the YOLO series—particularly YOLOv7, YOLOv9, and the proposed algorithm—exhibit outstanding results in both detection accuracy and overall metrics. The proposed algorithm leads in recall and mAP values, albeit with a slightly lower inference speed, making it highly suitable for practical applications that prioritize detection performance.

According to the comparative results of Average Precision (AP) across different models presented in Table 2, notable performance disparities are observed among various algorithms spanning multiple object categories. The proposed algorithm exhibits outstanding effectiveness across several key categories. In the domain of ground vehicles, the proposed method achieves an AP of 94.8% for "Car," surpassing all competing models and underscoring its superior detection accuracy. Similarly, for "Truck," it attains an AP of 88.8%, exceeding YOLOv9 (87.2%) and YOLOv7 (86.0%). In the "Van" category, the proposed algorithm reaches a top AP of 87.5%, outperforming YOLOv9 (85.6%). For "Long Vehicle," it leads with 87.2% AP, notably higher than YOLOv9 (82.0%) and YOLOv7/YOLOv8 (approximately 84.8%). In the "Bus" category, the proposed algorithm matches YOLOv9 at 93.8% AP, surpassing YOLOv7 (93.3%) and YOLOv5l (92.6%). In aircraft-related categories, the proposed algorithm also demonstrates excellent performance. It achieves 97.2% AP for "Propeller Aircraft," slightly above YOLOv9's 97.1%. For "Trainer Aircraft," it leads with 98.2% AP, compared to 97.3% for YOLOv9 and 96.9% for YOLOv8l. Although it attains 94.0% AP for "Chartered Aircraft," which is marginally lower than YOLOv8l's 95.8%, it still maintains a high detection standard. In the challenging "Fighter Aircraft" category, the proposed algorithm matches both YOLOv7 and YOLOv9 at 99.5% AP, indicating highly robust performance. For specialized vehicle types, the proposed algorithm continues to show competitive results. It achieves 62.1% AP for "Stair Truck," exceeding YOLOv9 (61.2%) and YOLOv7 (59.3%). In the "Pushback Truck" category, it reaches 72.9% AP, substantially outperforming YOLOv9 (71.3%) and YOLOv8l (66.8%). In the "Helicopter" category, it attains 91.6% AP, slightly below YOLOv5s (94.5%) yet still at a high level. For "Boat," it matches YOLOv9 with 98.4% AP, indicating consistently strong detection performance. Conversely, traditional detectors such as Faster R-CNN and SSD exhibit considerably lower AP across most categories, particularly in "Pushback Truck" (26.1%) and "Stair Truck" (45.1%), falling far short of the proposed method. In summary, the proposed algorithm delivers remarkable performance across numerous key categories—especially "Car," "Truck," "Van," "Long Vehicle," "Stair Truck," and "Pushback Truck"—demonstrating a substantial improvement in detection accuracy over existing models. It also excels in multiple aircraft-related categories. Although it is slightly outperformed in certain specific categories by some YOLO variants, it maintains an overall leading stance. The comprehensive and robust performance of the proposed algorithm underscores its strong potential as an effective tool for multi-category object detection in diverse scenarios, thereby providing a solid foundation for future applications and broader adoption.

## Ablation study

The ablation study in this work aims to scientifically validate the impact of each architectural modification. By progressively incorporating the C3 module, P2 detection head, GIoU loss, and SE module, we meticulously analyze their individual contributions to the final performance. This process not only confirms the necessity of each improvement for addressing specific limitations but also provides a clear understanding of the source of the performance gains.

Based on the ablation studies presented in Table 3, which evaluate the effect of successive improvements on the YOLOv9 model, each modification is shown to distinctly influence key performance metrics—including precision, recall, mean Average Precision (mAP@0.5 and mAP@0.5–0.95), inference speed (FPS), computational complexity (GFLOPS), and parameter count (Params). The baseline YOLOv9 model demonstrates a balanced performance profile, with a precision of 83.9%, recall of 83.0%, mAP@0.5 of 85.7%, and mAP@0.5–0.95 of 70.8%. It achieves a real-time inference speed of 58.4 FPS, with a computational load of 236.8 GFLOPS and 50.73 million

parameters, reflecting moderate model complexity. Introducing the C3 module leads to a marginal decline in precision and recall, while mAP@0.5 improves to 86.1%. Notably, the inference speed rises substantially to 97.0 FPS, underscoring the efficacy of the C3 module in accelerating detection. Although GFLOPS increase to 263.0, the parameter count is reduced to 49.84 million, indicating structural refinement. Further incorporation of the P2 detection head results in a precision of 82.9% but a higher recall of 82.4%. The mAP@0.5 improves slightly to 86.2%, and mAP@0.5–0.95 reaches 70.9%, suggesting that the P2 head enhances detection consistency, particularly for smaller objects. This comes at the cost of a lower FPS and higher GFLOPS, though the parameter count is further optimized to 39.93 million. Replacing the loss function with GIoU leads to a marked shift in the precision–recall trade-off: precision declines to 81.6%, while recall rises to 83.5%. The mAP values remain largely stable, indicating that GIoU improves localization quality and recall without significantly altering overall accuracy. Inference speed sees a minor drop to 85.4 FPS, with GFLOPS and parameters unchanged. Finally, integrating the SE attention module slightly lowers precision to 81.5% but raises recall to 84.1%. The mAP@0.5 reaches 86.6%, and mAP@0.5–0.95 attains 71.5%—the highest among all configurations—demonstrating the module's role in enhancing feature discriminability. The FPS remains acceptable for real-time use at 84.0, with negligible increases in GFLOPS and parameters. In summary, the cumulative integration of the C3 module, P2 detection head, GIoU loss, and SE module progressively enhances the model's detection capability, particularly in terms of recall and overall mAP, while maintaining real-time inference speed. These refinements collectively contribute to a more robust and accurate detector, well-suited for practical deployment.

## Result visualization

To visually validate the detection performance of the enhanced YOLOv9 model, we employ visualization techniques such as heatmaps to illustrate the model's focus regions during inference. As evidenced by the comparative heatmaps under different improvement configurations (see Fig. 7), the incorporation of the SE (Squeeze-and-Excitation) module and the P2 detection head markedly improves the model's capacity for precise localization of small and densely clustered objects. The SE module enhances feature discriminability by selectively emphasizing informative spatial channels, allowing the model to concentrate on semantically critical regions. Concurrently, the P2 detection head strengthens feature representation at higher resolutions, significantly boosting recognition performance for small targets and rendering their boundaries more distinct, even against cluttered backgrounds.

Heatmap exemplars from complex scenes further demonstrate the model's robust differentiation ability among multiple proximate targets. In scenarios with high object density, the refined model accurately localizes various vehicle types—such as cars and trucks—with notably fewer missed detections or false positives. These visual findings corroborate the efficacy of the proposed architectural improvements and provide insightful references for subsequent research and practical deployment.

The following figure shows the detection results of YOLOv5s, YOLOv5l, YOLOv7, YOLOv8s, YOLOv8l, YOLOv9, and the improved algorithm proposed in this paper on the remote sensing image object detection task. By comparing these images, one can more intuitively observe the differences in target localization and recognition among the different algorithms.

As illustrated in Fig. 8, YOLOv5s represents a significant benchmark in the YOLO family, renowned for its effective balance between detection speed and accuracy. The results demonstrate that YOLOv5s achieves competent performance in multi-object scenarios, successfully identifying common categories such as cars, trucks, and vans. However, the model exhibits certain limitations under challenging conditions, including occasional missed detections in complex backgrounds, suboptimal performance on small objects, and imprecise boundary localization for some targets.

As depicted in Fig. 9, YOLOv5l represents a scaled-up version of the YOLOv5 architecture, engineered to enhance detection accuracy by increasing both the depth and width of the network. Comparative analysis reveals that YOLOv5l achieves superior performance over its smaller counterpart, particularly in detecting small objects and maintaining stability in dense scenes. The visualization confirms that it produces more precise bounding boxes that align better with the actual object contours. This gain in accuracy, however, comes at the cost of increased computational complexity, resulting in a lower inference speed than YOLOv5s. Despite this trade-off, YOLOv5l remains a robust and effective choice for applications where detection precision is prioritized.

As illustrated in Fig. 10, YOLOv7 represents a significant advancement in the YOLO series, achieving a refined balance between inference speed and detection accuracy through architectural optimizations. The results demonstrate that YOLOv7 excels in object detection tasks, particularly in complex scenarios, where it accurately identifies multiple objects with minimal omissions or errors. Furthermore, the model provides superior bounding box regression and exhibits enhanced robustness, proving especially effective in dense object distributions and for small target detection.

As delineated in Fig. 11, YOLOv8s serves as the lightweight variant within the YOLOv8 series, prioritizing computational efficiency and inference speed for deployment in resource-limited scenarios. The results indicate that while YOLOv8s achieves competent performance in general multi-object detection, its accuracy is somewhat compromised in complex environments. Specifically, the model demonstrates a heightened susceptibility to false positives and missed detections, with performance limitations becoming more pronounced for small objects and in densely packed areas.

As presented in Fig. 12, YOLOv8l, as a larger variant within the YOLOv8 series, exhibits enhanced detection accuracy and superior model capability compared to its lighter counterpart, YOLOv8s. The results demonstrate that YOLOv8l achieves more precise localization of multiple targets and delivers more robust performance in complex environments. Notably, it shows significant improvements in handling occluded or densely clustered objects and in detecting small targets. These performance gains, however, are accompanied by increased

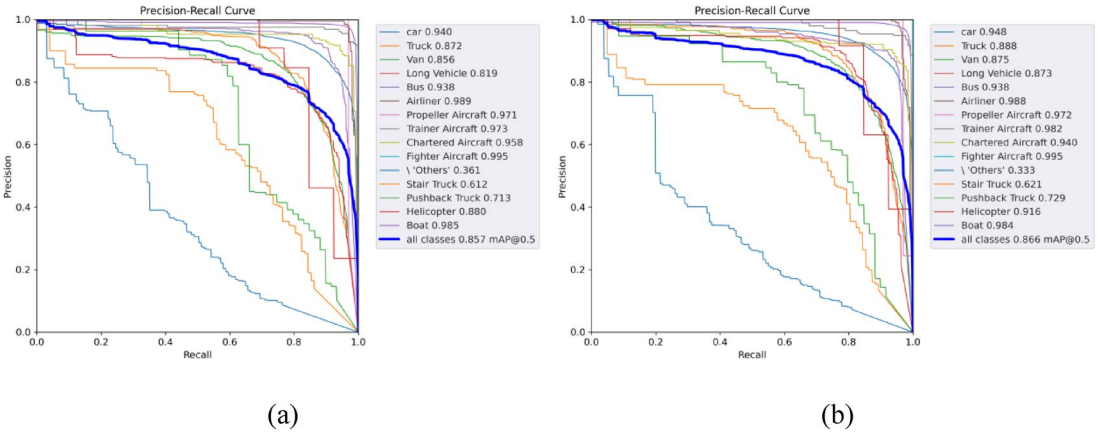


**Fig. 15**. Precision-Recall (P-R) curves: (**a**)YOLOv9; (**b**) Ours.

computational costs and longer inference times, reflecting the inherent trade-off between model capacity and operational efficiency.

As illustrated in Fig. 13, YOLOv9 represents a substantial evolution in the YOLO series, integrating advanced architectural refinements and optimization strategies. The detection results demonstrate its exceptional performance, particularly in complex and densely populated scenes, where it achieves a notable advancement in overall accuracy. In comparison to the YOLOv8 series, YOLOv9 exhibits a superior capability in resolving overlapping objects and delivers more precise bounding box localization, underscoring its enhanced perceptual and discriminative qualities.

As demonstrated in Fig. 14, the enhanced YOLOv9 algorithm proposed in this work—which integrates the C3 module, SE attention mechanism, P2 detection head, and GIoU loss function—achieves a performance level surpassing that of the standard YOLOv9. The results visually confirm that our method attains superior recognition accuracy in complex scenarios, exhibiting pronounced advantages particularly in detecting small objects and within densely packed areas. In direct comparison to YOLOv9, the proposed algorithm yields a reduced rate of false positives and delivers more distinctly defined detection boundaries, thereby providing compelling visual evidence for the collective efficacy of the introduced architectural refinements.

Figure 15 presents the precision-recall (P-R) curves for various target categories, enabling a clear visual comparison of performance differences between the two methods. Specifically, Fig. 15 (a) illustrates the performance of the baseline YOLOv9 model, while Fig. 15 (b) corresponds to the enhanced algorithm proposed in this work. The results demonstrate that the improved model attains a larger area under the P-R curve for most categories, indicating a consistent gain in detection performance. Notably, the proposed approach significantly enhances precision while maintaining high recall levels, thereby offering compelling quantitative evidence for the effectiveness of the algorithmic optimization introduced in this study.

As illustrated in the figure above, the Precision-Recall (P-R) curves for both the baseline and the improved algorithm are compared across multiple object categories. The enhanced model exhibits superior detection performance overall, achieving a mean Average Precision (mAP@0.5) of 0.866, which represents a modest yet consistent improvement over the 0.857 attained by the original algorithm. This advantage is particularly evident in several common vehicle categories. For instance, the improved model achieves an Average Precision (AP) of 0.948 for car, compared to 0.940 for the baseline. Similarly, higher AP values are observed for categories such as truck and van. A more pronounced improvement is noted in the long vehicle category, where the AP increases substantially from 0.819 to 0.873. Furthermore, the proposed algorithm demonstrates enhanced capability in detecting complex targets. In the helicopter category, the AP rises to 0.916 from 0.880. Similar gains are observed for pushback truck and stair truck, underscoring the robustness of the improvements across diverse object types. In summary, while the original algorithm maintains competitive performance in certain specific categories, the proposed improved algorithm delivers more balanced and superior results across the majority of detection tasks—especially for vehicle-related and complex objects. This leads to a higher overall mAP, confirming its stronger and more reliable detection capability.

## Discussion

### Result analysis

In this study, we introduce several enhancements to the YOLOv9 model, leading to notable improvements across multiple evaluation metrics. A key achievement is the observed gain in mAP@0.5–0.95, which we attribute primarily to the integration of the C3 module. By optimizing the feature extraction process, the C3 module reduces computational redundancy and strengthens feature representation, enabling more precise target localization in complex multi-object scenarios.

Furthermore, the incorporation of a P2 detection head substantially improves the detection of small objects—a common challenge in remote sensing imagery where conventional YOLOv9 exhibits limitations. Through multi-scale feature fusion, the P2 detection head enhances the model's sensitivity to fine-grained patterns, significantly boosting detection accuracy for small targets.

The adoption of the GIoU loss function also contributes critically to performance refinement. In contrast to the standard IoU loss, GIoU offers more informative gradient signals for bounding box regression, especially in cases of low overlap. This results in more stable training and a reduction in both false positives and missed detections under complex scene conditions.

Additionally, the embedding of the SE module allows for adaptive channel-wise feature recalibration, directing the model's attention to more discriminative features relevant to object detection. This mechanism further elevates the overall detection precision.

Despite these gains, the proposed model is not without limitations. The introduction of new modules has increased model complexity, leading to a measurable decrease in inference speed (FPS). Although the current FPS remains acceptable for many applications, further optimization is necessary for deployment in real-time scenarios. Moreover, while the P2 head improves small-object detection, some false positives and missed detections persist in highly crowded environments. Future work will focus on refining feature extraction and instance segmentation strategies to address these challenging cases.

### Comparison and improvement

In comparison with established object detection methods, the proposed model exhibits competitive advantages in several key metrics. For instance, relative to the YOLOv8 series, our enhanced YOLOv9 achieves higher performance in both mAP and Recall, which can be attributed to its more sophisticated network architecture and refined loss function. However, when compared to lightweight detectors such as YOLOv5s, the improved model falls short in inference speed, suggesting a potential trade-off between accuracy and efficiency in computational resource-limited or real-time-critical deployments.

Against conventional detectors like SSD and Faster R-CNN, the proposed model demonstrates clear superiority in both speed and detection accuracy. While methods such as SSD and Faster R-CNN often struggle with multi-object scenes and complex backgrounds, our model—through enhanced feature extraction and improved bounding box regression strategy—handles such scenarios more effectively. Furthermore, the incorporation of the SE attention module helps suppress false detections while preserving high precision, a capability less evident in earlier architectures including YOLOv8.

Nevertheless, despite its strong overall performance, the proposed model still exhibits limitations under specific challenging conditions. For example, in scenes with extremely small or heavily occluded objects, performance improvements remain constrained. This indicates the need for future work to explore more advanced network designs, data augmentation strategies, or feature fusion mechanisms to improve robustness in such cases.

In summary, the model introduced in this study achieves notable performance gains across multiple key benchmarks, particularly in complex environments and small object detection. However, the increased computational complexity necessitates future efforts to optimize inference efficiency without compromising accuracy, thereby better aligning the method with practical application constraints.

## Conclusion

This study focuses on enhancing the YOLOv9 model to improve its performance in remote sensing image object detection. To address the limitations of the original YOLOv9 in complex scenes and small object detection, we propose an improved architecture by integrating the C3 module, the SE (Squeeze-and-Excitation) attention mechanism, a P2 detection head, and the GIoU loss function. Extensive experiments were conducted to validate the effectiveness of the proposed model. The main findings are summarized as follows:

First, the incorporation of the C3 module significantly strengthens the feature extraction capability. By separating and fusing partial features, the C3 module reduces redundant computations and enhances representational capacity, leading to improved detection accuracy. In complex remote sensing scenarios, the upgraded model identifies targets more accurately, thereby reducing both missed and false detections.

Second, the introduction of the SE module improves the model's focus on informative features through adaptive channel-wise feature recalibration. Experimental results demonstrate that the SE module effectively boosts overall detection precision, particularly under complex backgrounds, where the model localizes target regions more accurately and suppresses false positives.

Moreover, the addition of a P2 detection head specifically mitigates the weakness of YOLOv9 in detecting small objects. By incorporating an extra low-level feature branch, the P2 head enhances multi-scale feature extraction, which proves especially beneficial in scenarios with dense small targets. Experiments confirm that the P2 head substantially elevates small object detection accuracy and effectively reduces omission errors.

Finally, the adoption of the GIoU loss function refines bounding box regression. In comparison with the conventional IoU loss, GIoU offers more meaningful gradient information in cases of low overlap between predicted and ground-truth boxes, facilitating faster convergence and improving localization accuracy in challenging scenes. The use of GIoU further stabilizes and refines bounding box regression, particularly for overlapping objects.

Despite the strong performance of the proposed model across multiple key metrics, certain limitations remain. For instance, the integration of new modules increases model complexity and slightly extends inference time. Future work should therefore aim to optimize computational efficiency to satisfy real-time application requirements.

In summary, the improved YOLOv9 model introduced in this study achieves notable performance in remote sensing object detection, exhibiting distinct advantages in complex environments and small target recognition. This work not only provides valuable insights for optimizing YOLO-series models but also offers practical support for remote sensing detection tasks. Future research may explore more efficient network architectures and training strategies to further advance detection accuracy and speed for diverse application needs.

## Data availability

The datasets can be downloaded in https://github.com/ihians/simd.

Received: 31 August 2024; Accepted: 11 November 2025
Published online: 29 December 2025

## References

1. Zeng, B. et al. Detection of military targets on ground and sea by UAVs with low-altitude oblique perspective. *Remote Sens.* **16** (7), 1288 (2024).
2. Chen, W. et al. Review of airborne oceanic lidar remote sensing. *Intell. Mar. Technol. Syst.* **1** (1), 10 (2023).
3. Jocher, G. et al. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. Zenodo. (2022).
4. Rajendran, G. B., Srinivasan, G. T., & Rathakrishnan, N. Adaptive YOLOv6 with spatialTransformer Networks for accurate object detection and Multi-Angle classification in remote sensing images. *Expert Systems withApplications*, **282**, 127796 (2025).
5. Wang, C. Y., Bochkovskiy, A. & Liao, H. Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (7464–7475). (2023).
6. Sohan, M., Ram, S. & Rami Reddy, C. V. T., A review on yolov8 and its advancements. In International Conference on Data Intelligence and Cognitive Informatics (529–545). (Springer, Singapore, 2024).
7. Wang, A. et al. Yolov10: Real-time end-to-end object detection. *Adv. Neural. Inf. Process. Syst.* **37**, 107984–108011 (2024).
8. Song, Y., Wei, S., Tian, B., Xu, S. & Zhang, L. LPI radar target detection performance optimization based on joint cognitive frequency transmission and power allocation. *Sig. Process.* **202**, 108736 (2023).
9. Su, J., Qin, Y., Jia, Z. & Liang, B. MPE-YOLO: enhanced small target detection in aerial imaging. *Sci. Rep.* **14** (1), 17799 (2024).
10. Lv, H., Qian, W., Chen, T., Yang, H. & Zhou, X. Multiscale feature adaptive fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
11. Hou, Y. et al. R-YOLO: A YOLO-based method for arbitrary-oriented target detection in high-resolution remote sensing images. *Sensors* **22** (15), 5716 (2022).
12. Li, Z. et al. RSI-YOLO: object detection method for remote sensing images based on improved YOLO. *Sensors* **23** (14), 6414 (2023).
13. Lu, X., Ji, J., Xing, Z. & Miao, Q. Attention and feature fusion SSD for remote sensing object detection. *IEEE Trans. Instrum. Meas.* **70**, 1–9 (2021).
14. Tu, Z., Yang, X., He, X., Yan, J. & Xu, T. RGTGAN: Reference-based gradient-assisted texture-enhancement GAN for remote sensing super-resolution. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–21 (2024).
15. Cai, S., Zhang, X. & Mo, Y. A lightweight underwater detector enhanced by attention mechanism, GSConv and WIoU on YOLOv8. *Sci. Rep.* **14** (1), 25797 (2024).
16. Zhao, Z. Q., Zheng, P., Xu, S. T. & Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Networks Learn. Syst.* **30** (11), 3212–3232 (2019).
17. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2961–2969). (2017).
18. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (1440–1448). (2015).
19. 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.* **39** (6), 1137–1149 (2016).
20. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proc. IEEE.* **111** (3), 257–276 (2023).
21. Soria, L. M., Ortega, F. J., Alvarez-Garcia, J. A., Velasco, F. & Fernandez-Cerero, D. How efficient deep-learning object detectors are? *Neurocomputing* **385**, 231–257 (2020).
22. Guo, Z., Zhang, W., Liang, Z., Shi, Y. & Huang, Q. Multi-scale object detection using feature fusion recalibration network. *IEEE Access.* **8**, 51664–51673 (2020).
23. Yun, H. & Park, D. Efficient object detection based on masking semantic segmentation region for lightweight embedded processors. *Sensors* **22** (22), 8890 (2022).
24. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 779–788 (2016).
25. Redmon, J. & Farhadi, A. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 7263–7271 (2017).
26. Redmon, J. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. (2018).
27. Wang, C. Y., Yeh, I. H. & Liao, H. Y. M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*. (2024).
28. Bochkovskiy, A. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. (2020).
29. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** (9), 1904–1916 (2015).
30. Cao, F., Xing, B., Luo, J., Li, D., Qian, Y., Zhang, C., … Zhang, H. (2023). An Efficient Object Detection Algorithm Based on Improved YOLOv5 for High-Spatial-Resolution Remote Sensing Images. Remote Sensing, **15** (15), 3755.
31. Liu, P., Wang, Q., Zhang, H., Mi, J. & Liu, Y. A lightweight object detection algorithm for remote sensing images based on attention mechanism and YOLOv5s. *Remote Sens.* **15** (9), 2429 (2023).
32. Ahmed, M., El-Sheimy, N., Leung, H. & Moussa, A. Enhancing object detection in remote sensing: A hybrid YOLOv7 and transformer approach with automatic model selection. *Remote Sens.* **16** (1), 51 (2023).
33. Shen, L., Lang, B. & Song, Z. DS-YOLOv8-Based object detection method for remote sensing images. *Ieee Access.* **11**, 125122–125137 (2023).
34. Liu, Y., He, G., Wang, Z., Li, W. & Huang, H. NRT-YOLO: improved YOLOv5 based on nested residual transformer for tiny remote sensing object detection. *Sensors* **22** (13), 4953 (2022).
35. Xie, T., Han, W. & Xu, S. Yolo-rs: a more accurate and faster object detection method for remote sensing images. *Remote Sens.* **15** (15), 3863 (2023).
36. Ma, C. et al. YOLO-UAV: object detection method of unmanned aerial vehicle imagery based on efficient multi-scale feature fusion. *IEEE Access* **11** 126857–126878 (2023).
37. Su, Z., Yu, J., Tan, H., Wan, X. & Qi, K. Msa-yolo: a remote sensing object detection model based on multi-scale strip attention. *Sensors* **23** (15), 6811 (2023).
38. Peng, G., Yang, Z., Wang, S. & Zhou, Y. AMFLW-YOLO: A lightweight network for remote sensing image detection based on attention mechanism and Multi-scale feature fusion. *IEEE Trans. Geoscience Remote Sensing* **61** 1–16 (2023).
39. Zhang, J., Chen, Z., Yan, G., Wang, Y. & Hu, B. Faster and lightweight: an improved YOLOv5 object detector for remote sensing images. *Remote Sens.* **15** (20), 4974 (2023).
40. Liu, J. & Liu, Z. YOLOv5s-BC: an improved YOLOv5s-based method for real-time Apple detection. *J. Real-Time Image Proc.* **21** (3), 1–16 (2024).

41. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 7132–7141 (2018).
42. Roy, A. G., Navab, N. & Wachinger, C. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I* (421–429). Springer International Publishing. (2018).
43. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (3–19). (2018).
44. Rezatofighi, H. et al. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 658–666 (2019).
45. Zheng, Z. et al. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence* **34** (07), 12993–13000 (2020)
46. Haroon, M., Shahzad, M. & Fraz, M. M. Multisized object detection using spaceborne optical imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **13**, 3032–3046 (2020).
47. Cui, Z., Wang, X., Liu, N., Cao, Z. & Yang, J. Ship detection in large-scale SAR images via Spatial shuffle-group enhance attention. *IEEE Trans. Geosci. Remote Sens.* **59** (1), 379–391 (2020).

## Acknowledgements
The authors received no financial support for the research, authorship, and/or publication of this article.

## Author contributions
Chunping Wang and Ming Chen conceived and designed the methodology and wrote the initial draft of the manuscript. Ying Yu was primarily responsible for creating the figures and provided significant suggestions on the theoretical and methodological sections of the paper. Yirui Li and Yifan Li were mainly responsible for data collection and validation. Xianglong Chen was primarily responsible for data analysis.

## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to C.W.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.